Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 3203065, 18 pages
https://doi.org/10.1155/2023/3203065

Research Article

A Modified Latent Dirichlet Allocation Topic Approach for Driving Style Exploration Using Large-Scale Ride-Hailing GPS Data

Ye Li,[1] **Yiqi Chen**,[1] **Jie Bao**,[2] **Lu Xing**,[3] **Jinjun Tang**,[1] **Changyin Dong**,[4] **and Ruifeng Gu**[1]

[1]*School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan 410075, China*
[2]*Civil Aviation College, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China*
[3]*School of Traffic and Transportation Engineering, Changsha University of Science and Technology, Changsha, Hunan 410075, China*
[4]*Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Ye Li; yelicsu@csu.edu.cn

Received 23 February 2023; Revised 30 August 2023; Accepted 21 October 2023; Published 4 November 2023

Academic Editor: Jing Dong-O'Brien

Driving style identification is of vital importance for intelligent driving system design and urban traffic management. This study aims to identify and analyze driving styles using large-scale ride-hailing GPS data taking different time periods, traffic, and weather conditions into account. The large-scale GPS data are collected and preprocessed, and then, the k-means clustering is implemented to acquire driving behavior. The modified latent Dirichlet allocation topic approach is applied to extract the driving states as the latent variables behind driving behaviors and finally recognize driving styles. The results show that driving styles are composed of five driving states with different probability combinations. Different driving styles in different situations are further analyzed and compared. When considering the impact of peak periods on the driving style, it indicates that styles tend to be conservative in the morning peak, free and dispersed in the evening peak, and diverse in the off-peak hours. While comparing styles regarding the influence of workdays, drivers act more cautiously and conservatively on weekdays but freer on weekends without the pressure of peak hours. The weather factor is also explored and rainy days are verified to be the resistance of driving so that most drivers become cautious and conservative. Finally, two aberrant driving styles are discovered and countermeasures are suggested to improve traffic safety.

1. Introduction

Intelligent driving systems are becoming increasingly popular in our lives, and autonomous driving technologies are being developed. However, the most challenging problem is not only technology but also people's acceptance of autonomous vehicles (AVs) [1]. Personalization would be an advisable measure to improve acceptance. One of the personalization measures is to simulate human driving behavior by switching to the appropriate driving style in certain scenarios. In the mixed driving environment, simulating human behavior for AVs can make driving behavior on the road more consistent, leading to safer driving and fewer crashes [2].

The driving style is a relatively stable state of individual driving behaviors [3]. Recognizing the driving style can help autonomous vehicles be more intelligent and user-friendly. The study of the driving style emerged in the 1990s. Existing studies on driving styles use questionnaires, driving simulators, and many other methods and the results have important implications for autonomous driving nowadays. There are great discrepancies in driving styles among different drivers due to various factors, such as gender, age, driving experience, psychological state, traffic situation, and even weather [4]. By identifying and storing different driving styles, an intelligent driving system can form a database of diverse styles, from which passengers can choose their favorite ones. This measure can also improve people's acceptance towards autonomous vehicles. In addition, aberrant driving styles may yield negative influence on road traffic throughput and safety [5–7]. For example, French et al. [8] found that dangerous driving styles are associated with serious traffic crashes. Generally, aberrant driving style recognition has several potential applications in practice as follows: (1) helping traffic managers analyze driving behaviors and the contributing factors of road crashes and identify driving risks in a timely manner, (2) enabling AVs to take timely measures to avoid conflicts with aberrant-style driving and potentially dangerous vehicles, and (3) assisting traffic managers in taking preventive measures for certain accidents in advance to improve traffic efficiency and traffic safety. Therefore, studies on recognizing driving styles and extracting aberrant ones are of great importance.

The primary objective of this study is to explore the driving style of ride-hailing drivers based on a modified LDA method using large-scale GPS data. The large-scale GPS ride-hailing data are firstly collected and preprocessed. Then, the k-means clustering algorithm is employed for acquiring driving behaviors. A modified latent Dirichlet allocation (LDA) topic approach is implemented to extract the driving states determined by a bag of driving behavior and discover various driving styles. The differences in driving styles are compared under various conditions, including weekdays or not, peak hours or not, and rainy or sunny day. Finally, two kinds of aberrant driving styles are screened out and corresponding countermeasures are suggested to improve traffic safety.

The main contributions of this paper are listed in the following: (1) applied the modified LDA algorithm with a new encoding method to extract driving behaviors from large-scale trajectory data and build a driving style library containing realistic driving styles for each driver individually; (2) revealed and compared the effects of different external factors on driving styles, including working days, peak hour congestions, and weather impact, and explored how the composition of driving styles shifted in different driving environments; and (3) identified abnormal and potentially dangerous driving within numerous styles in various driving environments based on the composition of each style from the driving style library.

The rest of the paper is organized as follows. Section 2 reviews previous studies related to this research. Section 3 describes the collection and preprocessing of large-scale ride-hailing GPS data. Section 4 introduces the clustering algorithm and LDA model, and the results are discussed in Section 5. Finally, the conclusions are drawn in Section 6.

2. Literature Review

Previous studies mainly collected data for studying the driving style and behavior by questionnaires. The driver behavior questionnaire (DBQ) was first proposed by Reason et al. [9], which classifies driving behaviors by features such as dangerous errors. It is still prevalent and its modified version has been implemented in a wide range of studies. For example, Dotse and Rowe [10] creatively applied the Manchester Driver Behavior Questionnaire in Ghana with 28 items to characterize aberrant driving behavior. To discover factors influencing driving methods, Deng et al. [11] also designed a 28-scale driver behavior questionnaire and improved vehicle lateral instability. Another novel research is to modify the driver behavior questionnaire with factors related to new issues such as texting, social media use, and the consumption of drug and alcohol [12]. The validation effectiveness of these questionnaires was also tested. For instance, Useche et al. [13] adopted an abbreviated version of a 9-item driver behavior questionnaire on freight drivers and tested its validation. They found that it can be well applied to obtain access to traffic violations and errors among this specific group of drivers.

In recent years, using the dynamic characteristics of vehicles to study driving styles has attracted researchers' attention and data collected from driving simulators have become prevalent [14]. Driving simulators can simulate abundant driving scenarios and record dynamic data collected while driving [15, 16]. Wang et al. [17] allowed the participants to drive vehicles in the simulator and collected these dynamic data. Then, the results were compared with previously collected and determined driving patterns by questionnaires to verify the effectiveness of their approach. To reduce the workload of labeling training data, Wang et al. [18] proposed a semisupervised method to classify drivers into aggressive and normal styles using simulation data. The environment in the driving simulator was set in advance by researchers. The simulator-based studies usually lead to two drawbacks. One is the lack of diversity and authenticity of the driving environment, and the other is that artificial settings will bring a certain degree of subjectivity. Besides, experiments carried out by driving simulators are usually costly and time consuming [19]. Therefore, GPS data collected from the real-world driving environment are being prevalent.

The amount of data collected from GPS devices is huge and covers a large area and thus quickly gained popularity among researchers. Compared to questionnaires and driving simulators, data collected from GPS devices are able to acquire a large number of real data, which could be employed for driving style analysis in a large scale. Studies using GPS data have been carried out by many researchers. For example, Aljaafreh et al. [20] classified driving styles into below normal, normal, aggressive, and very aggressive with data from GPS tracker devices. Ma et al. [21] identified three kinds of driving styles (aggressive, normal, and cautious) based on ride-hailing GPS data. GPS data integrated with data collected from monitoring cameras were applied by Qi et al. [22] to categorize driving styles into aggressive, cautious, and moderate types. Another analysis of driving styles was also implemented by Qi et al. [23] using GPS data and interactive data from the Chinese driving behavior database. They recognized diverse driving styles by topic models and verified the effectiveness of their method.

According to the aforementioned literature review, the advancement of data collection techniques makes it possible and promising to investigate driving styles on a larger scale and obtain more samples. The impact of the driving style on traffic safety is still a notable issue for researchers. Previous studies mainly focus on how to acquire driving styles, paying little attention to the difference under different driving situations. Comprehensive factors should be further considered, including the day of the week (weekday/weekend), time of the day (peak hour/off-peak hour), and weather conditions (rainy/sunny). In addition, aberrant and individual driving styles have not been fully explored in previous studies. Hence, this study intends to identify the driving style under diverse driving situations and explore and analyze the aberrant driving style using the large-scale GPS data.

3. Data Preprocessing

Data used in this research are Chengdu ride-hailing GPS data [24–27], which contain orders and vehicle trajectories for the entire month of November 2016. As shown Table 1, the GPS data include drivers' identity, orders' identity, time, and the vehicle location, with a sampling interval from two to four seconds. The whole dataset covers a part of the area in Chengdu, China. Its boundaries for a given longitude and latitude are as follows: (30.727818, 104.043333), (30.726490, 104.129076), (30.655191, 104.129591), and (30.652828, 104.042102). Note that the identity of drivers and their orders are anonymized and the coordinates are in the GCJ-02 coordinate system.

The data preprocessing procedures are shown in Figure 1, which include five major steps.

