Transportation Research Part C 128 (2021) 103142

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Understanding vehicles commuting pattern based on license plate recognition data

Wenbin Yao [a], Maolei Zhang [b], Sheng Jin [a,d,*], Dongfang Ma [c,d,*]

[a] Institute of Intelligent Transportation Systems, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China
[b] Alibaba Cloud Computing Co. Ltd., Hangzhou 311100, China
[c] Institute of Marine Information Science and Technology, Ocean College, Zhejiang University, Hangzhou 310058, China
[d] Pengcheng Laboratory, Shenzhen 518052, China

ARTICLE INFO

Keywords:
License plate recognition
Commuting vehicles
Travel behavior
Mobility pattern

ABSTRACT

Commuting vehicles are one of the most important vehicles on the road network, especially during morning and evening peak hours. Analysis of commuting vehicles can lay foundation for policy formulation, urban planning, business decision-making. A large number of license plate recognition (LPR) detectors installed on the road network have accumulated a large amount of LPR data that contain the spatio-temporal information on vehicles. Therefore, using the LPR data, the commuting behavior of vehicles can be mined and analyzed. This study takes the LPR data of Hangzhou, China as an example, a series of algorithms are used to extract a total of nine features reflecting the commuting travel behavior using the one-month LPR data, and then the factor analysis method is used to integrate the nine extracted features into three factors. The commuting behavior is characterized from three perspectives: the stability of the vehicle travel behavior during the high frequency travel time period on working days, the stability of the vehicle travel behavior during the non high frequency travel time period, the stability of the origin and destination (OD) of a vehicle during the high frequency travel time period. Based on these three factors, three methods, including the density-based spatial clustering of applications with noise algorithm (DBSCAN), the iterative self-organizing data analysis technique algorithm (ISODATA), and the fast search and find of density peaks clustering algorithm, are used to identify commuting pattern vehicles and their results are compared. Finally, the decision tree algorithm is used to extract the commuting rule and identify commuting vehicles from the collected data. After that, The influence of different high frequency travel time period on commuting rule extraction and commuting pattern vehicle recognition is carefully analyzed. The influence of randomness on the results is also tested. The proposed commuting pattern vehicle recognition algorithm is proved to be robust. The detection frequency and the first and last detected times of commuting vehicles and non commuting vehicles are then analyzed. 78.0% of commuting vehicles were first detected between 06:30 AM and 10:30 AM on each workday on average. And 77.8% of commuting vehicles were last detected between 03:30 PM and 09:30 PM on each workday on average. However, for non commuting vehicles, the phenomenon that start or end the travel during the high frequency travel time period is not obvious. This study lays a foundation for travel policy formulation. For instance, it can provide a reference for the determination of the restricted periods when implementing the travel restriction policy. In addition, the identification of

* Corresponding authors at: Institute of Intelligent Transportation Systems, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China.
E-mail addresses: jinsheng@zju.edu.cn (S. Jin), mdf2004@zju.edu.cn (D. Ma).

https://doi.org/10.1016/j.trc.2021.103142
Received 13 June 2020; Received in revised form 1 March 2021; Accepted 8 April 2021
Available online 5 May 2021

commuting vehicles is extremely useful for public transit network design and optimization, and it can help the decision-making process of a customized commuter bus. In addition, these strategies are expected to reduce the use of private cars and encourage travelers to switch to public transportation.

1. Introduction

Commuting is one of the main travel purposes of most travelers. During the morning and evening peak hours, commuting travel accounts for a larger proportion than during the off-peak hours, so commuting vehicles have a very large impact on traffic conditions. Understanding commuting behavior can help to determine the occupational and residential distribution in a certain area (Zhou et al., 2014; Huang et al., 2019), as well as provide support for urban planning, which is of great significance to urban planners. For instance, Ma et al. (2017) analyzed transit commuting behavior in Beijing based on the transit smart card data and found that there had been a serious imbalance between occupational and residential areas in Beijing. In the later stage of urban planning, planners can focus on adding some residential areas in commercial areas, while adding some commercial land in residential areas in order to reduce the imbalance between occupational and residential areas. The analysis of commuting vehicle behavior can also help to understand the travel behavior on the individual level, which can be helpful for policy formulation (Li et al., 2019), such as policy formulation of travel restriction. Namely, the policy formulation aims to reduce the demand for vehicles use during peak hours through the introduction of travel restriction policies and encourage people to switch to public transportation or choose to travel during off-peak hours. Analysis of commuting vehicle behavior lays the foundation for specifying the restricted periods in the travel restriction policy and assessing the effects of travel restriction policies (Yao et al., 2018). The analysis of commuting travel behavior can also be helpful to companies to make commercialization decisions. For instance, Dong et al. (2018) used taxi trip records and internet-based ride-sharing trip records provided by DiDi company to analyze the difference in travel behavior between online car-hailing passengers and taxi passengers, and determined commuting travelers through cluster analysis. Finally, it was found that the proportion of commuting trips of online car-hailing passengers had been relatively higher, but in both taxi and online car-hailing, the proportion of commuters had not been high in general. Therefore, analyzing the origin and destination (OD) of commuters is helpful for online car-hailing companies and taxi companies to allocate vehicles. Besides, the analysis of commuting behavior is also useful for bus route planning and optimization (Ma et al., 2012). Transit agencies usually hope that they can attract passengers as more as possible using a series of different marketing strategies (Strathman et al., 2008). Since commuters are one of the main target groups of bus companies, analysis of commuters can help bus companies develop business strategies specifically for this group. In addition, the analysis of commuting travel behavior can also provide support for the determination of bus stations and bus route planning of a customized commuter bus companies(Qiu et al., 2018). The researchers have been studying the behavior of commuters and the factors that affect the commuting travel behavior with the aim to develop strategies to guide commuting green travel (Gutiérrez et al., 2020; Ye et al., 2020; Jia et al., 2018).

Therefore, the analysis of commuting behavior is urgently needed. Such an analysis can provide the necessary support to urban planners and policy-makers, as well as assist business decision-making. It also plays a significant role in bus route optimization, customized commuting bus station determination, and new bus route planning. However, there has been no general definition of commute. Namely, commute can be defined as a trip between the same origin and destination during a long period. The origin and destination usually denote the home and workplace, respectively (Kung et al., 2014). Ma et al. (2017) defined public transport commuter as a regular transit rider who performs periodically recurring travels between home and other non-residential locations. In this study, we define the commute as a type of a regular trip, usually occurring during the morning or evening peak hours, during which the commuter travel between the same origin and destination over a long period of time.

This article aims to solve the three following problems: (1) what are the main behavior characteristics of commuting vehicles, and how to extract the features that reflect the commuting behavior based on license plate recognition (LPR) data; (2) how to design a more efficient algorithm that can be used for commuting pattern recognition; namely, such an algorithm needs to identify the commuting pattern as accurately as possible and must be able to meet the requirements of commuting pattern recognition for a large number of samples; (3) which conclusions can be drawn after screening out the commuting vehicles and analyzing their travel behaviors, and how can these conclusions be applied to engineering practice. In order to solve the above three problems, first, the high frequency travel time period in Hangzhou are identified, and an algorithm for identifying OD of a vehicle during the high frequency travel time period is designed. After that, based on a detailed analysis of commuting behaviors, a total of nine features that reflect commuting behaviors are extracted, and significant outliers are removed. Correlation analysis is performed after data processing. Due to the high correlation between the features, the factor analysis method is used to reduce the dimensionality of the features, and finally, three independent factors are obtained. One percent of the entire dataset is sampled, and then using three different clustering methods, the DBSCAN algorithm(Ester et al., 1996), the ISODATA algorithm(Ball and Hall, 1965), and the fast search and find of density peaks clustering algorithm(Rodriguez and Laio, 2014) on the one percent sub-dataset from random sampling, after that, the clustering results are analyzed. The Silhouette coefficient is used to compare the clustering results of each algorithm. Based on the cluster analysis results, it is proposed to use the decision tree algorithm to extract the commuting rule, which not only helps to understand the commuting behavior but also helps to extract the commuting identification rules that are conducive to be implemented in the relational database. In that way, the commuting vehicles in the entire data can be identified. After identifying the commuting vehicles in the entire data, a more detailed analysis of their travel behaviors and travel characteristics is conducted.

The remaining of this paper is organized as follows. Section 2 reviews relevant literature about traffic state estimation, travel

behiviour analysis and commuting behavior analysis based on LPR data. Section 3 introduces the data used in this study. Section 4 uses three clustering methods to cluster vehicles and compares the advantages and disadvantages of the three methods. Section 5 uses a decision tree model to extract commuting rules. Section 6 analyzes the travel characteristics of commuting vehicles and discusses the robustness of the proposed commuter pattern vehicle recognition algorithm. Finally, Section 7 draws the main conclusions and discusses future research directions.

2. Literature review

Many studies have analyzed the commuting behavior, but there have been some difficulties in analyzing individual commuting behavior. Namely, it is difficult to obtain a large amount of data needed for analysis. The previous studies have been mostly based on questionnaire surveys. However, the amount of questionnaire survey data is usually relatively scarce because the collection of such data requires much manpower and material resources, and the collection is very inconvenient. Door to door surveys are usually required. Also, there are often abnormal data in the questionnaire survey data because respondents can be sensitive to certain questions (Chen et al., 2017). Therefore, the analysis based on questionnaire survey data can often be unreliable due to the small amount of data and the existence of abnormal data.

At present, a large number of LPR detectors have been installed on the road network in many cities in China. These detectors capture passing vehicles and record the information on them, such as license plate number, vehicle type, driving lane, and time of detection. The LPR detectors accumulate a large amount of LPR data, which can be used for estimation and prediction of traffic state parameters, as well as for analysis of vehicle travel behavior and traffic demands (Luo et al., 2019; Ma et al., 2017a; Shen et al., 2020; Wang et al., 2016). The LPR data have received widespread attention because of its wide coverage, large amount, and high quality. The main problem is that there exist privacy issues, and it is difficult to obtain and use LPR data in practice. However, once these data are applied to engineering, it will play a significant role, and it is very promising to use the LPR data for solving the problems in transportation in the future.

Traffic state estimation based on license plate recognition data is an important research direction, which is always combined with artificial intelligence methods such as deep learning. Mo et al. (2017) proposed a modified car-following model corresponding to LPR data characteristics, and the model can fully capture the pattern of ground truth speed profile. LPR data can also be used to estimate the real-time lane queue length. Zhan et al. (2015) proposed a hybrid framework that combines both machine learning techniques and traffic flow theory, which can obtain detailed lane level queue length estimation efficiently in real time. Shao and Chen (2018) proposed a LPR data and collaborative tensor decomposition (CTD)-based method to estimate the sparse traffic volume data. There are many research results in this area, and this is an important application direction of LPR data. These researches provide a lot of algorithms and ideas of LPR data processing.

Because LPR data contains the spatiotemporal information of each individual vehicle, travel behavior analysis based on LPR data is another important research direction. Chen et al. (2017) extracted vehicle travel behavior features, classified the vehicles based on the extracted features, and analyzed their travel behaviors based on the LPR data. Liu et al. (2018) explored the changes in traffic conditions under two different vehicle travel restriction policies: One Day Per Week (ODPW) and Odd And Even (OAE), and they found that vehicle restriction stimulates more “illegal” travels and higher travel intensities. Chang et al. (2018) studied the impact of vehicle travel restriction policies on the behavior of vehicles with non-Shanghai license plates and analyzed their travel behaviors. Rao et al. (2018) proposed a trajectory reconstruction method based on the particle filter, and the OD patterns are estimated based on the trajectory using LPR data.

Many models describing vehicle travel behavior and predicting traffic state parameters have been proposed, which can lay a foundation for mining LPR data.However, there have been fewer studies on commuting behavior analysis based on the LPR data. Chen et al. (2017) extracted ten spatial and temporal variables describing vehicle travel behavior on weekdays and weekends, successively, based on the LPR data, and then used the *k*-means algorithm to classify the vehicles. The cluster numbers for weekdays and weekends were six and three, respectively. The clustering results using ten spatial and temporal variables of the working days showed that the second cluster might be commuting vehicles. However, the selected spatial and temporal variables in the mentioned study are not completely designed for identifying commuting vehicles, so they may not fully reflect commuting characteristics. In addition, the *k*-means algorithm is very sensitive to the selection of the initial clustering center, and requires cluster number be determined in advance. Thus, the selection of the initial cluster number has a significant impact on the results. Besides, the complexity of the *k*-means algorithm is relatively high, and often results cannot be obtained within an acceptable time for a large number of samples. Chang et al. (2018) proposed a two-step *k*-means clustering algorithm to cluster vehicles, where in the first step, commuting use of vehicles was identified based on its spatiotemporal pattern, and in the second step, the non-commuting vehicle use was analyzed. The *k*-means clustering algorithm was utilized in the two-step mining process. This method can obtain accurate results in describing the commute characteristics, but the above-mentioned problems of the *k*-means algorithm still exist.

In a word, LPR data has the advantage of high quality and wide coverage, which contains the spatiotemporal information of vehicles. Therefore, the recognition of commuting vehicles based on LPR data can be relatively accurate, and the result almost cover all commuting vehicles. However, due to the huge amount of LPR data, the traditional data processing methods are difficult to successfully process all LPR data. The method proposed in this study can solve the problem that the complex algorithm can not give result in an acceptable time in the huge amount of LPR data. Though LPR data has so many advantages, the uneven distribution of the detectors may lead to the deviation of the description of vehicle behavior, and then make the recognition of commuting vehicles biased.

Table 1
Detailed Information of LPR detectors.

Name	Information	Explanation
DEV_ID	2,150,091	Serial number of the detector
DEV_NAME	Intersection of YUGULU Road and QIUSHILU Road	Name of detector's location
XY	120.127709,30.265471	Coordinate of the detector

Table 2
Detailed Information of passing cars.

Name	Information	Explanation
ID	8,071,219,675	Serial number of the record
DEV_ID	2,150,091	Serial number of the detector
DEV_NAME	Intersection of YUGULU Road and QIUSHILU Road	Name of detector's location
WAY_ID	1	Lane
CAR_NUM	ZHE AT****	License plate number
CAR_TYPE	1	Type of car
CAP_DATE	2016/6/13 0:09:15	Time of record

3. Data description

3.1. Data source

The data used in this study denote the LPR data that includes the records on passing cars in Hangzhou City, Zhejiang Province, China, in June 2016. In 2016, Hangzhou city had a resident population of 9.188 million, and the number of motor vehicles was 2.29 million. Besides, in 2016, on the Hangzhou city's streets, there were 1,472 LPR detectors. The data collected by these LPR detectors consists of two parts, the first part denotes the LPR detector information, and the second part is the information of passing cars. The details and examples about the data are shown in Table 1 and Table 2. The average daily number of detected vehicles in June 2016 was 1.255 million, and the average daily number of detection records was 8.81 million.

LPR detectors are mainly distributed in the core area of Hangzhou. The core area of a city represents an area where the city's political, economic, and cultural activities are most concentrated, with a certain agglomeration effect.

3.2. Data cleaning

After analyzing the data, it was found that the quality problems of the data mainly related to the lack of license plates, incorrect license plate number recognition, and repeated detection by LPR detectors.

As for the lack of license plates, the main problem is that the 'CAR_NUM' item of the data with missing license plate was zero. The average daily missing data volume was about 400,000 rows, while the average daily total detection volume was about 9.6 million rows, so the missing amount accounted for about 4.17%. The missing data had a small effect on the whole, so the abnormal data was removed.

As for incorrect license plate number recognition, the main problem was that the number of characters in the license plate number was incorrectly identified. In general, a license plate number consists of seven characters. After data filtering, it was found that the number of license plates whose number of characters differed from seven was about 3000 records a day. Since the number of these license plates was very small (accounting for about 0.03%), all license plates whose number of characters differed from seven were removed.

In addition, during the actual shooting by the LPR detectors, there are cases in which two or more adjacent lanes are equipped with cameras. Thus, these cameras capture not only vehicles in their lane, but also vehicles in adjacent lanes. As a result, some vehicles have been detected two or more times at the same time, which accounting for about 1% of the total number of detections. For repeated detected records, randomly select a record to keep.

4. Commuting pattern recognition based on clustering

4.1. Identification criteria for high frequency travel time period

The high frequency travel time period(HFTTP) is defined as the period that vehicles gather to travel, and it also represents the period of serious traffic congestion and high incidence of traffic accidents. Therefore, it is very important to identify and analyze these hours, not only to lay the foundation for subsequent commuting pattern recognition but also to help understand the traffic state of an area based on which the corresponding traffic control measures can be defined.

Weng and Lv (2019) calculated the average daily card-swiping time of subway passengers on weekdays and identified the morning peak hours in Guangzhou, China, as 07:00AM–10:00AM and the evening peak hours as 05:00 PM – 08:00 PM. Sun and Yang (2015)

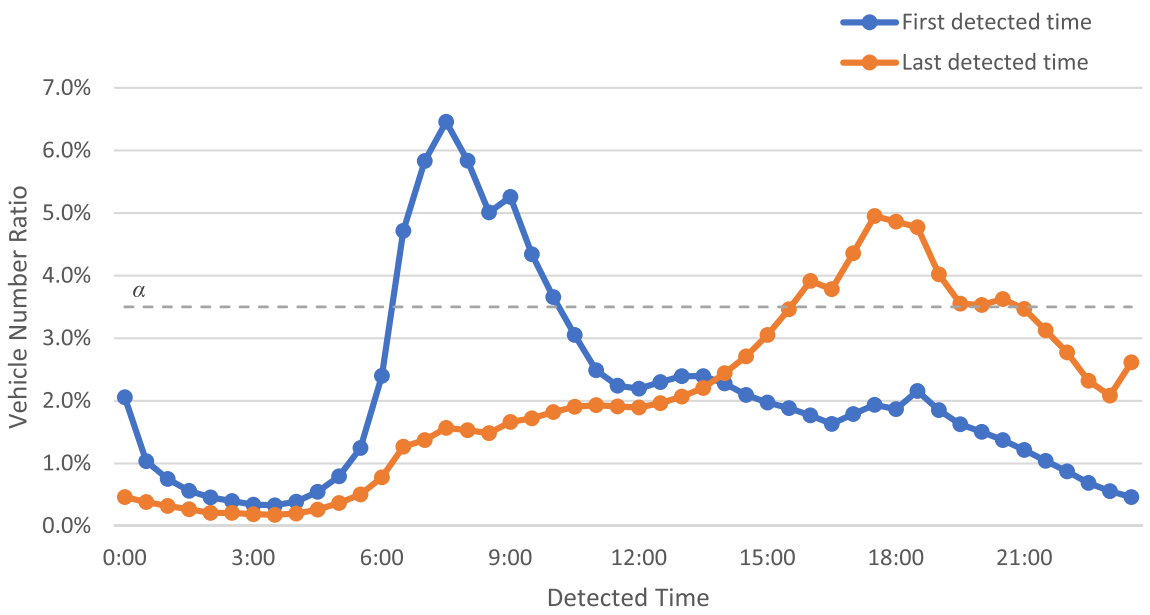


Fig. 1. Average first and last detected time distributions on all workdays.

Table 3
HFTTP in Hangzhou in June 2016 with different α values.

α	3%	3.5%	4%
HFTTP in the morning	06:30 AM – 11:00 AM	06:30 AM – 10:30 AM	06:30 AM – 10:00 AM
HFTTP in the evening	03:00 PM – 10:00 PM	03:30 PM – 09:30 PM	05:00 PM – 07:30 PM

used the "first card-swiping time" and "last card-swiping time" on weekdays to build a "card-swiping time pair" matrix, and conducted the statistical and measurement analysis on the first and last card-swiping times on weekdays using approximately 45% of the data including smart card and questionnaire survey in Xiamen, China in May 2014. They found that on weekdays, most bus passengers start their first trip between 06:30 AM and 09:00 AM and swipe the transit card for the last time between 04:00 PM and 09:00 PM. Ma et al. (2013) identified the travel mode of travelers using the DBSCAN clustering method. This method can also be used to find the routine travel time and return time of commuters. According to these studies, the first and last detected time on each working day in June 2016 are obtained to determine the HFTTP based on the LPR data. An interval of 30 min is used, so each day is divided into 48 intervals(Ma et al., 2017). We analyzed the average first and last detected time distributions of all workdays, and they are presented in Fig. 1.

Since the determination of HFTTP is not the focus of this study, a simple threshold method is used to determine the HFTTP. The vehicle number ratio greater than α is defined as the HFTTP. Different threshold will get different HFTTP, and the determination of the threshold value needs to be determined according to the local conditions in different regions. Table 3 gives the HFTTP when α is 3%, 3.5% and 4%.

In this study, 3.5% is taken as an example to carry out the follow-up study, and the influence of taking different thresholds on the results is discussed in detail in the **Discussion**.

4.2. Origin and destination identification of commuting trip

In this work, a commute is defined as a regular trip, usually occurring during the morning and evening peak hours, during which the commuter travels between the same origin and destination (usually home and work or school) over a long period of time. Therefore, determining the origin and destination of the trip to home or work is very important for characterizing the commuting trips. Ma et al. (2017) extracted the trip chains of travelers and regarded the transit riders' first and last trip as their commuting trips. Once the commuting trips are found, the origin and destination of the trip can be easily obtained. However, the last trip of the day can possibly be not the end of the commuting trip but can have entertainment or other travel purposes. Chen et al. (2017) regarded the detected interval of vehicles of less than an hour as the same trip so that it could break the trip chains and get trips of travelers, after that, the origin and destination of commuting trips could be identified.

As already mentioned, in this study, the LPR data is used for analysis. Due to the limitations of these data, it is impossible to get the actual OD of a traveler. Using the LPR data, only the information on some detected locations on the travel trajectory can be obtained, but not the location information of the traveler at each moment along the travel trajectory. Thus, the real OD information of commuting trips can not be obtained, but only the information related to the first and last detected locations. However, the purpose of

this study is not to discover the actual OD information of travelers, and the OD of commuting trips is used only to measure the stability of the travel behavior of a traveler from the spatial perspective. Therefore, the deviation of the actual OD and the extracted OD caused by the limitations of LPR data does not have a great impact on the recognition accuracy of the commuting pattern. In this study, the first and last detected locations are denoted as the origin and destination of a trip, respectively.

Based on the identified HFTTP, a simple but efficient algorithm for obtaining the origin and destination of commuting trips, which can also be used to analyze the stability of commuting trip's OD, is proposed. The proposed algorithm combines the characteristics of LPR data and considers the travel route selection problem of commuting trips, as well as the possible problems caused by missing capture of LPR detectors, in order to determine the OD and its stability coefficient of a commuting trip.

The OD stability coefficient of a commuting trip is defined as follows.

Assume that the origin and the destination of a commuting trip at the HFTTP in the morning are denoted as O_1 and D_1, respectively, and the ones at the HFTTP in the evening as O_2 and D_2, respectively. The stability coefficient of O_i is defined as Eq. (1).

$$F_{O_i} = N_{O_i}/N_{\mathrm{w}} \tag{1}$$

where N_{oi} denotes the number of days when the vehicle is first detected at O_i at the HFTTP in the morning/evening, and N_w denotes the number of working days.

Similarly, the stability coefficient of D_i is defined as Eq. (2).

$$F_{D_i} = N_{D_i}/N_{\mathrm{w}} \tag{2}$$

where N_{Di} denotes the number of days when the vehicle is last detected at D_i at the HFTTP in the morning/evening.

The algorithm for obtaining the OD of a commuting trip and calculating its stability coefficients is shown in **Appendix A**.

4.3. Commuting features extraction

As already mentioned, commuting trips are defined as a type of regular trips, usually occurring during morning and evening peak hours, and during which the commuter travels between the same origin and destination over a long period of time. Ma et al. (2017) proposed the most frequent departure time, the number of days when each transit rider travels and route-level and stop-level similarities to describe commuting behavior based on smart card data. Chang et al. (2018) described commuting behavior using number of days on which each vehicle was detected during the peak morning hours and peak evening hours and the number of days on which each vehicle was detected during the off-peak hours based on LPR data. Based on the features proposed by the existing research, a total of nine features are adopted to characterize the commuting pattern from the time and space perspectives. The first four features mainly describe travelers' travel behavior from the time dimension, and the remaining five features mainly describe travelers' travel behavior from the spatial dimension. Travel frequency coefficient describes travel behavior stability in a long period of time. Travel frequency coefficient at HFTTP mainly describes travel behavior stability at HFTTP, average detected frequency at non-HFTTP and standard deviation of detected frequency at non-HFTTP mainly describe the travel behavior stability at non-HFTTP, and the remaining features mainly describe the stability of the origin and destination at HFTTP. There is information overlap among the nine features, but no two features contain the completely same information. In order to describe the commuting behavior as comprehensively as possible, all features are taken into account in this study, and then dimension reduction method is used to eliminate the correlation.

Travel frequency coefficient at HFTTP

Firstly, the travel characteristics of commuting vehicles are analyzed from the perspective of daily travel time. As well-known, commuting vehicles have regular daily travel characteristics. Thus, from the temporal perspective, commuting vehicles appear on the road at a consistent time each day and generally travel at HFTTP. Therefore, the first travel feature of commuting vehicles is defined as the travel frequency coefficient at HFTTP, and it is calculated as Eq. (3).

$$F_1 = \frac{N_1}{N_w} \tag{3}$$

where N_1 denotes the number of working days when a car travels at HFTTP, and N_w denotes the total number of working days.

(2) Travel frequency coefficient

Next, the stability of commuter vehicle behavior during a long period is examined. Ideally, a commuter vehicle should travel at HFTTP on each working day. However, considering that commuters can have overtime work, early departure and time off work on working days, and not all commuters have the same number of working days, commuters may not travel on each working day, and even if they do, they may not travel at HFTTP. Therefore, the second feature named the travel frequency coefficient is used to measure the long-term travel stability of vehicles, and it is calculated as Eq. (4).

$$F_2 = \frac{N_2}{N_w} \tag{4}$$

where N_2 denotes the number of days, a commuter vehicle is detected on the road.

Average detected frequency at non-HFTTP

The commuting behavior is also examined from the perspective of travel regularity and travel frequency at non-HFTTP. Commuting vehicles generally do not frequently travel at non-HFTTP because these periods usually include the working time or lunchtime of workers, and they generally do not drive during that time, most workers would like to have lunch in company or nearby. However, some of the workers can choose to go out to eat or to go home for lunch, which includes the driving; but generally, a small number of

Table 4
KMO and Bartlett's test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0.827	
Bartlett's Test of Sphericity	Approx. Chi-Square	34095.030
	df	36
	Sig.	0.000

workers drive at non-HFTTP. Therefore, the average detected frequency at non-HFTTP is introduced to reflect the travel frequency at non-HFTTP, and it is calculated as Eq. (5).

$$M = \frac{\sum_{i=1}^{N_w} f_i}{N_w} \tag{5}$$

where f_i denotes the detected frequency of a vehicle at non-HFTTP on day i.