(1) Data selection. We select ride-hailing GPS data on the 6th, 7th, and 14th of November 2016, forming dataset 1, dataset 2, and dataset 3, respectively. The reason for selecting these three days is to cover both weekdays and weekends as well as sunny days and rainy days. Furthermore, we divide each dataset into three subdatasets by different time periods as follows: morning peak (7:00–10:00), evening peak (17:00–20:00), and off-peak (6:00–7:00, 10:00–17:00 and 20:00–22:00) [28]. Different time periods could manifest impacts of traffic congestion conditions on driving styles.

(2) Feature extraction. Since the raw data only contain the drivers' information about timestamp and location, which are not enough to analyze the driving style and behavior, it is necessary to extract more features to characterize movements of vehicles. Therefore, the distance, velocity, acceleration, and jerk are further calculated. Specifically, the formula of distance is given as follows [29]:

$$b_j^i = \left[\sin\left(\frac{\mathrm{lat}_j^i - \mathrm{lat}_{j-1}^i}{2}\right)\right]^2 + \cos\left(\mathrm{lat}_{j-1}^i\right) \times \cos\left(\mathrm{lat}_j^i\right) \times \left[\sin\left(\frac{\mathrm{lng}_j^i - \mathrm{lng}_{j-1}^i}{2}\right)\right]^2,$$
$$d_j^i = \begin{cases} 2 \times \sin^{-1}\sqrt{b_j^i} \times R \times 1000; \ j > 1, \\ 0; \ j = 1, \end{cases} \tag{1}$$

where lat_j^i and lng_j^i denote the latitude and longitude of point j of driver i, respectively; d_j^i represents the distance of driver i moving from point $j-1$ to point j; and R represents the radius of the Earth and $R = 6,371$ km here. The velocity and acceleration at point j can be further calculated as follows:

$$v_j^i = \begin{cases} \dfrac{d_j^i}{t_j - t_{j-1}}; \ j > 1, \\ 0; \ j = 1, \end{cases}$$
$$a_j^i = \begin{cases} \dfrac{v_j^i}{t_j - t_{j-1}}; \ j > 1, \\ 0; \ j = 1, \end{cases} \tag{2}$$

where $\mathrm{v}_\mathrm{j}^\mathrm{i}$ and $\mathrm{a}_\mathrm{j}^\mathrm{i}$ denote the velocity and acceleration of driver i at point j, respectively; and t_j represents the timestamp at point j. The jerk, also known as variable acceleration, is the change rate of acceleration over time [30]. When the vehicle accelerates or push brakes suddenly, drivers in the vehicle will have a strong sense of discomfort due to the large jerk and the degree to which different drivers can withstand is different. Therefore, to fully describe the driving behavior of drivers, jerk is also considered one of the key characteristics [31].

$$\mathrm{jerk}_j^i = \begin{cases} \dfrac{a_j^i}{t_j - t_{j-1}}; \ j > 1, \\ 0; \ j = 1. \end{cases} \tag{3}$$

(3) Outlier elimination. Due to the quality of positioning devices and potential errors of data acquisition, transmission, and storage processes, GPS data inevitably have outliers that will negatively affect the analysis results. Therefore, it is imperative to identify and eliminate abnormal data. The steps for outlier detection and elimination are as follows: (i) filter out

Table 1: GPS data description.

Features	Types	Samples
Driver ID	String	glox.jrrlltBMvCh8nxqktdr2dtopmlH
Order ID	String	jkkt8kxniovIFuns9qrrlvst@iqnpkwz
Timestamp	String	1501584540
Longitude	Float	104.04392
Latitude	Float	30.727818

repeatedly sampled data and keep only one valid point; (ii) remove data not within the sampling frequency; (iii) the outliers that are beyond the reasonable intervals are eliminated. The reasonable intervals of velocity, acceleration, and jerk are (0, 33.33 m/s), ($-8\,\text{m/s}^2$, $4\,\text{m/s}^2$), and ($-11\,\text{m/s}^3$, $11\,\text{m/s}^3$), respectively [32, 33]; and (iv) remove the data where the running distance is zero.

(4) Data normalization. It aims to standardize data to form dimensionless data and speed up the initialization as well as iteration of the data analysis algorithm. The commonly used Z score normalization is applied in this study as follows:

$$x' = \frac{x-\mu}{\sigma}, \tag{4}$$

where x' denotes the normalized data and μ and σ represent the mean and standard deviation of the original data, respectively.

(5) Driver selection. Drivers with the number of GPS points from 500 to 1000 are selected since the sample size of most drivers is among this range. Ultimately, 389,971 sample data were adopted for further analysis.

4. Methodology

This study aims to analyze the driving style through driving behavior. Previous studies obtained the driving style mainly by clustering trajectory data directly [21, 34–37]. However, in this study, it is considered that each GPS data point, containing the information of speed, acceleration, and jerk, only reflects the driver's action and some actions with similar values can be regarded as one type of behavior [23]. This means that driving action is the basic unit of driving behavior. Therefore, the actions should be aggregated to represent driving behaviors before driving behavior is used to recognize driving styles.

Therefore, concepts of driving action, driving behavior, driving state, and the driving style are defined as follows: (1) driving action: each GPS data point, the basic unit of driving behavior; (2) driving behavior: an aggregated form of driving action; (3) driving state: an intermediate variable between driving behavior and style; and (4) driving style: a manner of driving.

Figure 2 displays the framework for recognizing driving styles using large-scale ride-hailing GPS data. First, GPS data are processed by extracting features, eliminating outliers and normalizing. Second, the driving behavior is obtained by the k-means clustering algorithm. Then, the modified LDA topic model is implemented to recognize the driving style. The recognition results of different datasets, including at peak hours or off-peak hours, on weekdays or weekends, and on a sunny day or a rainy day, are further compared. Finally, corresponding countermeasures are discussed to reduce aberrant driving styles based on modeling results.

4.1. K-Means Clustering Algorithm. Clustering is a data analysis method that classifies a given sample into a specific category according to the similarity or distance among features [38]. The principle of classification is to divide data into a set of groups by minimizing within-group distances and maximizing between-group distances. K-means clustering is one of the most widely used clustering methods [39]. It divides the sample dataset into k subsets denoting k categories and then assigns samples into these categories, making each sample have the smallest distance to the category center to which it belongs [38].

Given a sample dataset $D = \{x_1, x_2, x_3, \ldots, x_m\}$, assuming that the set of divided k categories is $C = \{C_1, C_2, C_3, \ldots, C_k\}$, the goal of k-means clustering is to minimize the square error. The objective function of k-means clustering can be written as follows [40]:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - \mu_i \right\|^2, \tag{5}$$

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x, \tag{6}$$

where μ_i is the mean vector of the class C_i. Equation (5) depicts the tightness of intracluster sample data, in which a smaller E represents a higher similarity and a smaller distance.

To minimize equation (5), k-means clustering uses an iterative algorithm. A parameter k denoting the number of clusters or categories should be given in advance. There are many methods to determine the parameter k, for example, using metrics such as the silhouette coefficient index [41] and the Davies–Bouldin index [42] or through experiments [23]. At the beginning of the iterative algorithm, k class centers are randomly selected, followed by calculating the distance between sample points and each cluster center. Sample points are then assigned to the nearest cluster. For the current classification, a mean vector μ_i of cluster C_i, indicating the location of the current cluster center after the first classification, is recalculated. Afterwards, we replace the previous cluster center with the new one. Iteration is continued until the maximum number of iterations or the minimum adjustable amplitude threshold is reached. Finally, each sample point obtains a category label [43].

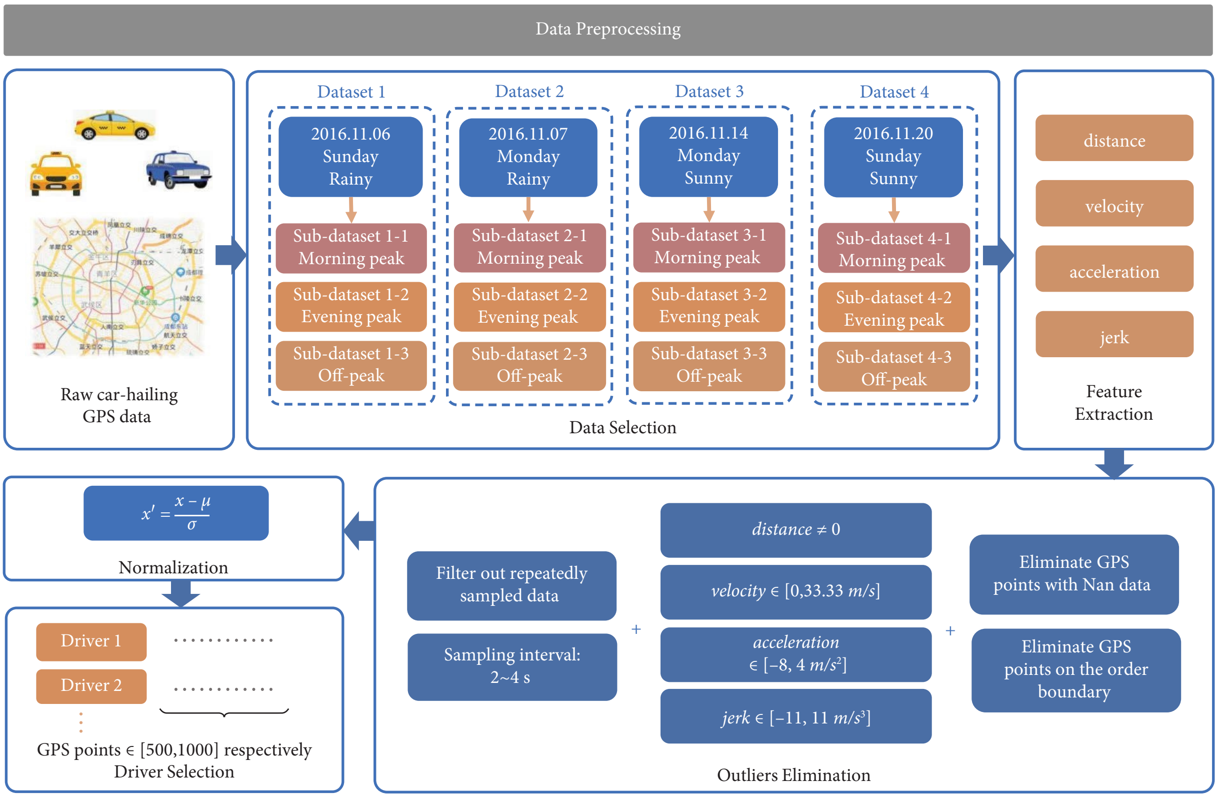


Figure 1: The flowchart of data preprocessing procedures.

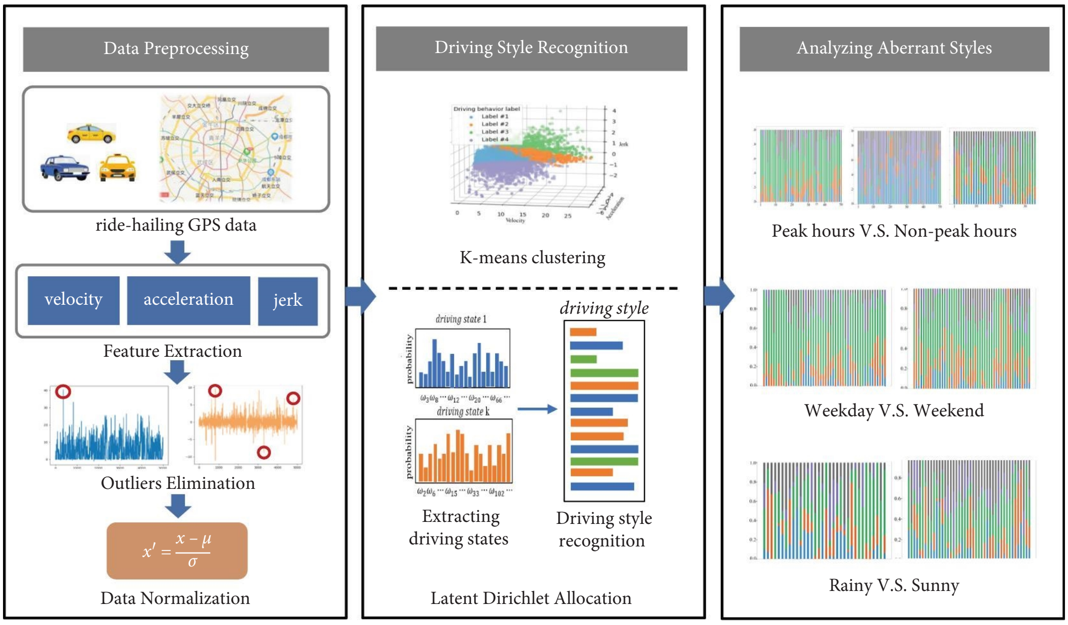


Figure 2: Framework for recognizing driving styles using large-scale GPS data.

4.2. Latent Dirichlet Allocation. The LDA topic model is a generative probability model based on Bayesian learning [44]. Different from previous topic analysis models such as probabilistic latent semantic analysis, the LDA is characterized by using the Dirichlet distribution as the prior distribution of the multinomial distribution, which prevents the overfitting problem that often occurs in machine learning. It has been widely applied by researchers, especially in the scopes of image classification, search engine optimization, capturing hot topics, visual scene classification, and so forth [45–50].

The LDA defines document collection as $w = \{w_1, \cdots, w_m, \cdots, w_M\}$, where w_m denotes a document in position m and the total number of documents is M. They consist of the word sequence $w_m = (\omega_{m1}, \cdots, \omega_{mn})$, in which ω_{mn} denotes the word in the position n in w_m. In addition, topic vectors are defined as $z = \{z_1, \cdots, z_t, \cdots, z_T\}$, where z_t represents the topic in position t and the total number is T [44]. Each document w_m is decided by a conditional probability distribution $p(z \mid w_m)$ of topics. The distribution $p(z \mid w_m)$ follows a multinomial distribution with the parameter θ_m, which is generated by the Dirichlet distribution as a prior distribution with the hyperparameter α. Their relationship can be written as $z_t \sim \text{Mult}(\theta_m)$ and $\theta_m \sim \text{Dir}(\alpha)$. Likewise, for topic-word distribution $p(\omega \mid z_t)$, there are also relationships as follows: $\omega_{mn} \sim \text{Mult}(\varphi_t)$ and $\varphi_t \sim \text{Dir}(\beta)$ [51]. Among these notations, only ω_{mn} is the observed variable and θ_m, φ_t, and z_t are all hidden variables, while α and β are both hyperparameters.

Applying LDA for topic analysis estimates the posterior probability distribution through a given document collection, thereby learning the topic distribution $p(z \mid w_m)$ of each text and the word distribution $p(\omega \mid z_t)$ of each topic [52]. According to Figure 3, in this research, we modify the LDA approach for driving style analysis and assume that the driving style collection is $w = \{w_1, \cdots, w_m, \cdots, w_M\}$, where M is the number of styles as well as drivers. The GPS data points are clustered and form a collection $C = \{c_1, c_2, c_3, \ldots, c_k\}$, where k indicates the number of clusters. The driving behavior sequence of each driver $w_m = (c_{m1}, \cdots, c_{mn})$ is the observed variable sequence, elements of which are from the set C according to the real driving situation. In addition, latent topics z are recognized as driving states of drivers. Therefore, by computing the driving state distribution of each style $p(z \mid w_m)$ and the driving behavior distribution of each state $p(\mathrm{c} \mid z_t)$, the driving styles can be analyzed.

The primary difference between the modified LDA model (Figure 4) and the original LDA model (Figure 5) is the introduction of driving behavior categories obtained from k-means clustering. The word for the original LDA is ω_{mn}, while in the modified LDA proposed in this study, x denotes each sample of the original data which is a vector about the speed, acceleration, and jerk. The driving behavior category c_k is the counterpart of ω_{mn}. A mapping from raw data x to the category of driving behavior c_k is added to our method. Therefore, the original time-series data x is transformed into the driving behavior group sequence $C = \{c_1, c_2, c_3, \ldots, c_k\}$, and thus, it forms the corpus of each driver.

The variational inference expectation-maximum (EM) algorithm is employed in this study to solve the LDA model. This algorithm defines the variational distribution $q(\varphi, z, \theta \mid c, \alpha, \beta)$ to approximate the posterior probability distribution $p(\theta, \varphi, z \mid \lambda, \phi, \gamma)$, in which λ, ϕ, and γ are the variational parameters. The goal of solving this problem is to maximize the objective function as follows [53]:

$$L(\lambda, \phi, \gamma;\ \alpha, \beta) = E_q[\log p(\theta, \varphi, z, c \mid \alpha, \beta)] - E_q[\log q(\varphi, z, \theta \mid \lambda, \phi, \gamma)], \tag{7}$$

where the mathematical expectation is the definition of distribution $q(\varphi, z, \theta \mid c, \alpha, \beta)$, written as $E_q[\bullet]$ for convenience; $L(\lambda, \phi, \gamma;\ \alpha, \beta)$ presents the maximum likelihood function; λ, ϕ, and γ are the variational parameters; and α and β are the parameters of the LDA model.

After inputting driving behavior sequences of all drivers $w = \{w_1, \cdots, w_m, \cdots, w_M\}$ into the LDA model, the following E-step and M-step are iterated alternately until the model converges [54]:

E-step: use the current estimated model parameters α and β to estimate the variational parameters λ, ϕ, and γ by maximizing equation (7).

M-step: use the current variational parameters λ, ϕ, and γ to estimate model parameters α and β by maximizing objective equation (7).

After completing the iteration of the E-step and M-step, the optimal parameters α and β could be obtained as well as the driving state-behavior distribution $p(\mathrm{c} \mid z_t)$ and driving style-state distribution $p(z \mid w_m)$.

5. Results and Discussion

5.1. Results of k-Means Clustering. Considering that the role of k-means clustering is to transform the inputs of LDA into categorical variables [55], the parameter k should be taken as large as possible within a reasonable range to avoid homogenization of driving styles. Besides, in previous studies of topic models, it is acceptable if k is taken between 120 and 136 [22, 23, 55]. After multiple trials, it was found that $k = 120$ contributes to stability and interpretability of model results, as well as transforming the original redundant vectors into understandable driving behaviors. Therefore, the parameter k of k-means clustering is taken as 120 in this study.