Standard deviation of detected frequency at non-HFTTP

Besides the average detected frequency at non-HFTTP, it is also needed to measure the travel regularity at non-HFTTP, so the standard deviation of the detected frequency at non-HFTTP is defined to reflect the travel regularity of a vehicle at non-HFTTP, and it is calculated as Eq. (6).

$$\sigma = \sqrt{\frac{1}{N_w}\sum_{i=1}^{N_w}(f_i - M)^2} \tag{6}$$

where M denotes the average detected frequency at non-HFTTP.

OD stability coefficient of commuting trip

The commuting behavior is analyzed from the perspective of OD stability over a long period of time. Commuting vehicles generally travel between home and working place or school, so the origin and destination are relatively stable. The aforementioned features of a commuting trip, namely F_{O1}, F_{D1}, F_{O2}, and F_{D2} reflect the stability of the origin and destination of travelers who travel at HFTTP. The calculation of these four features is shown in Eqs. (1) and (2).

OD stability coefficient from spatial perspective

The commuting behavior is analyzed from the perspective of spatial locations of O_1 and D_2, as well as O_2 and D_1. Ideally, O_1 and D_2 of the commuter should denote the same location, as well as O_2 and D_1. However, in actual situations, the travel routes of the commuter from home to work and from work to home are not necessarily the same. Therefore, O_1 and D_2 can differ, as well as O_2 and D_1. However, for a commuter vehicle, O_1 and D_2 should be relatively close in space, as well as O_2 and D_1. That means the distances between them should be small, so the stability coefficient of OD from the spatial perspective is defined to measure the above relationship that O_1 and D_2 should be relatively close in space, as well as O_2 and D_1. Before the calculation formula of this coefficient is given, it is needed to introduce the OD distance matrix, which is defined as Eq. (7):

$$\overrightarrow{D} = \begin{array}{c} \\ O_1 \\ O_2 \end{array} \begin{array}{c} \begin{array}{cc} D1 & D2 \end{array} \\ \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \end{array} \tag{7}$$

where x_{ij} denotes the linear distance between O_i and D_j.

After the distance matrix is obtained, the OD stability coefficient from the spatial perspective can be calculated as Eq. (8).

$$F_s = \frac{\left|\overrightarrow{D}\right|}{(max\{x_{11}, x_{22}\})^2} - \frac{x_{12} + x_{21}}{max\{x_{11}, x_{22}\}} \tag{8}$$

This feature can help us measure the spatial stability between the origin and destination. When O_1 and D_2 are the same points, as well as O_2 and D_1, the ideal situation is reached, and $F_s = 1$. The longer the distances between O_1 and D_2, and O_2 and D_1 are, the more serious F_s deviates from one, and a vehicle is less likely to be a commuter vehicle. The extracting method of the OD of a vehicle from the LPR data is as described above.

4.4. Dimension reduction

After getting all the features of vehicles, the box-and-whisker plot method (Wickham and Stryjewski, 2011; Santoyo, 2017) is used to clean the data, then the KMO and Bartlett's Test are conducted to check whether the data is suitable for structure detection, and the obtained results are shown in Table 4.

According to the KMO and Bartlett's Test, there is a strong correlation between features. The factor analysis method (Fruchter, 1954) is used to remove the correlation, and identify the underlying relationships between features. This method is a commonly used dimension reduction method. It can also be used to extract some common factors. In this method, every two common factors are linearly independent of each other, and each original observed variable can be expressed as a linear combination of common factors. The factor analysis method has been often used as a statistical analysis method in psychology, physiology, sociology, history, and other

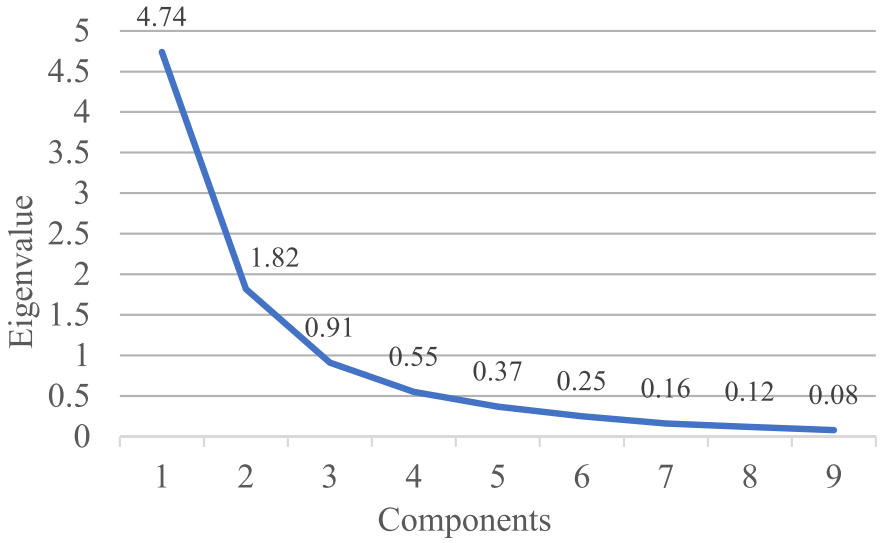


Fig. 2. The Scree plot.

Table 5
Factor loading.

	Factor 1	Factor 2	Factor3
F_1	0.912	0.064	0.152
F_2	0.802	0.284	−0.050
F_{O1}	0.892	−0.031	0.106
F_{D1}	0.886	−0.080	0.136
F_{O2}	0.873	−0.102	0.105
F_{D2}	0.872	−0.059	0.106
F_s	0.186	−0.007	0.977
σ	0.027	0.931	−0.052
M	−0.037	0.918	0.039

fields (Cattell, 2012), but currently, it is less commonly used in fields such as physics, biology, and chemistry. Generally speaking, when all observed variables can be summarized by fewer potential variables, then the factor analysis method is appropriate for dimension reduction. In this study, all features describe the travel behavior of vehicles from time or space perspectives. They can be roughly divided into three aspects: travel regularity at HFTTP, travel regularity at non-HFTTP, and OD regularity, so it is considered that three common factors can be extracted from nine observed features.

Since the entire dataset is too large, applying the factor analysis method and cluster analysis successively will consume much time and resources. For more complex algorithms, such as the ISODATA algorithm (Memarsadeghi et al., 2007), maybe it is unavailable to get a result in an acceptable time in the entire dataset. Therefore, before performing the factor analysis method, a sub-dataset containing 4,623 vehicles (1% of the entire dataset) was randomly selected. In this way, subsequent factor analysis method and cluster analysis can be completed in a shorter time. In addition, more complex algorithms can be applied to this small sub-dataset in order to get more accurate results in an acceptable time. Finally, three random sampling experiments were conducted to verify the robustness of the results.

In this work, the factor analysis method is used to reduce the dimensions of the original nine features. First, the number of common factors is needed to be determined. There are a number of different methods for determining the number of factors, but none of the methods to determine the number of factors is considered to be the best. And therefore the decision must be based on the preponderance of evidence from several perspectives on the issue (Brown, 2009). Kaiser criterion, Scree plot and Variance explained criteria (Warne and Larsen, 2014) are used to determine the number of factors. The eigenvalues of each component and the Scree plot are given in Fig. 2. As shown in Fig. 2, the eigenvalues of the first two factors are>1, and the third one is 0.91, which is close to 1. Therefore, the number of factors can be selected as 2 or 3. The percent of variance accounted for by a two-factor solution is 57.40%, while the percent of variance accounted for by a three-factor solution is 77.57%. The Scree plot first sorts the components according to the value of the eigenvalues and then plots the components on the x-axis and the corresponding eigenvalues on the y-axis. When the eigenvalues decline slowly (the curve has an inflection point), the components before the infection point are used as factors. All in all, three factors are selected in this study.

The factor loading can show the relationship between the original observed features and the latent factors, helping to understand the actual meaning of the latent factors. The factor loading of the three factors is shown in Table 5:

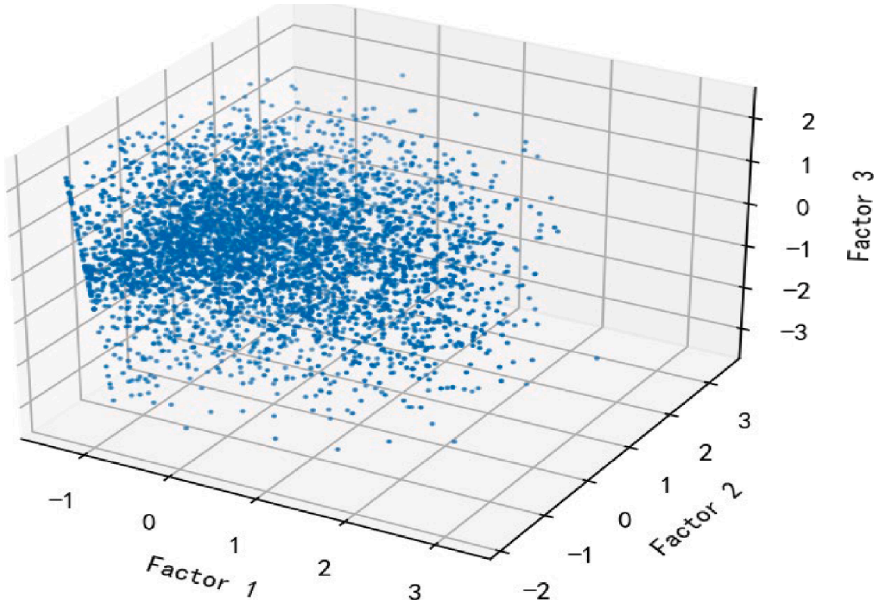


Fig. 3. Sample distribution after dimension reduction.

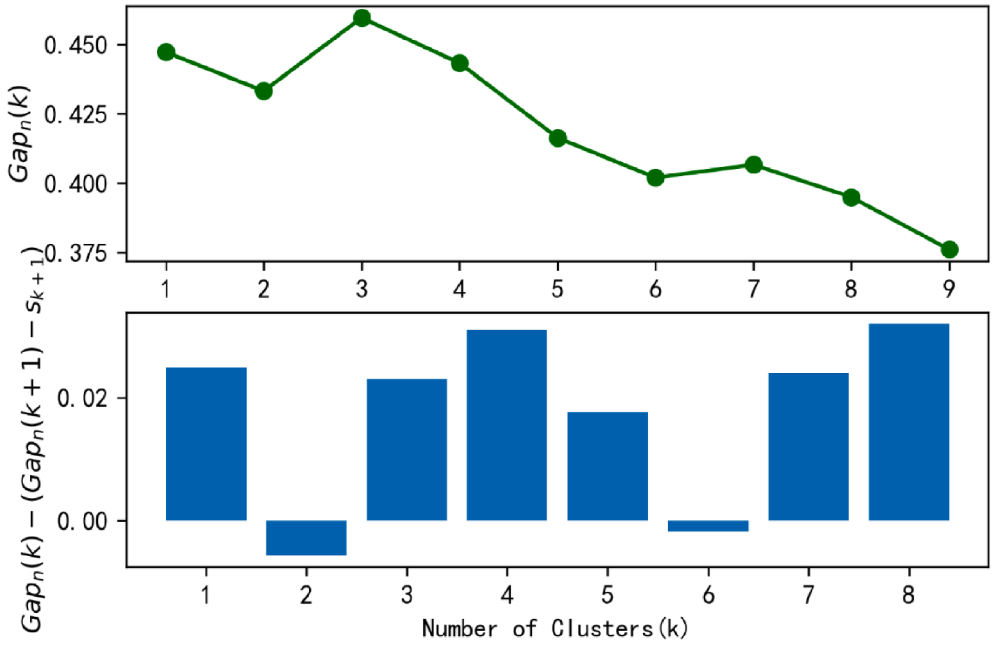


Fig.4. The changing trends of $Gap_n(k)$ and $Gap_n(k)$-($Gap_n(k+1)$-s_{k+1}) withcluster number.

As given in Table 5, Factor 1 mainly reflects the information on F_1, F_2, F_{O1}, F_{D1}, F_{O2}, and F_{D2} , and these six features mainly describe the stability of the vehicle travel behavior at HFTTP on the working days. Factor 2 mainly reflects the information on σ and M, and these two features mainly describe the stability of the vehicle travel behavior at non-HFTTP. Factor 3 mainly reflects the information on F_s, which shows the stability of the OD at HFTTP, which can also be seen in the definition of F_s given by Eq. (8).

Therefore, after dimension reduction using the factor analysis method, the original nine features are reduced to three independent factors. The sample distribution after dimension reduction is presented in Fig. 3.

4.5. Pattern recognition

The three selected factors are used as a manifestation of travel behavior. After that, we first need to determine how many categories

the vehicles in this sub-dataset can be divided into, which lays the foundation for the understanding of the data distribution and for subsequent cluster analysis. Tibshirani et al. (2001) proposed the gap statistic method, which can be used to estimate the optimal number of clusters.