The result of clustering driving action in order to acquire driving behavior is shown in Figure 6. It shows that actions are composed of velocity, acceleration, and jerk and that driving behavior is represented by 120 groups in different colors.

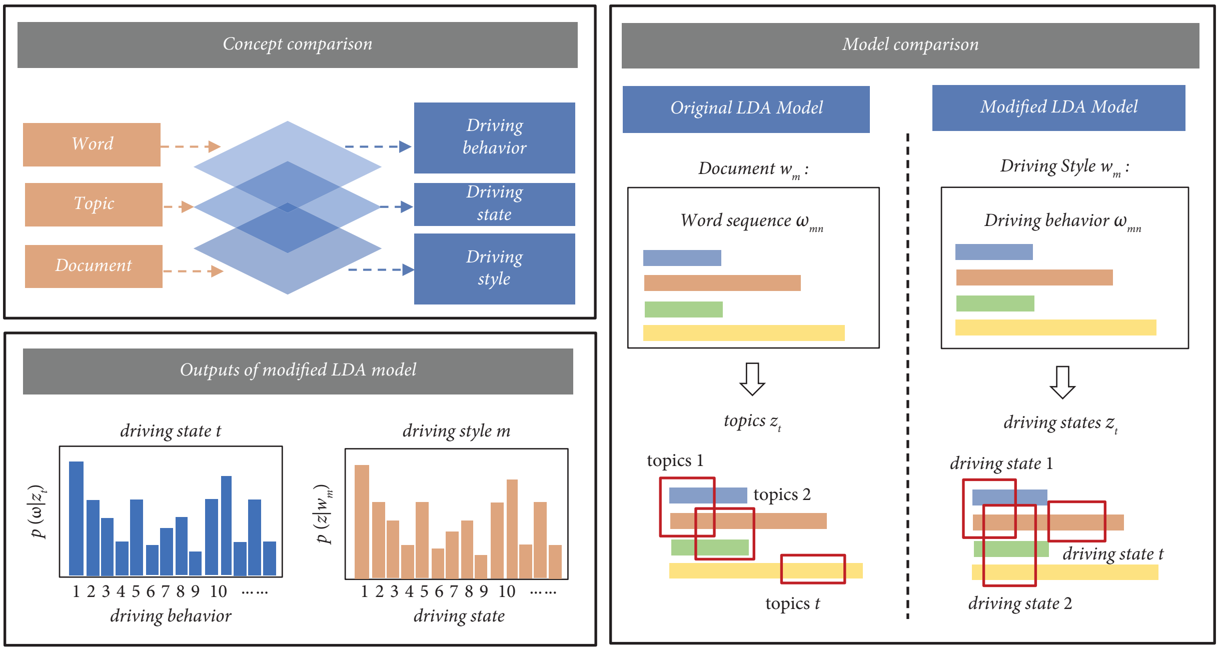


FIGURE 3: The schematic comparison between the original and modified LDA models.

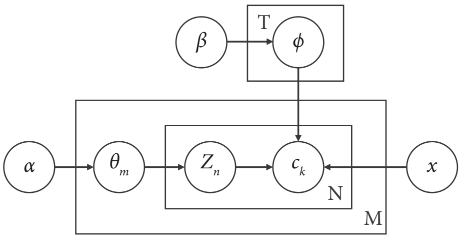


FIGURE 4: Graphical model of the modified LDA.

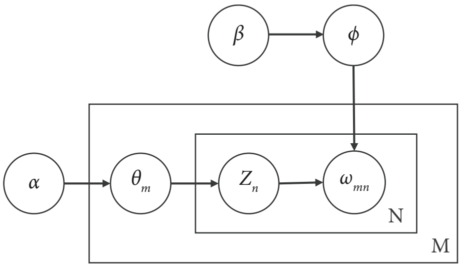


FIGURE 5: Graphical model of the original LDA.

5.2. Results of Driving State Recognition. Then, the recognition of the driving style is carried out. Before implementing the LDA model, the hyperparameters α and β and the number of topics T should be specified. α controls the sparsity of topics in the document. A low α value means that the document only covers a small number of topics, so α is usually set to a small fraction (1/T) of the number of topics. Likewise, β plays a role in the sparsity of words in the topic. The smaller it is, the more uneven the word probability distribution is, meaning that each topic will be more specific. Therefore, in this study, α is taken as the default 1/T and $\beta = 0.01$ based on preliminary tests [55]. The semantic coherence score is an indicator which is commonly used to evaluate the topic coherence. The calculation of this index is shown in the following equation [56–58]:

$$C_{\text{UMass}} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{D\left(w_i, w_j\right) + \varepsilon}{D\left(w_j\right)}, \qquad (8)$$

where $D(w_j)$ denotes the document frequency of word w_j, $D(w_i, w_j)$ represents the frequency of word w_i, and w_j is the co-occurring in the same document. N denotes the number of the most probable words in topic t, and $N = 20$ here [59]. ε is a constant to ensure the logarithm is not zero and $\varepsilon = 1$ in this study [56]. This indicator is always negative where the higher represents the higher quality of the topic [58, 60]. As is shown in Figure 7, the coherence of each topic is calculated and reaches the highest when $T = 5$.

In addition, the perplexity score is also utilized to figure out the optimal number of topics. A smaller perplexity indicates a better performance, which can be calculated as follows [49]:

$$\text{perplexity}(D) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p\left(w_d\right)}{\sum_{d=1}^{M} N_d}\right\}, \qquad (9)$$

where N_d represents the number of words in document w_d in position d and $p(w_d)$ denotes the probability of w_d. As shown in Figure 8, the perplexity of each topic is calculated, where the number of topics ranging from one to fifty. The result shows that when $T = 5$, the perplexity reaches the minimal value. Both the results of coherence and the perplexity index suggest that the most appropriate number of topic is five. Therefore, $T = 5$ is selected, which means the number of potential driving states is five.

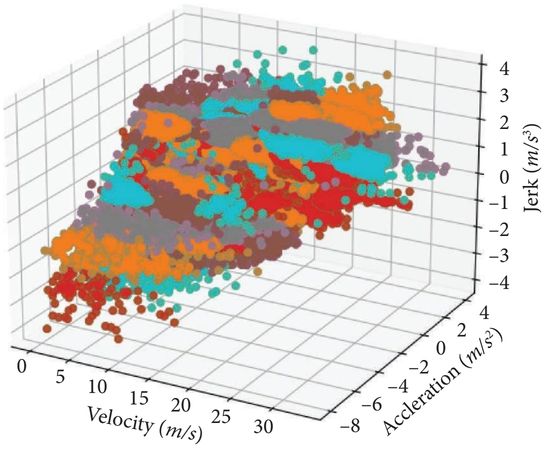


Figure 6: Results of the k-means clustering algorithm.

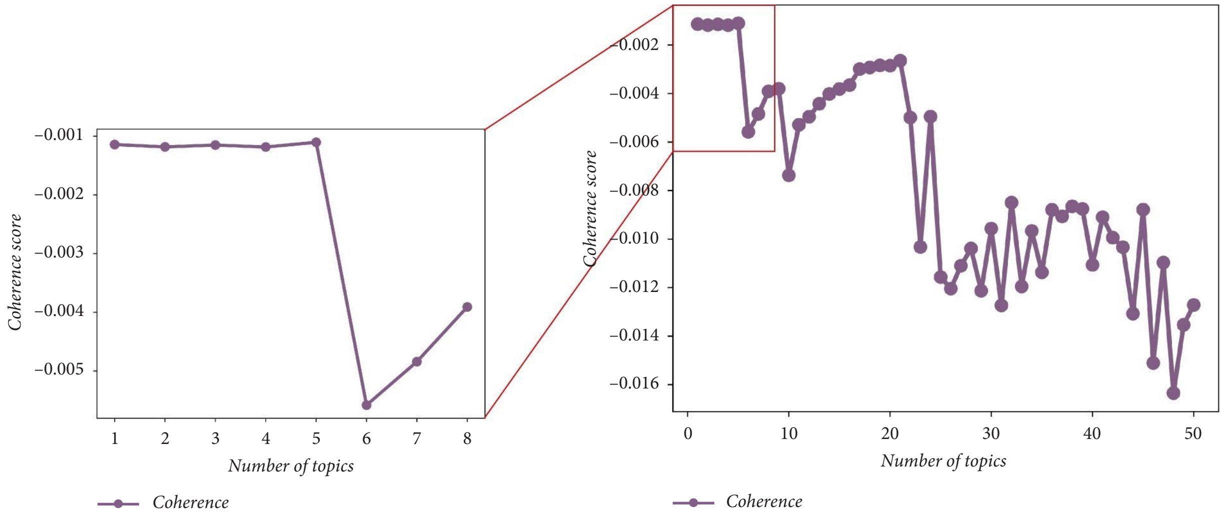


Figure 7: Coherence score of each topic.

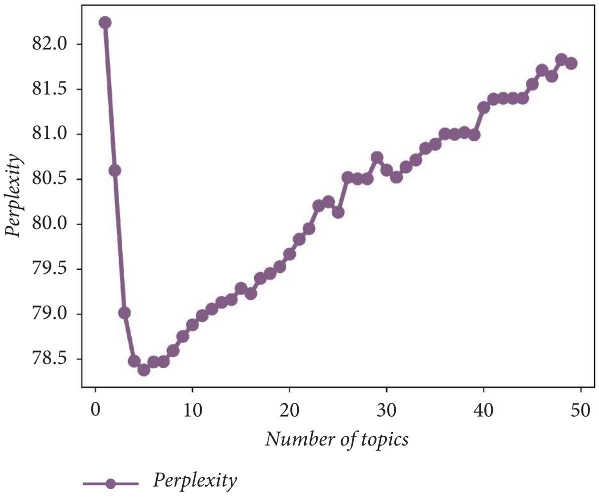


Figure 8: Perplexity of each topic.

The results of implementing the LDA model is shown in Table 2 after determining the parameters α, β, and T. Five driving states are obtained. Each driving state is composed of different proportions of 120 driving behavior (DB) groups. Table 2 shows each driving state's top 10 driving behavior groups with the largest proportions. For example, the driving behavior group that occurs most frequently in the driving state #1 is group 3, and its probability is 0.0237. The second most frequent one is group 108, and its probability is 0.0210. Comparing the percentages of these driving behaviors in each state, we can see that the probabilities of the components differ significantly in driving state #2, #3, or #5, respectively, which means that they each have an explicit topic. However, the probabilities of each component are relatively similar in state #1 or #4.

To understand the meaning of these driving states more intuitively, the relationship between the driving states and the main driving behaviors is shown in Figure 9. Driving states #1–#5 are denoted by the colors blue, orange, green, gray, and purple. A larger point in this graph represents a larger probability for a driving behavior to occur in a certain driving state. From Figure 9, it can be distinguished easily that most driving behavior groups in the driving state #3 (green) are in the high-speed and low-stability areas. Also, the most frequent driving behaviors in the driving state #2 (orange) are concentrating on medium-speed and medium-stability areas. It is found that it is complex to distinguish driving states #1, #4, and #5. Therefore, we enlarged the red-boxed area of Figure 9(c) and plotted it in Figure 10(a). And, the driving behaviors of each driving state, excluding states #2 and #3, are shown in Figures 10(b)–10(d).

In Figure 10, we can compare the behaviors in the driving states #1 (blue), #4 (gray), and #5 (purple). In comparison, #5 is more concentrated on low to medium speeds, with most of the acceleration and jerk being relatively low, and can be considered to be in a very stable state. The blue and gray dots are more concentrated in the low-speed area, where the blue dots have mostly larger acceleration and jerk than the gray ones, implying more instability. Therefore, it can be assumed that the blue state represents a low-speed and low-stability state, while the gray one is a low-speed and medium-stability state.

The meanings of each driving state are summarized in Figure 11, and cells in the table containing a color indicate that the state contains this meaning. Driving state #1 represents a low-speed and low-stability state. Driving state #2 denotes a medium-speed and medium-stability state and state #3 denotes a high-speed and low-stability one. In addition, driving state #4 is a low-speed and medium-stability state while driving state #5 is a low-to-medium speed and high stability one.

5.3. Results of Driving Style Recognition. Figures 12–15 show the composition of each driver's style. Each bar in the figure represents a driver's style, one of which is made up of five different colored driving states. The meaning of each color corresponding to a driving state has been discussed above and is shown in summary in Figure 11. The horizontal axis is the drivers' identification.

The first ten drivers in Figure 12(a) are the same ones as the first ten drivers in Figure 12(c), respectively. This also occurs in Figures 13–15, implying that these drivers passed through the sampling sites in both the morning peak and the off-peak period. For example, no. 301 in Figure 12(a) is the same person as no. 401 in Figure 12(c) and no. 455 in Figure 13(a) and no. 555 in Figure 13(c) are the same driver. Apart from these, all other drivers are distinct from each other. After data analysis, it is found that drivers in the evening peak generally do not drive through the same locations in the morning peak and off-peak hours, so drivers the in evening peak period are unique.

Comparing the three subplots in Figure 12 reveals how whether or not it is peak hour has an effect on drivers' driving style on a sunny Monday. First, Figure 12(a) shows the driving style of drivers in the morning peak. The majority of their style composition is the driving state # 4 (gray) and state #5 (purple), and there is also a more driving state #2 (orange) than in the off-peak period (Figure 12(c)), representing the fact that driving styles of the morning peak is mostly low speed. The high stability of purple indicates lower acceleration and jerk, probably due to the fact that the morning peak hours are usually congested. These data were collected in one of the most congested areas of Chengdu during the morning peak, so it can be inferred that the low acceleration and jerk of drivers during the morning peak is due to traffic jam without much room for movement. While during the evening peak hours (Figure 12(b)), there is a larger composition of driving state #3(green) and lesser state #2 than in the morning peak but still more in state #1(blue) than in state #5(purple) if compared to the flat peak. This means that the driving styles at the evening peak is similar to the morning peak but with slightly higher speed, acceleration, and deceleration.

There is a higher probability of driving states #1 and #3 and lower probability of states #4 and #5 in the off-peak period (Figure 12(c)) compared to the morning peak and evening peak hours. This indicates generally higher speeds and less stability. This is perhaps due to the fact that during the flat-peak period, the roads are less congested, giving drivers more freedom to drive, resulting in a diversity of driving operations. The trend of individual driving styles is the same as overall. There is a fewer proportion of gray, purple, and orange driving states and larger proportion of blue and green states during off-peak period, indicating generally less stable. There are likely more drivers changing lanes freely, accelerating and decelerating sharply. During this period, drivers could display a various range of driving styles and being less influenced by the traffic flow.

In addition, we found that the drivers would keep a consistent style at different times of the day. For example, driver 305 in Figure 12(a) and driver 405 in Figure 12(c), technically the same person; his off-peak style (id = 405) is mostly composed of blue and green driving state, indicating no matter how fast he drives, he would often accelerate and decelerate aggressively and result in a high jerk. When it comes to the morning peak hour style (id = 305), although he performs low speed with low acceleration and jerk, some of the high-speed instability is still retained.

Table 2: Results of the LDA model and latent driving states (DB: driving behavior group).

Driving state #1		Driving state #2		Driving state #3		Driving state #4		Driving state #5	
DB	$p(\omega\|z_k)$	DB	$p(\omega\|z_k)$	DB	$p(\omega\|z_k)$	DB	$p(\omega\|z_k)$	DB	$p(\omega\|z_k)$
3	0.0237	5	0.0974	33	0.0590	16	0.0483	62	0.0624
108	0.0210	90	0.0590	49	0.0511	62	0.0374	16	0.0554
90	0.0210	3	0.0378	5	0.0293	117	0.0339	3	0.0512
62	0.0188	29	0.0354	98	0.0255	65	0.0292	108	0.0499
115	0.0166	47	0.0335	90	0.0221	54	0.0282	54	0.0433
72	0.0162	42	0.0309	42	0.0213	53	0.0277	90	0.0404
56	0.0152	68	0.0270	29	0.0201	6	0.0261	72	0.0383
89	0.0151	23	0.0257	3	0.0182	26	0.0238	115	0.0353
54	0.0149	115	0.0242	105	0.0163	72	0.0225	65	0.0328
53	0.0148	119	0.0191	110	0.0153	10	0.0220	117	0.0315