In the gap statistic method, the statistic $Gap_n(k)$ is defined as Eq. (9).

$$Gap_n(k) = E_n^*\{logW_k\} - logW_k \tag{9}$$

where W_k denotes the pooled within-cluster sum of squares around the cluster means, E_n* denotes the expectation of a sample of size n from the reference distribution obtained by random sampling B times. The optimal cluster number k denotes the value that maximizes $Gap_n(k)$. That is to say, this method standardizes the graph of $\log(W_k)$ by comparing it with its expectation under an appropriate null reference distribution of the data. In addition, the smallest k that makes $Gap_n(k)$-($Gap_n(k+1)$-s_{k+1}) be greater than zero is considered as the optimal initial cluster number, where $s_{k+1} = \sqrt{\frac{1+B}{B}}sd(k+1)$, and $sd(k+1)$ denotes the standard deviation of $logW_{k+1}$ of the reference distribution obtained by random sampling B times. In this study, the gap statistic method (Tibshirani et al., 2001) is employed to determine the optimal number of classifications. After reducing the dimensions to three factors, the changes in $Gap_n(k)$ and ($Gap_n(k)$-($Gap_n(k+1)$-s_{k+1})) with the number of clusters k are obtained, and they are shown in Fig. 4.

In Fig. 4, it can be seen that when $k = 3$, $Gap_n(k)$ has the maximum value, while when $k = 1$, $Gap_n(k)$-($Gap_n(k+1)$-s_{k+1}) has the minimum value that is larger than zero. The reason for this is that the distribution of the sub-dataset is relatively tight. Considering the actual situation, the sub-dataset should not be divided into only one category. Chang et al. (2018) used the LPR data of Shanghai to classify non-local vehicles on the road network of Shanghai into five categories: commuting vehicles, vehicles used with low intensity on workdays and non-workdays, vehicles used with high intensity on both workdays and non-workdays, vehicles used with higher intensity on workdays, and vehicles used with higher intensity on non-workdays. Chen et al. (2017) used the LPR data of Shenzhen to classify vehicles on the road network on working days into six categories, including the commuting vehicles and some other occasional leisure travel vehicles. In this study, the focus is mainly on commuting vehicles and nine features mentioned above. The authors consider that vehicles on the road network can be divided into at least three categories. The first category represents vehicles with a strong regularity of travel behavior during both HFTTP and non-HFTTP and stable OD. This category includes commuting vehicles that are the focus of this work. The second category represents vehicles that are detected only on a few days in the month, and their travel regularity is poor. This type of vehicle mainly includes non-local vehicles (the license plate is non-Zhe A) or local vehicles traveling in non-core areas, such as Yuhang District, Xiaoshan District, and Fuyang District. The third category mainly includes vehicles having many trips on working days, but the regularity of trips is not very strong. These vehicles mainly include household vehicles for business, corporate vehicles for business, and many others.

As shown in Fig. 4, $Gap_n(k)$ has the maximum value when $k = 3$, and $Gap_n(k)$ decreases when $k > 3$, so in this work, the optimal initial cluster number k is set to three.

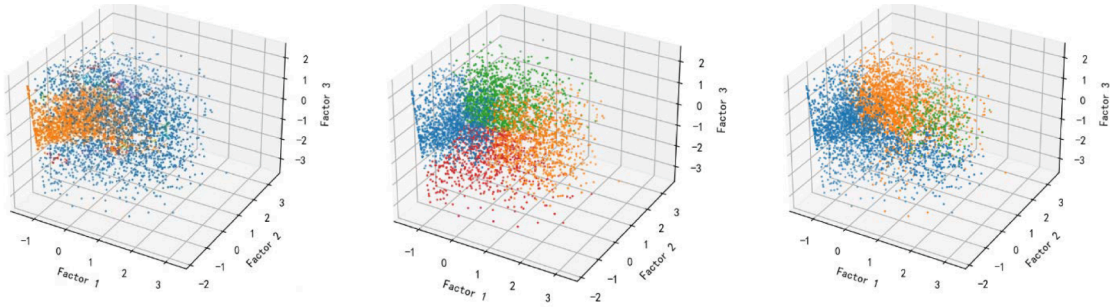
After obtaining the optimal number of initial clusters, three methods, the Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) (Ester et al., 1996), the Iterative Selforganizing Data Analysis Techniques Algorithm (ISODATA) (Ball and Hall, 1965), the fast search and find of density peaks clustering (Rodriguez and Laio, 2014) are used to perform cluster analysis on the dataset.

The DBSCAN algorithm represents a density-based clustering algorithm. This method does not need to determine the initial number of clusters; it can automatically determine the number of clusters based on the sample density. In addition, it can automatically delete outliers. The DBSCAN algorithm divides all sub-dataset points into three categories. The first category includes the core points. A point is regarded as a core point if more than *minPts* points are within distance *eps* from the core point. Also, the point is reachable from the core point if the distance between them is smaller than *eps*. The second category includes density-reachable points. A point is regarded as a density-reachable point if there is a path p_1,p_2,⋯, p_n, and where the distance between p_{i+1} and p_i is smaller than *eps*. The third category includes outliers. A point is regarded as an outlier if it cannot be reachable from any other point. Based on these three categories of points, the dataset can be automatically divided into different clusters.

The ISODATA algorithm represents an unsupervised classification algorithm, and it can automatically split the clusters if the number in the clusters is too large, while it can merge the clusters if the number in the clusters is too small. Additionally, human assistance in adjusting the algorithm parameters can be performed in the process of the ISODATA algorithm, which means we can modify the parameters of the ISODATA algorithm to help it perform better according to the results of each iteration. However, the ISODATA algorithm is usually run slowly because of its complexity. So if a dataset is very large, the ISODATA algorithm will not terminate in an acceptable period of time. Fortunately, we have previously sampled the entire dataset, so the ISODATA algorithm can be successfully applied to the sub-dataset. From this point of view, it is necessary to sample the entire dataset.

The fast search and find of density peaks clustering algorithm represents a new clustering method proposed by Alex Rodriguez and Alessandro Laio in 2014. It is different from the classic clustering method, and its main innovation lies in its characterization of the clustering center; namely, the density of the clustering center is large, and the density of its surrounding neighbors is less than that of the clustering center. The distance between clustering centers and other denser data points is relatively large. In other words, the clustering centers are pretty dense and far apart from each other.

Using the above three different types of clustering algorithm to cluster the dataset, expect to identify commuting pattern vehicles. The Silhouette coefficient (Rousseeuw, 1987) is used to evaluate the clustering results. It is necessary to calculate the mean intra-cluster distance ($a(i)$), and the mean nearest-cluster distance ($b(i)$) of point p_i and then the Silhouette coefficient of point p_i can be calculated by Eq. (10).



Silhouette coefficient: -0.2470

a) Result of the DBSCAN algorithm (The blue points denote the outliers)

Silhouette coefficient: 0.2854

b) Result of the ISODATA algorithm

Silhouette coefficient: 0.1757

c) Result of the fast search and find of density peaks clustering algorithm

Fig. 5. Results of the clustering algorithm.

Table 6
The descriptive statistics of the features of commuting vehicles obtained by three clustering methods.

The descriptive statistics of the features of commuting vehicles obtained by DBSCAN

	F_1	F_2	F_{O1}	F_{D1}	F_{O2}	F_{D2}	F_s	σ	M
count	2056	2056	2056	2056	2056	2056	2056	2056	2056
mean	0.143	0.488	0.143	0.129	0.124	0.136	−1.183	2.219	3.710
std	0.142	0.270	0.150	0.137	0.121	0.128	0.458	1.600	1.990
min	0.048	0.048	0.048	0.048	0.048	0.048	−2.464	0.000	0.000
25%	0.048	0.238	0.048	0.048	0.048	0.048	−1.440	0.957	2.400
50%	0.095	0.476	0.095	0.048	0.095	0.095	−1.142	2.163	3.600
75%	0.190	0.714	0.143	0.143	0.143	0.143	−1.000	3.355	5.000
max	0.905	1.000	0.952	0.905	0.857	0.810	0.990	7.204	11.000

The descriptive statistics of the features of commuting vehicles obtained by ISODATA

	F_1	F_2	F_{O1}	F_{D1}	F_{O2}	F_{D2}	F_s	σ	M
count	1012	1012	1012	1012	1012	1012	1012	1012	1012
mean	0.631	0.882	0.593	0.576	0.512	0.517	−0.375	2.084	3.470
std	0.203	0.109	0.195	0.194	0.191	0.192	0.855	1.461	1.724
min	0.048	0.476	0.048	0.048	0.048	0.048	−2.378	0.000	0.000
25%	0.476	0.810	0.476	0.429	0.381	0.381	−1.049	1.000	2.327
50%	0.667	0.905	0.619	0.571	0.524	0.524	−0.446	2.049	3.333
75%	0.762	1.000	0.714	0.714	0.667	0.667	0.411	2.968	4.638
max	1.000	1.000	1.000	1.000	1.000	1.000	0.996	8.083	10.750

The descriptive statistics of the features of commuting vehicles obtained by fast search and find of density peaks clustering algorithm

	F_1	F_2	F_{O1}	F_{D1}	F_{O2}	F_{D2}	F_s	σ	M
count	318	318	318	318	318	318	318	318	318
mean	0.735	0.911	0.664	0.662	0.576	0.591	0.492	2.110	3.645
std	0.149	0.087	0.162	0.159	0.171	0.171	0.377	1.465	1.703
min	0.286	0.619	0.190	0.238	0.143	0.095	−0.593	0.000	0.000
25%	0.619	0.857	0.571	0.571	0.476	0.476	0.197	1.155	2.617
50%	0.762	0.952	0.667	0.667	0.571	0.619	0.562	2.082	3.563
75%	0.857	1.000	0.762	0.762	0.714	0.714	0.833	3.021	4.771
max	1.000	1.000	1.000	1.000	1.000	0.952	0.989	6.626	8.500

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{10}$$

Finally, the Silhouette coefficients of all samples are averaged to get the Silhouette coefficient of the dataset using the clustering algorithm. The range of the Silhouette coefficient is [-1, 1]; the closer the value of the Silhouette coefficient is to one, the better the clustering effect is. If the value is less than zero, it means that the clustering effect is not good, and many points are classified incorrectly.

The results of the three clustering methods and the Silhouette coefficients are shown in Fig. 5.

Table 7
The average of features of the ISODATA algorithm results.

	F_1	F_2	F_{O1}	F_{D1}	F_{O2}	F_{D2}	F_s	σ	M	Number
Cluster 1	0.111	0.374	0.119	0.114	0.108	0.113	−0.779	1.109	2.657	1458
Cluster 2	0.631	0.882	0.593	0.576	0.512	0.517	−0.375	2.084	3.470	1012
Cluster 3	0.222	0.641	0.188	0.160	0.143	0.167	−0.809	4.530	6.431	1335
Cluster 4	0.185	0.622	0.203	0.178	0.169	0.181	−2.525	2.462	3.802	818

The descriptive statistics of the features of commuting vehicles obtained by three clustering methods is shown in Table 6. A total of 34 clusters are obtained, and 47.67% of the sub-dataset points are identified as outliers in the results of DBSCAN algorithm. the cluster with the most sample points, which is the cluster formed by the red points in Fig. 5(a), is analyzed. Its descriptive statistics are given in Table 6. the commuting pattern of this cluster is not obvious, and the recognition effect of the commuting vehicles is not good, which is consistent with the result of the Silhouette coefficient.

The descriptive statistics of commuting vehicles obtained by ISODATA algorithm are shown in Table 6, and the average value and sample number of each cluster are shown in Table 7. The average values of F_1, F_2, F_{O1}, F_{D1}, F_{O2}, F_{D2}, σ, and M of the first cluster are all smaller than the average values of those of the entire dataset, while the value of F_s is larger than the average value of that of the entire dataset. Therefore, these vehicles have fewer travel days and lower detection frequency. This type of vehicle mainly includes transit vehicles and vehicles coming to Hangzhou for tourism, as well as non-local vehicles whose license plate numbers are non-Zhe A or local vehicles in the non-core areas, such as Yuhang District, Xiaoshan District, and Fuyang District. Transit vehicles are defined as vehicles passing through the core area of Hangzhou, but their destination is not in the core area. The values of F_1, F_2, F_{O1}, F_{D1}, F_{O2}, F_{D2}, and F_s of the second cluster are all greater than the average values of those of the entire dataset, while the values of σ and M are smaller than the average values of those of the entire dataset. Therefore, this kind of vehicle has stable daily travel behavior and a strong travel regularity; thus, they are mainly composed of commuting vehicles. The average values of F_1 and F_2 of the third cluster are approximately the same as the average values of F_1 and F_2 of the entire dataset; average values of F_{O1}, F_{D1}, F_{O2}, and F_{D2} are smaller than the average values of those of the entire dataset; average values of F_s are greater than the average values of that of the entire dataset; average values of σ and M are slightly larger than the average values of those of the entire dataset. In conclusion, this type of vehicle travel in the road network morn than half of the working days, but the travel regularity is not high. According to the travel behavior, it is assumed that these vehicles are mainly used for going to work or getting off work, but their travel time is flexible, and they are not traditional commuting vehicles. The values of F_1, F_{O1}, F_{D1}, F_{O2}, F_{D2}, F_s, σ, and M of the fourth cluster are all smaller than the average values of those of the entire dataset, while the value of F_2 is larger than the average value of that of the entire dataset. This kind of vehicle mainly includes household business vehicles and corporate business vehicles, which are not commonly used for commuting but for entertainment.