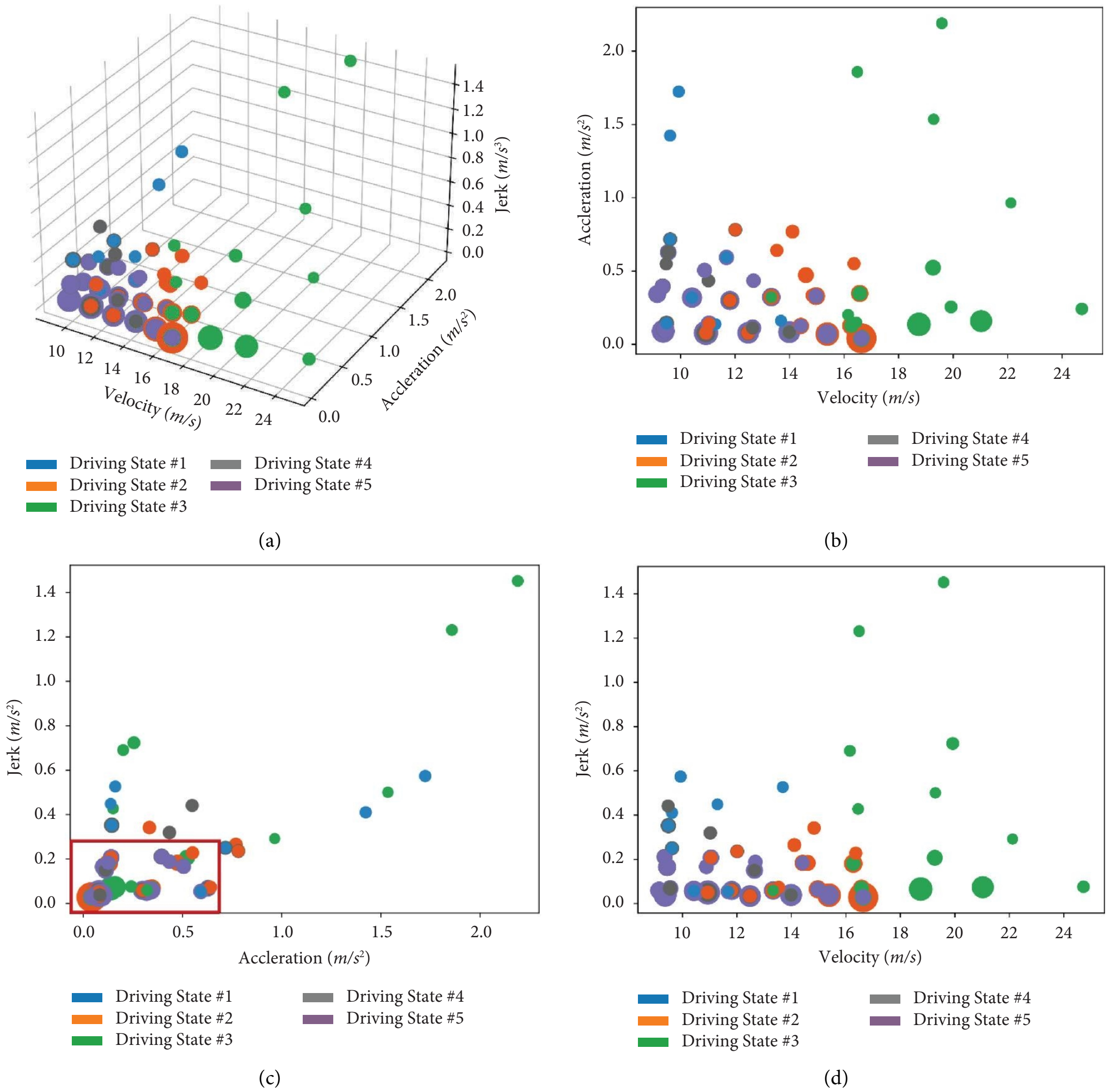
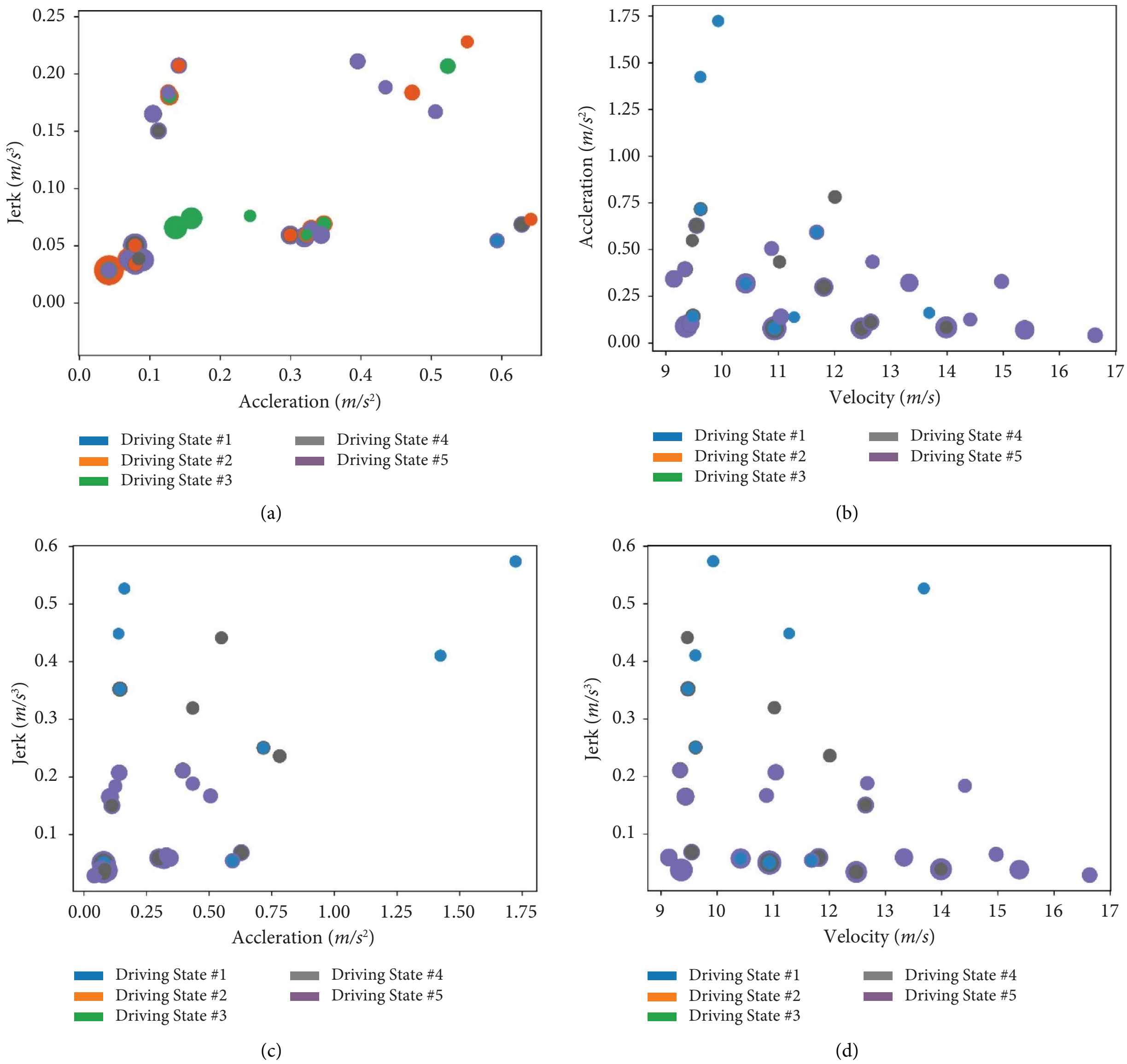


Figure 9: Scatter plot of driving states.

However, there are also some drivers who display different driving styles at different times of the day, such as 455 in Figure 13(a) and 555 in Figure 13(c). This driver's style on a sunny Sunday during the flat peak (id = 555) consists almost exclusively of driving states #4 and #5, implying a relatively conservative, low to medium speed, and stable



(a)

(b)

(c)

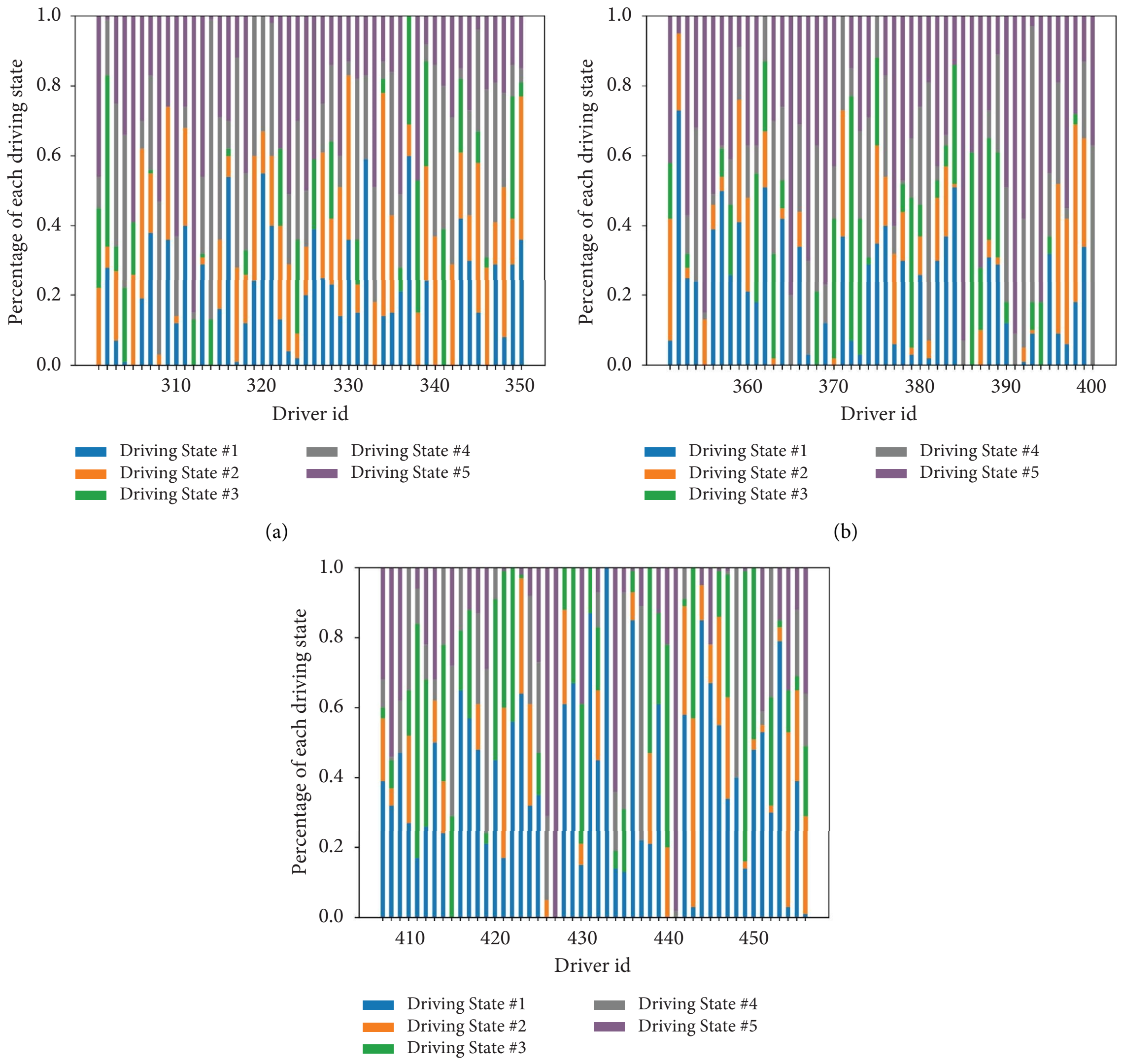
(d)

Figure 10: More details about the driving-state scatter plots.