In this work, the focus is mainly on commuting vehicles. According to the travel characteristics reflected by various features, it can be concluded that vehicles in the cluster, including the commuting vehicles, travel on 88.2% of the working days on average, and on 63.1% of the working days, they travel at HFTTP. Also, the regularity of OD is strong. There are fewer trips at non-HFTTP, and the regularity of travel behavior at non-HFTTP is strong, which is consistent with the definition of commuting vehicles.

The commuting vehicles identified by fast search and find of density peaks clustering algorithm travel on 91.1% of working days, and on 73.5% of the working days, they travel at HFTTP. The regularity of the OD is strong. There are fewer trips at non-HFTTP, and the regularity of the travel behavior at non-HFTTP is strong, which is consistent with the definition of commuting vehicles.

By comparing the results obtained by the three clustering methods, it can be found that the DBSCAN algorithm is not applicable to this case because it is impossible to use the DBSCAN algorithm to identify the commuting vehicles. From the perspective of data distribution, this is because the distribution of the data is extremely uneven, and even the density of sample points in the same cluster varies greatly. Considering the behavior of real-life vehicles, this is because the travel characteristics of vehicles, even for vehicles of the same type, can differ significantly. Both the ISODATA algorithm and the fast search and find of density peaks clustering algorithm can successfully identify commuting vehicles, and according to the Silhouette coefficient, the results obtained by the ISODATA algorithm have higher credibility. Among these three methods, the optimal method is the ISODATA algorithm, so the ISODATA algorithm is chosen as the final commuting pattern recognition method.

5. Commuting rule extraction and commuting vehicles identification

Although the ISODATA algorithm can successfully identify commuting vehicles, the identified commuting vehicles belong only to a small sub-dataset of only 1% of the total data. The commuting vehicles in the entire data have not been identified yet. Due to the complexity of the ISODATA algorithm, it is difficult to use it for commuting pattern recognition on the entire dataset containing travel behavior information on 462,288 vehicles, so it is still needed to study how to extract the commuting rule form such a large dataset. The extracted commuting rule can be easily implemented in the traditional relational databases (Ma et al., 2017). Therefore, the commuting rule can be used to identify the commuting vehicles in the entire dataset in a relational database. In addition, the extraction of a commuting rule can also help to deepen the understanding of commuting travel behavior.

After using the ISODATA algorithm to classify a small sub-dataset (1% of the entire dataset, which included the information on 4,623 vehicles), each sample in each cluster is labeled separately. Since the ISODATA algorithm divides the samples into four clusters and the focus of this work is on commuting vehicles, these vehicles are labeled as 1, and all other vehicles are labeled as 0. Then the

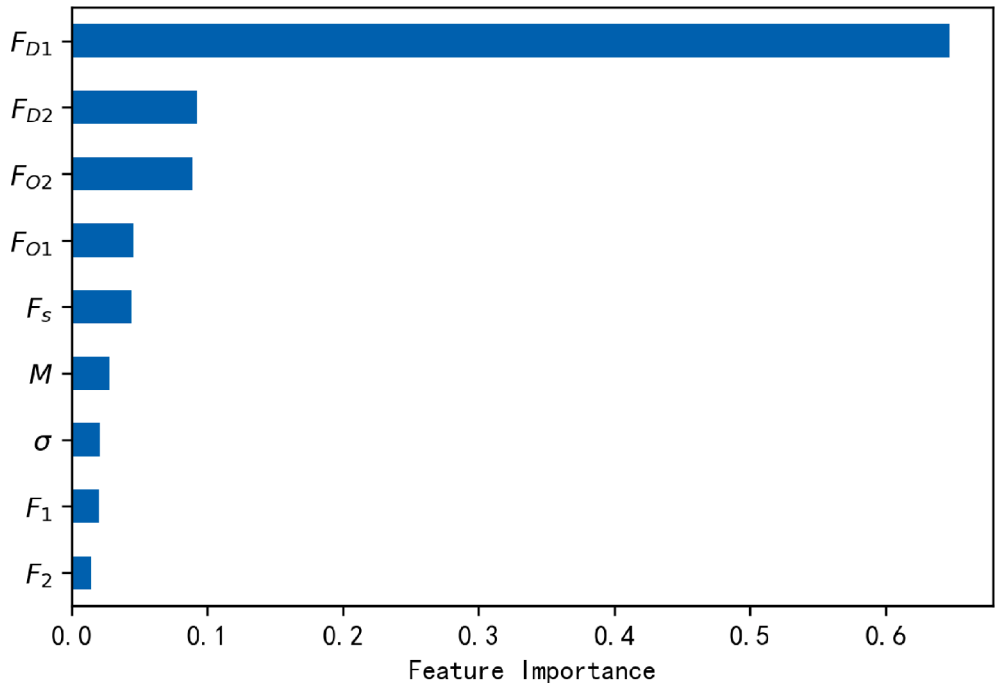


Fig. 6. Feature importance ranking by the decision tree algorithm.

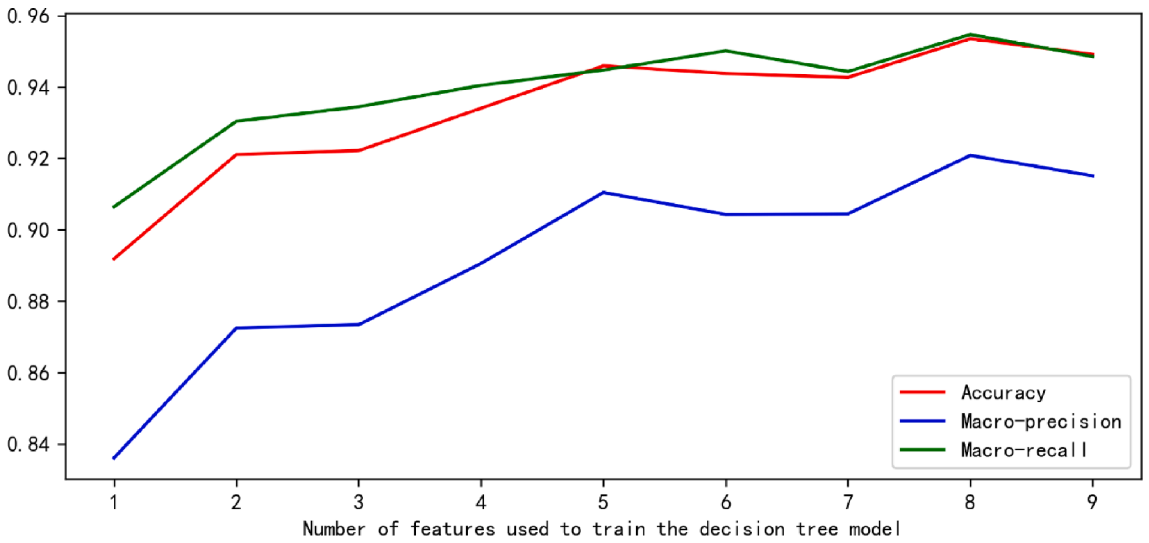


Fig. 7. Performance of the decision tree algorithm trained with different numbers of features.

decision tree algorithm (Loh, 2011) is employed to extract the commuting rule.

The decision tree algorithm constructs a tree that describes the relationships between the features and target values. If the target values are discrete, then the model is called the classification tree. In such a tree structure, leaves are the labels of samples. On the other hand, if the target values are continuous, then the model is called the regression tree. The decision tree is simple to understand and implement, and it can perform well with a large dataset. In addition, it is similar to a human's decision-making process. Hence, the commuting pattern can be well understood using the decision tree algorithm.

The decision tree algorithm is used to extract the commuting rules. However, if all nine features are used to build a decision tree model, the tree model will be very large and will not be conducive to the extraction of commuting rules. Through the correlation analysis, it is found that there is a certain correlation between these features. Thus, it is not necessary to use all nine features to build a decision tree. Therefore, the decision tree model is first used for feature importance ranking. In the decision tree constructing process, the decision tree algorithm sequentially selects the features that can make the best classification effect, and divide the feature, building

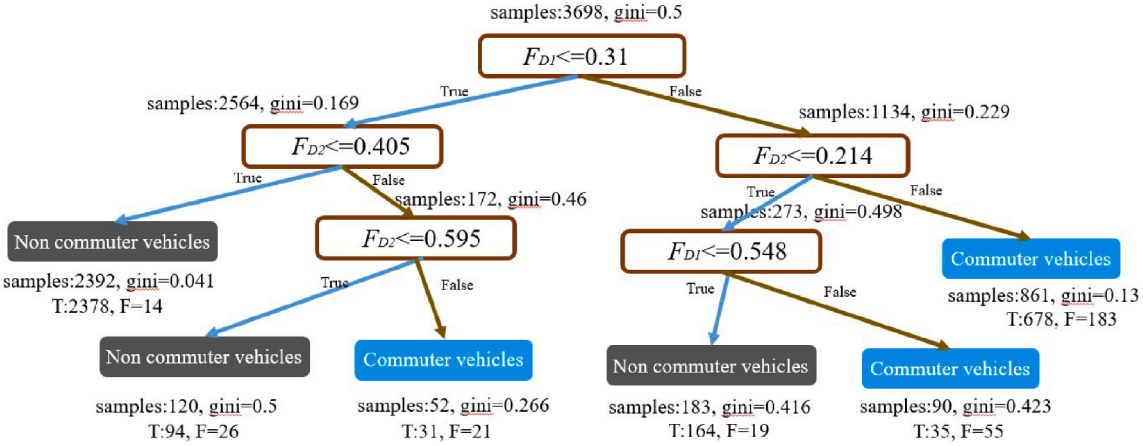


Fig. 8. Result of the decision tree model.

a decision tree; thus, the decision tree model denotes a classic feature selection algorithm. In this work, the decision tree model is used to rank the feature importance of nine features. The ranking results are shown in Fig. 6. The greater the feature importance is, the more important the feature is.

Based on the obtained feature importance values, the most important 1–9 features are selected to build the decision tree, successively. The effect of the trained decision tree model for a different number of features is shown in Fig. 7. In this work, the dataset is divided into training and test sets according to the ratio of 4:1. When a different number of features in the training set are used to train the decision tree model, the grid search method is used to optimize the decision tree model parameters. When the grid search method is used to search for the optimal parameters, the model evaluation is conducted by the five-fold cross-validation, and the average accuracy is used as the evaluation index of the model performance. The grid search method separately searches the three parameters of the decision tree as follows: *max_depth*: [2, 3, 4, 5, 6, 7, 8, 9, 10], *min_samples_leaf*: [2, 5, 10, 20, 40], and *min_samples_split*: [20, 50, 80, 100], each combination of the above three parameters will be used to search for the optimal parameters of the decision tree model, where *max_depth* denotes the maximum depth of the decision tree model, *min_samples_leaf* denotes the minimum number of the samples in the decision tree's node, and *min_samples_split* denotes the minimum number of samples that is needed to split the decision tree's node. The optimal decision tree model obtained by training with different numbers of features is evaluated using three evaluation indicators: accuracy, macro-precision, and macro-recall, which are respectively calculated as Eqs. (11), (12) and (13).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Macro - precision = \frac{1}{C}\sum_{i}\frac{TP_i}{TP_i + FP_i} \tag{12}$$

$$Macro - recall = \frac{1}{C}\sum_{i}\frac{TP_i}{TP_i + FN_i} \tag{13}$$

where$i \in C$, representing all possible classes; C denotes the total number of classes, TN_i stands for true negative in class i, TP_i stands for true positive in class i, FP_i represents false positive in class i, and lastly, FN_i denotes false negative in class i.

As shown in Fig. 7, as the number of features increases, the prediction effect of the model gradually increases. When the number of features is equal to five, the model performance is almost the best, and the model performance improvement is not obvious by further increasing the number of features. However, the optimal model constructed with five features is still very large, and the rules are not intuitive. In Fig. 7, it can be seen that when two features are used, the model performance is quite good, and as the number of features increases, the accuracy and macro-recall basically remain unchanged, while the macro-precision shows a certain degree of improvement. By weighing the effectiveness, simplicity, and interpretability of the model, the two most important features are chosen to build a decision tree model. The results obtained by this model in Fig. 8.

Each rectangle in Fig. 8 represents a node; *gini* in the node indicates the gini impurity of the node corresponding to the decision tree. The gini impurity is a measure of how often the element that is randomly selected from the dataset is labeled wrongly if it was randomly labeled based on the distribution of labels in the dataset. The minimum value of the gini impurity is zero when all the samples in the node belong to the same class. The more uneven the labels in a node are, the larger the gini impurity will be. Further, *Samples* in a node denotes the number of sample points contained in this node; *T* represents the number of samples that are correctly classified; *F* denotes the number of samples that are misclassified. In addition, it should be pointed out that in the decision tree model training, due to the imbalance of samples in various classes, the ratio of non-commuting vehicles to commuting vehicles is 3.57, so different weights are assigned to the samples in the two classes. Specifically, the weights are adjusted inversely proportional to the class frequencies. Therefore, in the final decision tree model, the number of correct samples is less than the number of wrong samples at

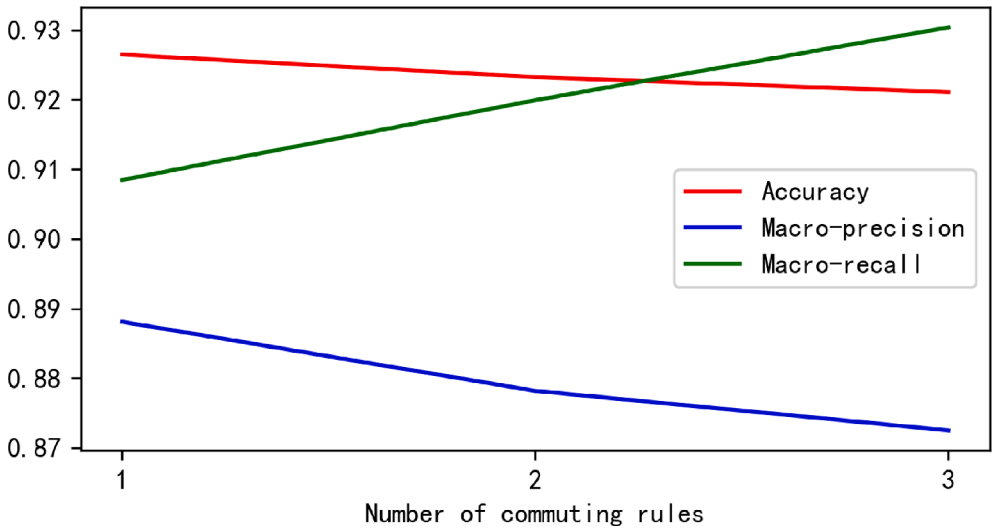


Fig. 9. The changing trends of the accuracy, macro-precision, and macro-recall with the number of commuting rules.