Driving state	Speed			Stability		
	low	medium	high	low	medium	high
#1	🟦			🟦		
#2		🟧			🟧	
#3			🟩	🟩		
#4	▨				▨	
#5	🟪	🟪				🟪

Figure 11: Meanings of five driving states.

driving style. However, he was in a medium-high and unstable state for half of the time during the morning peak (id = 455). This finding is similar to the conclusions of other researches [23, 61], which have found that there would be changes in individual driving styles sometimes and that are largely due to the driving environment.



(a)

(b)

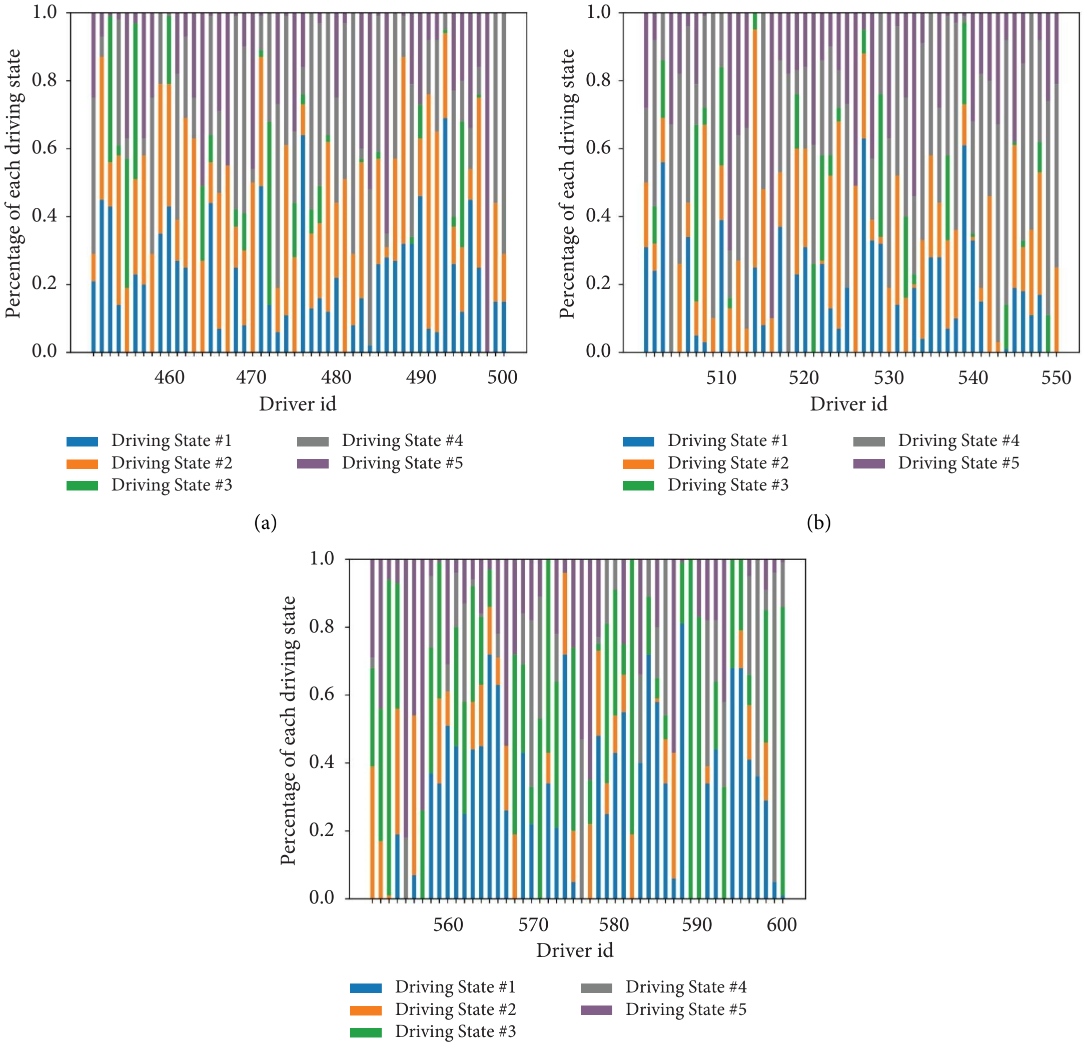
(c)

Figure 12: Driving style on sunny Monday. (a) Subdataset 3–1 (morning). (b) Subdataset 3-2 (evening). (c) Subdataset 3-3 (off-peak).

Figure 13 shows the driving styles of drivers on a sunny Sunday. Comparing Figures 12 and 13, it is possible to explore the effect of being on a weekday on the driving style. First, we make comparisons between Figures 12(a) and 13(a), which is the comparison between the morning peak on a weekday and a weekend. It can be seen that there is a significantly larger proportion of orange driving states on Sunday mornings than on Monday morning peaks, where the average proportion of the former is 0.285 compared to 0.2014 for the latter. Also, there are fewer purple driving states on weekend morning than on Monday, where the average proportion is 0.2698 on Monday and only 0.1766 on Sunday. Therefore, it can be concluded that the driving styles of drivers on weekend morning are at a more similar and focused speed with higher stability, implying a status with more freedom.

The difference between the evening peak on Monday and Sunday can be studied when comparing Figures 12(b) with 13(b). The average proportion of the driving state #5 in Figure 12(b) is 0.3838, while that of the driving state #4 in Figure 13(b) is 0.3658, and these two driving states dominate the driving style at that time of day, respectively. As the gray is low speed and medium stability, and the purple is low-to-medium speed and high stability, it can be assumed that drivers travelling in the weekend-evening peak tend to be more aggressive in acceleration and deceleration, although driving at low speed, compared to Monday evening.

In addition, we compare the Monday off-peak hours in Figure 12(c) and Sunday off-peak hours in Figure 13(c). They are both dominated by the driving state #1. However, the driving state #3 has a very different probability. It only accounts for 0.1942 on Monday but goes up to 0.2780 on Sunday, indicating that the driving style during the Sunday flat-peak hours is at higher speed and lower stability than that of Monday. This result makes sense because weekend travel destinations are more varied and the traffic is rarely congested. The driving styles performed would be similar to the one under the free-flow traffic condition.



(a)

(b)

(c)

Figure 13: Driving style on sunny Sunday. (a) Subdataset 4–1 (morning). (b) Subdataset 4–2 (evening). (c) Subdataset 4-3 (off-peak).

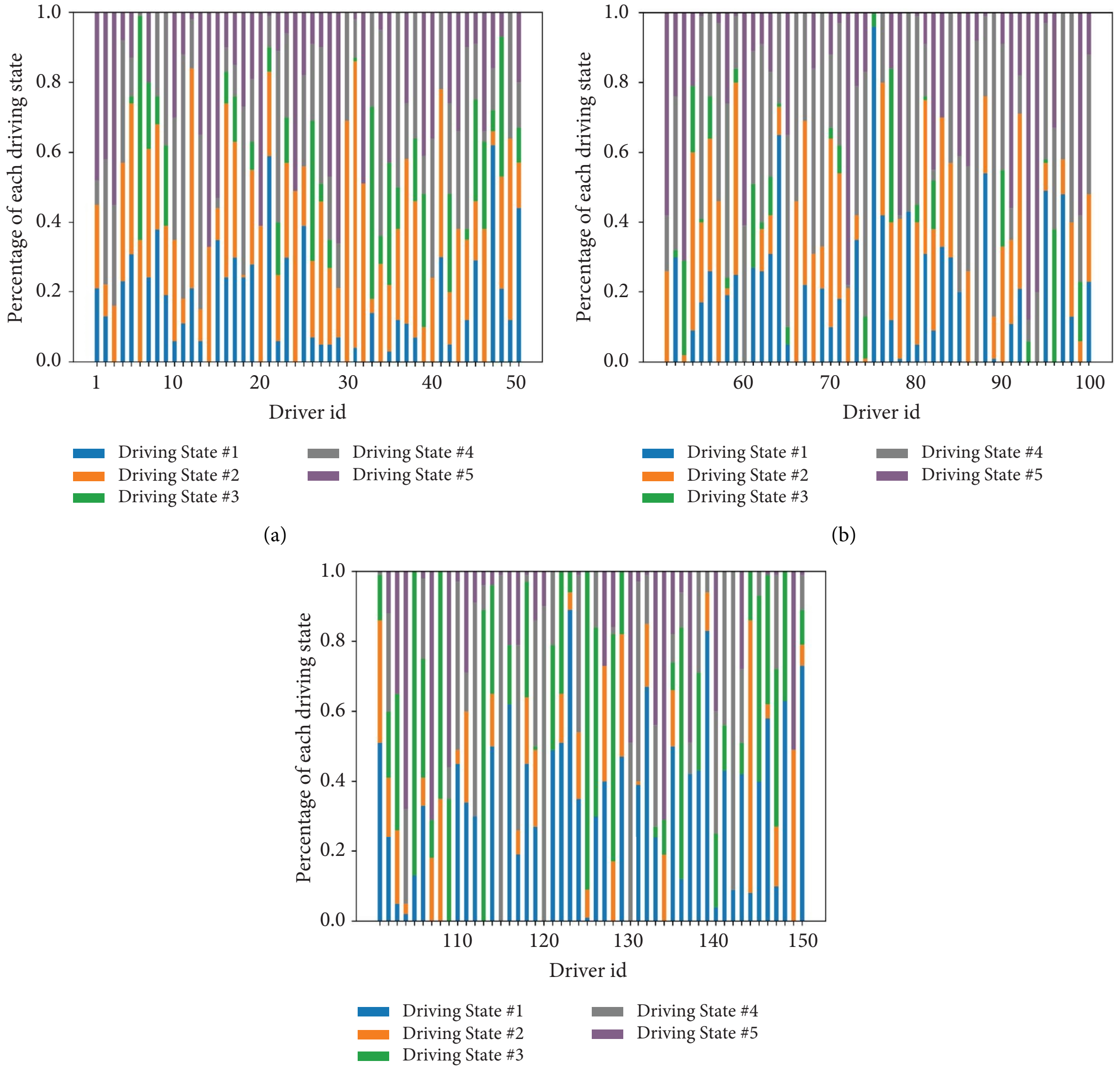
Furthermore, the influence of rainy weather on driving styles is also explored in this study by comparing Figures 13 and 14. First, it is shown in Figures 13(a) and 14(a) that the probability of driving states #3(green) and #5(purple) is greater on rainy Sunday than on sunny day. Under rainy weather, some drivers show more high-speed unstable driving and some show low-to-medium speed and stable driving.