Table 8
The descriptive statistics of the commuting vehicles.

	F_1	F_2	F_{O1}	F_{D1}	F_{O2}	F_{D2}	F_s	σ	M
count	114,439	114,439	114,439	114,439	114,439	114,439	114,439	114,439	114,439
mean	0.607	0.863	0.562	0.551	0.491	0.515	−0.434	2.335	3.875
std	0.203	0.129	0.196	0.185	0.194	0.175	1.050	1.744	2.121
min	0.048	0.333	0.048	0.048	0.048	0.238	−4.297	0.000	0.000
25%	0.476	0.810	0.429	0.429	0.333	0.381	−1.110	1.000	2.400
50%	0.619	0.905	0.571	0.524	0.476	0.524	−0.347	2.137	3.667
75%	0.762	0.952	0.714	0.667	0.619	0.619	0.508	3.445	5.121
max	1.000	1.000	1.000	1.000	1.000	1.000	1.000	8.921	11.182

some leaf nodes.

On the test set, the accuracy, macro-precision, and macro-recall of the decision tree model are 92.1%, 87.2%, and 93.0%, respectively.

The commuting rules can be obtained from the decision tree graph as follows:

$$Rule1: F_{D1} > 0.31 \& F_{D2} > 0.214 (gini = 0.13, samples = 861)$$
$$Rule2: F_{D2} > 0.595 \& F_{D1} \leqslant 0.31 (gini = 0.266, samples = 52)$$
$$Rule3: F_{D1} > 0.548 \& F_{D2} \leqslant 0.214 (gini = 0.423, samples = 90)$$

First, the rules are ranked according to their importance based on the gini impurities, then 1, 2, or 3 rules are selected successively according to the importance and used to identify commuting vehicles in the test set, and the three indicators (accuracy, macro-precision, and macro-recall) are calculated to evaluate the commuting pattern recognition performance of the model. The changing trends of the accuracy, macro-precision, and macro-recall change with the number of commuting rules can be seen in Fig. 9.

According to the results presented in Fig. 9, with the increase in the number of commuting rules, the accuracy and macro-precision decrease continuously, from 92.6% to 92.1% and from 88.8% to 87.2%, respectively, while the macro-recall increases from 90.8% to 93.0%. In order to make the indicators of the model performance as high as possible while keeping the rules as simple as possible, the commuting pattern recognition rules are defined as follows:

$$Rule1: F_{D1} > 0.31 \& F_{D2} > 0.214 (gini = 0.13, samples = 861)$$
$$Rule2: F_{D2} > 0.595 \& F_{D1} \leqslant 0.31 (gini = 0.266, samples = 52)$$

The rule uses only two features, F_{D1} and F_{D2}, to identify commuting vehicles, and in this case, the accuracy, macro-precision, and macro-recall on the test set reach the values of 92.3%, 87.8%, and 92.0%, respectively. However, it is surprising that using only the OD stability coefficient of a commuting trip commuting vehicles can accurately distinguish on the road network. This can be explained by the physical meaning of these two features; namely, when a vehicle destination during the HFTTP in the morning on a weekday is relatively fixed and the destination during the HFTTP in the evening is also relatively fixed (i.e., $F_{D1} > 0.31$ and $F_{D2} > 0.214$), a vehicle can be identified as a commuting vehicle. Further, when the destination during the HFTTP in the evening is pretty fixed, that is, $F_{D2} > 0.595$, it is not even needed to use other information on a vehicle to infer that the vehicle represents a commuting vehicle. This also shows that the regularity of traveling during the HFTTP of commuting vehicles is not characteristic for other types of vehicles.

Table 9
Comparison of the average value of features of the commuting vehicles and the average value of features of all other vehicles.

	F_1	F_2	F_{O1}	F_{D1}	F_{O2}	F_{D2}	F_s	σ	M
Commuting vehicles	0.607	0.863	0.562	0.551	0.491	0.515	−0.434	2.335	3.875
Other vehicles	0.158	0.517	0.159	0.137	0.124	0.132	−1.146	2.640	4.220
Relative change (%)	−74.0	−40.1	−71.7	−75.1	−74.7	−74.4	−164.1	+13.1	+8.9

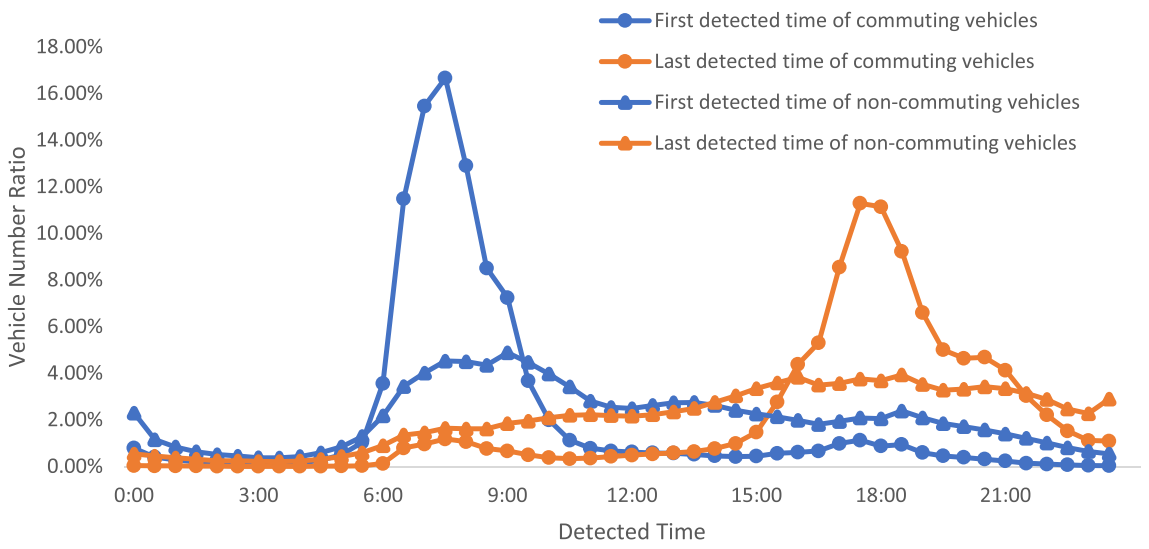


Fig. 10. Average first and last detected time distributions of all workdays.

After getting the identification rules of commuting vehicles, a total of 114,439 of commuting vehicles in Hangzhou's core area are accurately identified, including 111,945 Hangzhou local license plate vehicles (the license plate is Zhe A), and 2,494 non-local license plate vehicles (license plate is non-Zhe A).

6. Discussion

6.1. Analysis of the differences between the travel characteristics of commuting vehicles and those of all other vehicles

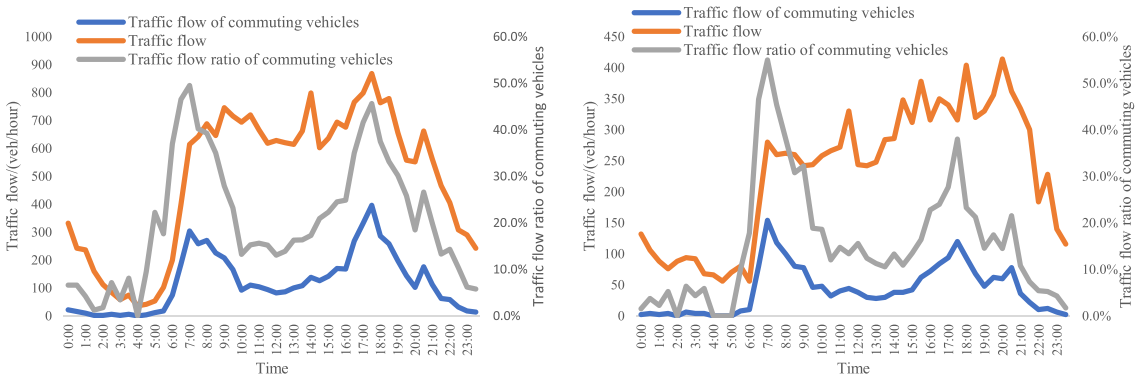
A total of 114,439 commuting vehicles in the core area of Hangzhou have been identified. The descriptive statistics of each feature of these vehicles are shown in Table 8.

The differences between the travel characteristics of commuting vehicles and those of all other vehicles are analyzed. The average value of each feature and the relative change is given in Table 9.

As shown in Table 9, F1, F2, F_{O1}, F_{D1}, F_{O2}, F_{D2}, and F_s of non-commuting vehicles are significantly smaller than those of the commuting vehicles, which means that in the traffic network, commuting vehicles travel more days and have stronger spatiotemporal regularity than non-commuting vehicles, and thus have a greater impact on the traffic conditions of the road network, which is analyzed more detailed in the following. It should be noted that the average value in Table 9 represent the information on the non-commuting vehicles whose feature F_s is not null. The values of σ and M of commuting vehicles are slightly lower than those of non-commuting vehicles, indicating that commuting vehicles travel less during the non-HFTTP, but have stronger regularity.

6.2. Analysis on travel regularity of commuting vehicles

The travel behavior of commuting vehicles was further analyzed. During 21 working days in June 2016 in Hangzhou's Core area, a total of 3,509,489 vehicles were recorded and saved as the LPR data. The LPR data included a total of 462,288 vehicles whose feature F_s was not NULL, and a total of 114,439 vehicles were recognized as commuting vehicles, accounting for 3.26% of all the recorded vehicles. The proportion of commuting vehicles was very low, because there were a large number of transit vehicles. Transit vehicles are defined as vehicles passing through the core area of Hangzhou, but their destination is not in the core area. In this study, many vehicles in the suburbs of Hangzhou entering into core area were transit vehicles. In addition, as the capital city of Zhejiang Province and a famous tourist city, Hangzhou had a large number of vehicles coming for business and tourism. Transit vehicles were often detected only a few times a day, most of them were 1–2 times. LPR data contained very little information about such vehicles, which usually failed to calculate the feature F_s successfully. Therefore, it could be found that LPR data included only a total of 462,288 vehicles which feature F_s was not null, and the F_s of 3,047,201 vehicles was null. In the subsequent analysis, in order to eliminate the



a) The intersection of Yugu road and Zheda Road b) The intersection of Huancheng West Road and Fengqi Road

Fig. 11. The changes of Q, Q_c and r on June 15, 2016.

interference of transit vehicles, the total vehicle refers to all vehicles whose F_s is not empty. During 21 working days in June 2016, there were a total of 77,240,905 records of LPR data, among which there were 21,225,036 records of commuting vehicles. The detected records of the commuting vehicles accounted for 27.48% of all the records. The average daily number of commuting vehicles and the daily detection frequency of commuting vehicles during the HFTTP on 21 working days in June 2016 were calculated. The average daily number of commuting vehicles traveling during the HFTTP was 104,874, and the average daily detection frequency was 772,806. On 21 working days in June 2016, the average daily number of vehicles traveling during the HFTTP was 304,127, and the average daily detection frequency was 2,249,097. During the HFTTP, the average daily commuting vehicles accounted for 34.48% of all the vehicles, and the detection frequency of commuting vehicles accounted for 34.36% of all the vehicles.

The first and last detected times of commuting and non-commuting vehicles were also analyzed. As previously mentioned, each interval lasted 30 min, so a day was divided into 48 intervals. We analyzed the average first and last detected time distributions of all workdays, and the results are shown in Fig. 10.

As shown in Fig. 10, the first time the commuting vehicles were detected was concentrated in the interval of 06:30 AM – 10:30 AM. And 78.0% of commuting vehicles were detected for the first time during this time period on average. The average last detected time distribution was relatively dispersed. On average, 77.8% of commuting vehicles were last detected between 03:30 PM and 09:30 PM on each workday. However, for non commuting vehicles, the phenomenon that start or end the travel during the morning and evening peak hours is not obvious.

We further analyze the travel behavior of commuting pattern vehicles within a day, and define the traffic flow(Q), traffic flow of commuting vehicles (Q_c), and traffic flow ratio of commuting vehicles (r) as Eqs. (14), (15) and (16).

$$Q = \frac{freq_i}{\Delta T} \tag{14}$$

$$Q_c = \frac{freq_{ci}}{\Delta T} \tag{15}$$

$$r = \frac{Q_c}{Q} \tag{16}$$

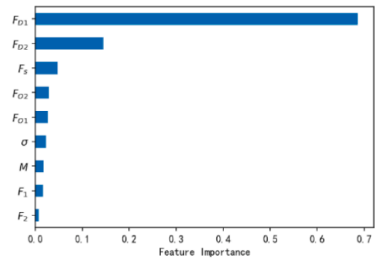
Where $freq_i$ is the total number of detected records of LPR detector i in ΔT, and $freq_{ci}$ is the total number of detected records of commuting pattern vehicles of LPR detector i in ΔT.

Two LPR detectors with large traffic flow are selected to show the changes of Q, Q_c and r on June 15, 2016, and ΔT is set to 30 min. Two LPR detectors are located at the intersection of Yugu road and Zheda Road and the intersection of Huancheng West Road and Fengqi Road, respectively. The results are shown in Fig. 11.

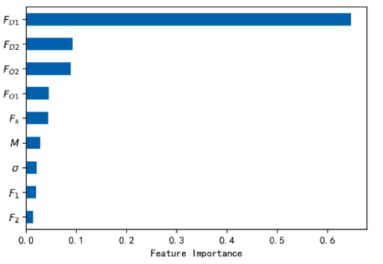
It can be seen that the traffic flow at the two LPR detectors is large all day, and most commuting pattern vehicles travel during the morning and evening peak hours. The traffic flow ratio of commuting vehicles in the morning peak is up to about 50%, and the traffic flow ratio of commuting vehicles in the evening peak is slightly lower that that in the morning peak.

6.3. *The influence of different HFTTP on commuting rule extraction and commuting vehicle recognition*

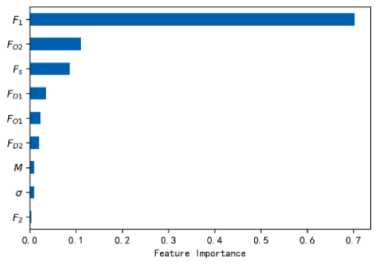
As mentioned above, different threshold will get different HFTTP, and the determination of the threshold value needs to be determined according to the local conditions in different regions. Table 3 gives the HFTTP when α is 3%, 3.5% and 4%. The above studies are all based on $\alpha = 3.5\%$, next, we will discuss the influence of different HFTTP on commuting rule extraction and commuter vehicle recognition.



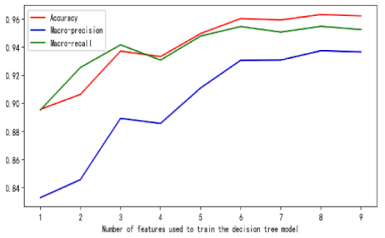
(a) Feature importance ranking by the decision tree algorithm with $\alpha = 3\%$



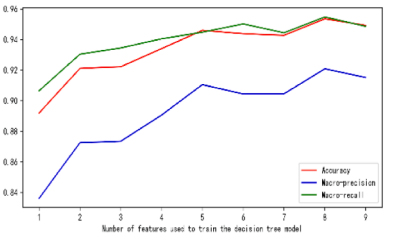
(b) Feature importance ranking by the decision tree algorithm with $\alpha = 3.5\%$



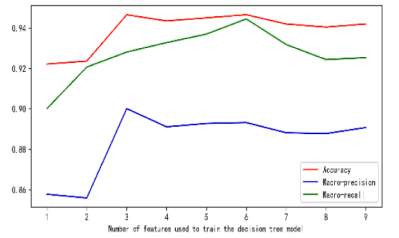
(c) Feature importance ranking by the decision tree algorithm with $\alpha = 4\%$



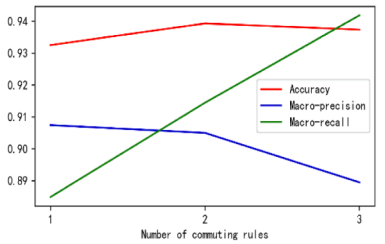
(d) Performance of the decision tree algorithm trained with different numbers of features with $\alpha = 3\%$



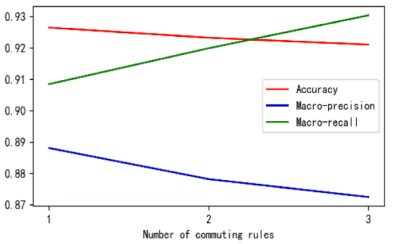
(e) Performance of the decision tree algorithm trained with different numbers of features with $\alpha = 3.5\%$



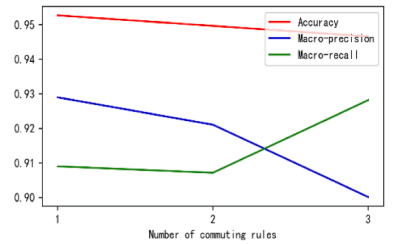
(f) Performance of the decision tree algorithm trained with different numbers of features with $\alpha = 4\%$



(g) The changing trends of the accuracy, macro-precision, and macro-recall with the number of commuting rules with $\alpha = 3\%$



(h) The changing trends of the accuracy, macro-precision, and macro-recall with the number of commuting rules with $\alpha = 3.5\%$



(i) The changing trends of the accuracy, macro-precision, and macro-recall with the number of commuting rules with $\alpha = 4\%$

Fig. 12. Extraction of commuting rules with different α values.

Table 10
Commuting rules with different α.

Commuting rules	*Gini*	*Samples*
The commuting rules when α is 3%		
$F_{D1} > 0.31\&F_{D2} > 0.214\&F_s > -2.689$	0.087	880
$F_{D1} > 0.452\&F_{D2} \leqslant 0.214\&F_s > -1.691$	0.302	89
$F_{D1} \leqslant 0.31\&F_{D2} > 0.405\&F_s > -1.579$	0.375	134
The commuting rules when α is 3.5%		
$F_{D1} > 0.31\&F_{D2} > 0.214$	0.13	861
$F_{D2} > 0.595\&F_{D1} \leqslant 0.31$	0.266	52
$F_{D1} > 0.548\&F_{D2} \leqslant 0.214$	0.423	90
The commuting rules when α is 4%		
$F_1 > 0.357\&F_{O2} > 0.214\&F_s > -2.089$	0.069	497
$F_1 > 0.357\&F_{O2} \leqslant 0.214\&F_s > 0.205$	0.256	26
$F_1 \leqslant 0.357\&F_{O2} > 0.262\&F_s > -1.009$	0.397	69

The commuting features are extracted based on the HFTTP that is obtained by taking $\alpha = 3\%$, 3.5% and 4%, then the ISODATA algorithm is used to cluster the sub-dataset to identify the commuting vehicles. The sub-dataset is further used to train the decision tree to obtain the importance ranking of each feature, as shown in (a), (b) and (c) of Fig. 12. Using different number of features to train the decision tree, the performance of the model in the three cases is as (d), (e) and (f) of Fig. 12. When the number of features is 3, the performance of the model taking the $\alpha = 4\%$ is basically optimal. Increasing the number of features is not helpful to improve the model performance. Therefore, F_1, F_{O2} and F_s are used to identify commuting vehicles when the α is 4%. When α is 3% and 3.5%, the changing trends of model performance with the number of features are very similar. Therefore, by weighing the effectiveness, simplicity, and interpretability of the model, the three most important features are chosen to build a decision tree model when taking α is 3%, and the two most important features are chosen to build a decision tree model when taking α is 3.5%. The commuting rules of decision tree model are shown in Table 10 when the α is 3%, 3.5% and 4%, respectively.

First, the rules are ranked according to their importance based on the gini impurities, then 1 to 3 rules are selected successively according to the importance and used to identify commuting vehicles in the test set when α is 3% and 4%, and the three indicators (accuracy, macro-precision, and macro-recall) are calculated to evaluate the commuting pattern recognition performance of the model in the two cases. The changing trends of the accuracy, macro-precision, and macro-recall change with the number of commuting rules can be seen in (g), (h) and (i) of Fig. 12. In order to make the indicators of the model performance as high as possible while keeping the rules as simple as possible, two rules are kept when α is 3% and 3.5%, while only one rule is kept when α is 4%, and the final commuting rules are shown in Table 11 when the α is 3%, 3.5% and 4%, respectively.

Table 11
Commuting rules with different α values.

Commuting rules	***Gini***	***Samples***
The commuting rules when α is 3%		
$F_{D1} > 0.31 \& F_{D2} > 0.214 \& F_s > -2.689$	0.087	880
$F_{D1} > 0.452 \& F_{D2} \leqslant 0.214 \& F_s > -1.691$	0.302	89
The commuting rules when α is 3.5%		
$F_{D1} > 0.31 \& F_{D2} > 0.214$	0.13	861
$F_{D2} > 0.595 \& F_{D1} \leqslant 0.31$	0.266	52
The commuting rules when α is 4%		
$F_1 > 0.357 \& F_{O2} > 0.214 \& F_s > -2.089$	0.069	497

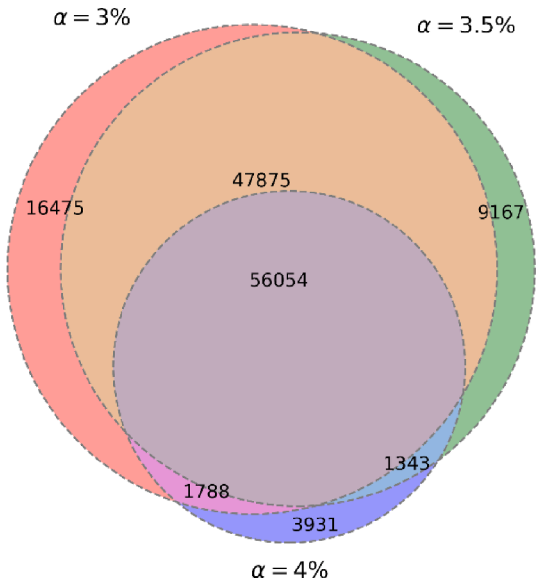


Fig. 13. The relationship between commuting vehicles identified by taking different α.

Table 12
The commuting rules obtained by three random sampling.

Commuting rules	*Gini*	*Samples*
First random sampling		
$F_{D1} > 0.31\&F_{D2} > 0.214$	0.13	861
$F_{D2} > 0.595\&F_{D1} \leqslant 0.31$	0.266	52
Second random sampling		
$F_{D1} > 0.262\&F_{O2} > 0.262$	0.088	867
$F_{D1} > 0.452\&F_{O2} \leqslant 0.262$	0.376	160
Third random sampling		
$F_{D1} > 0.262\&F_{D2} > 0.214$	0.101	945
$F_{D1} \leqslant 0.262\&F_{D2} > 0.643$	0.171	26

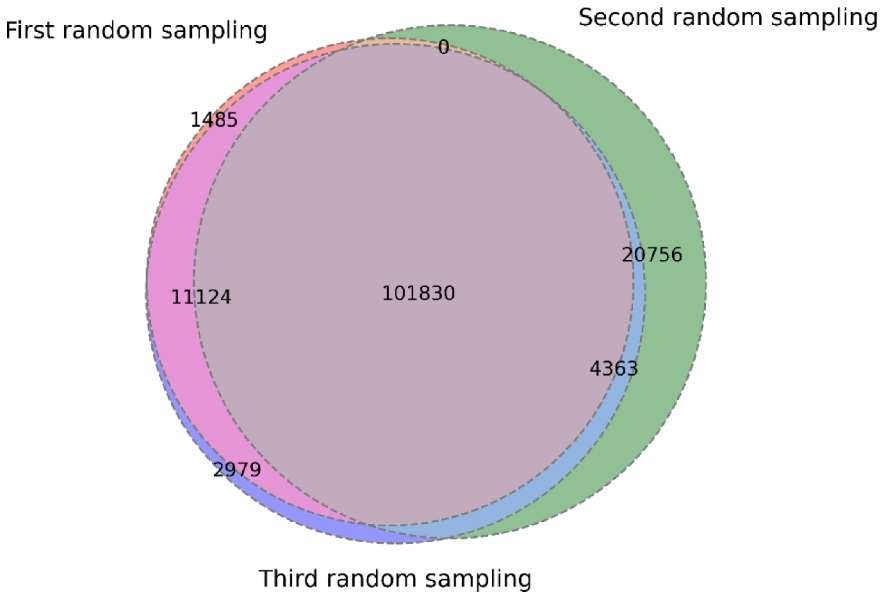


Fig. 14. The relationship between commuting pattern vehicles identified by three random sampling.

After getting the identification rules of commuting vehicles, the commuting vehicles in the whole dataset can be identified successfully. Although different HFTTP can be obtained when threshold values are 3%, 3.5% and 4%, the final commuting rules in the above three cases contain at least one of the OD stability coefficients of commuting trip.

Fig. 13 shows the relationship between commuting vehicles identified by taking different α. There are 122,192, 114,439 and 63,116 commuting vehicles identified when taking α as 3%, 3.5% and 4%, respectively. There are 103,929 commuting vehicles at the intersection of the commuting vehicles identified by taking 3% and 3.5% as the threshold respectively. There are 57,842 commuting vehicles at the intersection of the commuting vehicles identified by taking 3% and 4% as the threshold respectively. And there are 57,397 commuting vehicles at the intersection of the commuting vehicles identified by taking 3.5% and 4% as the threshold respectively. Besides, there are 56,054 commuting vehicles at the intersection of the commuting vehicles identified by taking 3%, 3.5% and 4% as the threshold respectively. When the α is 3% or 3.5%, the commuting rules are very similar, and the identified commuting vehicles have a high coincidence degree. When the α is 4%, the HFTTP become shorter than that when α is 3% or 3.5%, so the regularity of commuting vehicle travel behavior are higher than that of taking 3% or 3.5% as threshold. Therefore, the result of taking 4% as the threshold is a little different from the result of taking 3% or 3.5% as the threshold. But on the whole, the final commuting rules in the above three cases contain at least one of the OD stability coefficients of commuting trip.

6.4. Influence of random sampling on results

Because of the huge amount of LPR data, in order to get the results in an acceptable time, the whole dataset is randomly sampled before clustering algorithm. In order to verify the robustness of the algorithm proposed in this study, three sub-datasets are obtained by three random sampling. On the obtained sub-dataset, clustering algorithm and decision tree algorithm are run to get commuting rules and identify commuting pattern vehicles. The results based on three sub-datasets are compared to verify the robustness of the

algorithm. The commuting rules obtained by three random sampling are shown in Table 12, and the relationship between commuting pattern vehicles identified by three random sampling is shown in Fig. 14.

It can be found that the commuting rules obtained by three random sampling are all composed of OD stability coefficients of commuting trip. Using the commuting rules obtained by three random sampling to identify the commuting pattern vehicles, there are 101,830 vehicles in the intersection of the three sets, which shows that the algorithm proposed in this study is very robust to the randomness brought by random sampling.

7. Conclusions

Commute represents a very common and important travel behavior, especially during the morning and evening peak hours. The analysis of commute is helpful for urban planning and policy-makers in formulating travel policies such as travel restriction policy. Nowadays, there are numerous LPR detectors on the road network that collect a large amount of LPR data daily, thus acquiring the spatiotemporal information on travelers. Therefore, the LPR data can be used to analyze the travel behavior of travelers. In addition, compared to the questionnaire survey data, the LPR data have greater volume and longer tracking time, so the characterization of travel behavior based on the LPR data is more accurate. Based on the LPR data, this paper analyzes the commuting behavior of vehicles from both time and space perspectives. According to the analysis of the first and last detected times, the HFTTP in Hangzhou are identified, and a method for mining the OD of a vehicle during the HFTTP is proposed. Besides, a total of nine features that reflect the commuting pattern are extracted from the LPR data. Base on the correlation analysis, dimension reduction is performed on the nine features by the factor analysis method. Finally, three factors are obtained and used to characterize the commuting behavior from three perspectives: the stability of the vehicle travel behavior during the HFTTP on working days, the stability of the vehicle travel behavior during the non-HFTTP, the OD stability during the HFTTP. After sampling 1% of the entire LPR data, three different clustering methods, the DBSCAN algorithm, the ISODATA algorithm, the fast search and find of density peaks clustering algorithm, are applied to the sampled sub-dataset, and the obtained clustering results are analyzed and compared. The Silhouette coefficient is used to evaluate the clustering results of each algorithm. The results show that the DBSCAN algorithm cannot successfully recognize the commuting pattern, but the ISODATA algorithm and the fast search and find of density peaks clustering algorithm can successfully recognize the commuting pattern, and the ISODATA algorithm achieves the best commuting pattern recognition accuracy.

Based on clustering analysis, the decision tree algorithm is used to extract commuting rules, and finally, the commuting rule of ($F_{D1} > 0.31$ and $F_{D2} > 0.214$) or ($F_{D2} > 0.595$ and $F_{D1} \leq 0.31$) is obtained. Using this rule to identify commuting vehicles from the entire data, a total of 114,439 commuting vehicles in the core area of Hangzhou, including 111,945 vehicles with local license plates (license plate is Zhe A) and 2,494 vehicles with non-local license plates (license plate is non-Zhe A) are accurately identified. After that, The influence of different HFTTP on commuting rule extraction and commuting pattern vehicle recognition is carefully analyzed. The influence of randomness on the results is also tested. The proposed commuting pattern vehicle recognition algorithm is proved to be robust. By analyzing the travel behavior of commuting vehicles in more detail, it is found that the spatiotemporal regularity of commuting vehicles is much stronger than that of non commuting vehicles, the average of F_1, OD stability coefficients of commuting trip, F_s of non commuting vehicles are>70% smaller than commuting vehicles. In the collected LPR data, the detected records of the commuting vehicles accounted for 27.48% of all the records. During the HFTTP, the average daily commuting vehicles accounted for 34.48% of all the vehicles, and the detection frequency of commuting vehicles accounted for 34.36% of all the vehicles. In addition, it is also found that 78.0% of commuting vehicles were first detected between 06:30 AM and 10:30 AM on each workday on average. And 77.8% of commuting vehicles were last detected between 03:30 PM and 09:30 PM on each workday on average. However, for non commuting vehicles, the phenomenon that start or end the travel during the HFTTP is not obvious.

Future research can focus on the impact of transportation demand management measures on commuting travel behavior and the analysis of affecting factors on travel behavior, which can be used for further improving the effect of transportation demand management measures.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2018YFB1601000), and the National Natural Science Foundation of China (No. 92046011, and 91746105). This research is supported by LPR data from Hangzhou City Brain.

Appendix A

Determination of the origin of a commuting trip during the HFTTP in the morning and calculation of its stability coefficient

Algorithm: *Determination of the origin of a commuting trip during the HFTTP in the morning and calculation of its stability coefficient*

Input: LPR data;
Output: The origin of the commuting trip of the vehicle and its probability in the HFTTP in the morning.

(*continued on next page*)

(*continued*)

Algorithm: *Determination of the origin of a commuting trip during the HFTTP in the morning and calculation of its stability coefficient*

Methods:

Step 1: For each working day included in the LPR data, for each vehicle in the LPR data, get the first detected point in the HFTTP in the morning, get the second detected point in the HFTTP in the morning.

Step 2: Get the maximum probabilistic point of each first detected location of the vehicle and denote it as a. Get the second maximal probabilistic point of each first detected location of the vehicle and denote it as b. Get the maximum probabilistic point of each second detected location of the vehicle and denote it as c.

Step 3: For each working day included in the LPR data, for each vehicle in the LPR data, if the first detected point in the range$\{a, c\}$, then mark it as a.

Step 4: Get a after Step 3. Calculate the maximal probability p_1 of the first detected point. Calculate the second maximal probability p_2 of the first detected point.

Step 5: If p_2 is larger than the threshold k_1, then judge if $b != c$, then $p_1 = p_1 + p_2$, else if $b == c$, then $p_1 = p_1$.

Similarly, the origin and the corresponding stability coefficient during the HFTTP in the evening can be obtained by replacing the HFTTP in the morning with the HFTTP in the evening. And the destination and the corresponding stability coefficient can be obtained by replacing the first and second detected locations with the last and penultimate detected locations of a commuting trip.

References

Ball G H, Hall D J. ,1965 ISODATA a novel method of data analysis and pattern classification. Stanford research inst Menlo Park CA.

Brown, J., 2009. Choosing the right number of components or factors in PCA and EFA. JALT Testing & Evaluation SIG Newsletter 13 (2).

Cattell, Raymond B., 2012. The scientific use of factor analysis in behavioral and life sciences. Springer Science & Business Media.

Chang, Y., Duan, Z., Yang, D., 2018. Using ALPR data to understand the vehicle use behaviour under TDM measures. IET Intel. Transport Syst. 12 (10), 1264–1270.

Chen H, Yang C, Xu X., 2017. Clustering vehicle temporal and spatial travel behavior using license plate recognition data. Journal of Advanced Transportation 2017.

Dong, Y., Wang, S., Li, L., et al., 2018. An empirical study on travel patterns of internet based ride-sharing. Transportation research part C: emerging technologies 86, 1–22.

Ester, M., Kriegel, H.P., Sander, J., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise//Kdd 96 (34), 226–231.

Fruchter, B., 1954. Introduction to factor analysis. D. Van Norstrand Company.

Gutiérrez, M., Hurtubia, R., de Dios Ortúzar, J., 2020. The role of habit and the built environment in the willingness to commute by bicycle. Travel Behaviour and Society 20, 62–73.

Huang, J., Levinson, D., Wang, J., et al., 2019. Job-worker spatial dynamics in Beijing: Insights from Smart Card Data. Cities 86, 83–93.

Jia, N., Li, L., Ling, S., et al., 2018. Influence of attitudinal and low-carbon factors on behavioral intention of commuting mode choice–A cross-city study in China. Transportation research part A: policy and practice 111, 108–118.

Kung, K.S., Greco, K., Sobolevsky, S., et al., 2014. Exploring universal patterns in human home-work commuting from mobile phone data. PLoS ONE 9 (6), e96180.

Li, Y., Guo, Y., Lu, J., et al., 2019. Impacts of congestion pricing and reward strategies on automobile travelers' morning commute mode shift decisions. Transportation research part A: policy and practice 125, 72–88.

Liu, Z., Li, R., Wang, X.C., et al., 2018. Effects of vehicle restriction policies: Analysis using license plate recognition data in Langfang, China. Transportation Research Part A: Policy and Practice 118, 89–103.

Loh, W.Y., 2011. Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (1), 14–23.

Luo, X., Wang, D., Ma, D., Jin, S., 2019. Grouped travel time estimation in signalized arterials using point-to-point detectors. Transp. Res. Part B 130, 130–151.

Ma, D., Luo, X., Wang, D., Jin, S., Wang, F., Guo, W., 2017a. Lane-Based Saturation Degree Estimation at Signalized Intersections Using Travel Time Data. IEEE Intell. Transp. Syst. Mag. 9 (3), 136–148.

Ma, X., Liu, C., Wen, H., et al., 2017. Understanding commuting patterns using transit smart card data. J. Transp. Geogr. 58, 135–145.

Ma, X., Wang, Y., Chen, F., et al., 2012. Transit smart card data mining for passenger origin information extraction. Journal of Zhejiang University Science C 13 (10), 750–760.

Ma, X., Wu, Y.J., Wang, Y., et al., 2013. Mining smart card data for transit riders' travel patterns. Transportation Research Part C: Emerging Technologies 36, 1–12.

Memarsadeghi, N., Mount, D.M., Netanyahu, N.S., et al., 2007. A fast implementation of the ISODATA clustering algorithm. International Journal of Computational Geometry & Applications 17 (01), 71–103.

Mo, B., Li, R., Zhan, X., 2017. Speed profile estimation using license plate recognition data. Transportation research part C: emerging technologies 82, 358–378.

Qiu, G., Song, R., He, S., et al., 2018. Clustering passenger trip data for the potential passenger investigation and line design of customized commuter bus. IEEE Trans. Intell. Transp. Syst. 20 (9), 3351–3360.

Rao, W., Wu, Y.J., Xia, J., et al., 2018. Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. Transportation Research Part C: Emerging Technologies 95, 29–46.

Rodriguez, A., Laio, A., 2014. Clustering by fast search and find of density peaks. Science 344 (6191), 1492.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Santoyo S., 2017. A Brief Overview of Outlier Detection Techniques. Towards Data Science, September, 2017.

Shao, W., Chen, L., 2018. License Plate Recognition Data-Based Traffic Volume Estimation Using Collaborative Tensor Decomposition. IEEE Trans. Intell. Transp. Syst. 19 (11), 3439–3448.

Shen, X., Zhou, Y., Jin, S., Wang, D., 2020. Spatiotemporal influence of land use and household properties on automobile travel demand. Transp. Res. Part D 84, 102359.

Strathman, J.G., Kimpel, T.J., Broach, J., et al., 2008. Leverage ITS Data for Transit Market Research: A Practitioner's Guidebook. TCRP Report 126. Transportation Research Board, Washington, DC.

Sun, S., Yang, D., 2015. Identification of Transit Commuters Based on Naïve Bayesian Classifier. Journal of Transportation Systems Engineering and Information Technology 15 (6), 46–53.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2), 411–423.

Warne, R.T., Larsen, R., 2014. Evaluating a proposed modification of the Guttman rule for determining the number of factors in an exploratory factor analysis. Psychological Test and Assessment Modeling 56 (1), 104–123.

Wang, D., Fu, F., Luo, X., Jin, S., Ma, D., 2016. Travel Time Estimation Method for Urban Road Based on Traffic Stream Directions. Transportmetrica A 12 (6), 479–503.
Weng, X., Lv, P., 2019. Subway IC Card Commuter Crowd Identification Based on GBDT Algorithm. Journal of Chongqing Jiaotong University(Natural Science) 38 (5), 8–12.
Wickham, H., Stryjewski, L., 2011. years of boxplots. Am. Statistician. 40.
Yao, W., Ding, Y., Xu, F., et al., 2018. Analysis of cars' commuting behavior under license plate restriction policy: a case study in Hangzhou, China. 21st International Conference on Intelligent Transportation Systems (ITSC): 236–241.
Ye, R., De Vos, J., Ma, L., 2020. Analysing the association of dissonance between actual and ideal commute time and commute satisfaction. Transportation Research Part A: Policy and Practice 132, 47–60.
Zhan, X., Li, R., Ukkusuri, S.V., 2015. Lane-based real-time queue length estimation using license plate recognition data. Transportation Research Part C: Emerging Technologies 57, 85–102.
Zhou, J., Murphy, E., Long, Y., 2014. Commuting efficiency in the Beijing metropolitan area: An exploration combining smartcard and travel survey data. J. Transp. Geogr. 41, 175–183.