In terms of Figures 13(b) and 14(b), the driving state #1(blue) is 0.1588 on sunny Sunday and 0.1856 on rainy Sunday, while the average proportion of driving state #4(gray) is 0.3658 on a sunny day and 0.3282 on a rainy day, meaning that the driving style on rainy nights is still dominated by low-speed driving but with increasing instability.

Comparing Figures 13(c) with 14(c), the proportion of the driving state #3 decreases from 0.278 in sunny days to 0.2266 in rainy days. The proportion of the driving state #4 is 0.1366 when sunny but increases to 0.2056 under rainy condition, meaning that some drivers modified their driving style from high speed to low-speed behavior when it rains at night but with a slight increase in stability.

Similarly, comparing Figures 12 and 15 can also draw the conclusion of the difference in the driving style between rainy and sunny days. First, we compare Figures 12(a) with 15(a). The rainy Monday morning peak is clearly much more in the driving state #2 and less in the driving state #5, showing that the overall speed increases but reduces the stability. In terms of individuals, for example, comparing drivers 151 and 301, the driving style in rainy conditions (151) consist of more of driving state #1, which means lower speed and less stability. It is in line with the overall driving style as well as the common sense.

By comparing Figures 12(b) and 15(b), it can be seen that the driving style under rainy situation has an increasing percentage of driving states #2 and #4, with the former increasing from 0.1074 to 0.2290 and the latter from 0.1872 to 0.3107. In contrast, the probability of the driving state #5 decreases when it rains, from 0.3838 to 0.1788, meaning



(a)

(b)

(c)

Figure 14: Driving style on rainy Sunday. (a) Subdataset 1-1 (morning). (b) Subdataset 1-2 (evening). (c) Subdataset 1–3 (off-peak).

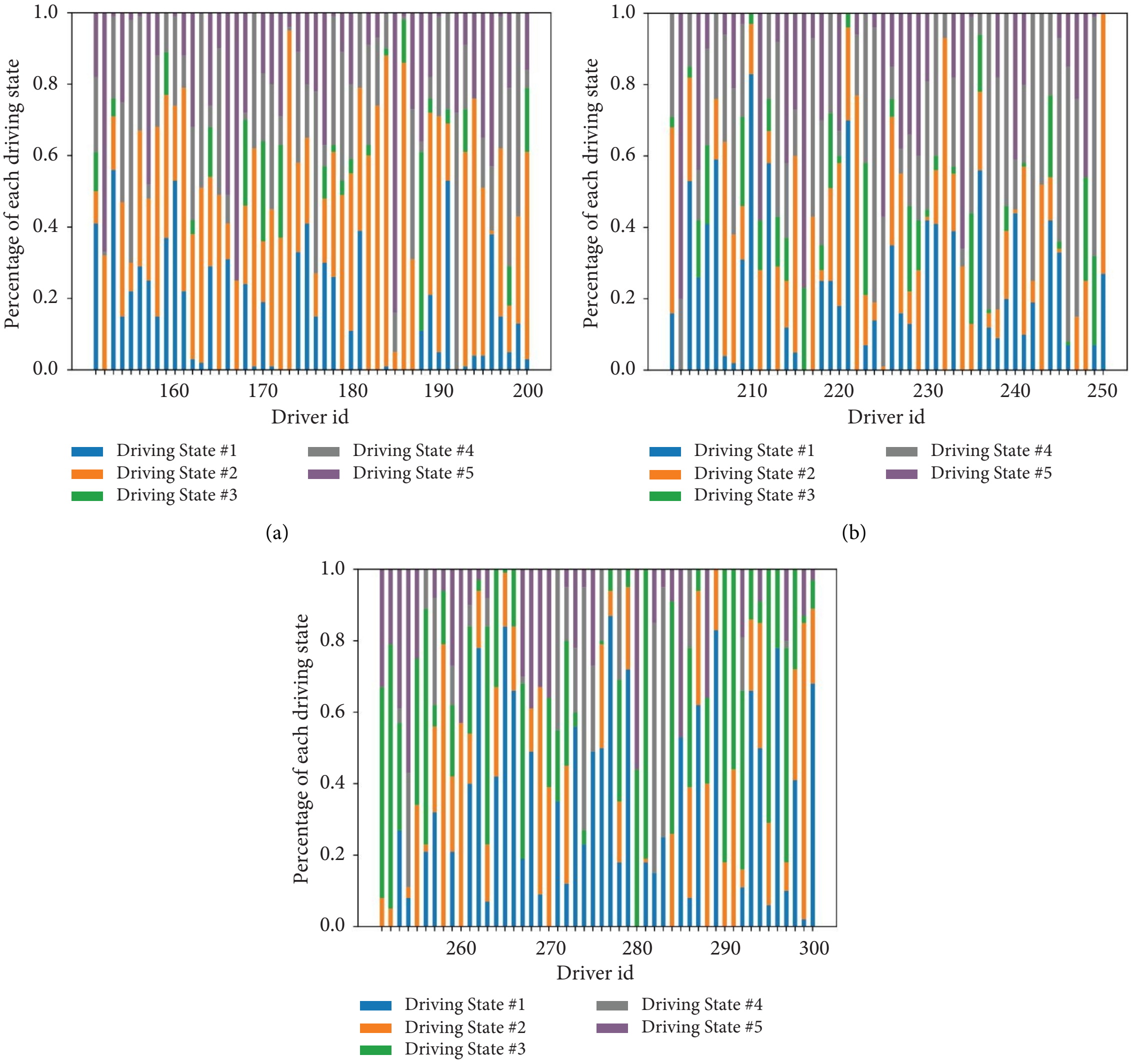
most drivers' driving stability was reduced on rainy Monday night.

Last but not least, we can see from Figures 12(c) and 15(c) that the average percentages of driving states #1 and #5 decreased significantly on rainy Monday while states #2 and #3 increased. It can be deduced that most drivers have higher speed but slightly less stability under rainy situation. Therefore, rainy weather would cause drivers to drive more unstably. This conclusion is similar to the one by comparing Figures 13 and 14.

In summary, it is concluded that peak hours have an effect on driving styles, making drivers to lower the speed and drive cautiously with smaller acceleration and jerk. It is indicated that there would be calmer and less aggressive driving styles when driving at high traffic volumes [62]. At weekends, there are multiple styles in one time slot, probably because there is not as much concentrated commuting as on weekdays. Both morning and evening peak hours have a large number of high-speed and high-stability or low-stability driving styles. The rainy weather, on the other hand, mainly affects driving stability. Comprehensively, the peak period has the most obvious and greatest influence on driving styles, which means that the driving styles difference between during peak and off-peak hours is the most significant.

5.4. Results of the Abnormal Driving Style.

It is apparent that not all driving styles have the same composition of driving states even under the same conditions of days, periods, and weather. Therefore, the different combinations of these five driving states form the unique driving style of each driver. It is worth noting that there are two aberrant types of driving styles that should be given more attention among all these styles. The first type is bold and aggressive during rush hours. For example, driving style 188 in Figure 15(a), mainly with 50% driving state #3 (green, high speed, and unstable) and 37% driving state #5 (purple, low-medium speed, and



(a)

(b)

(c)

Figure 15: Driving style on rainy Monday. (a) Subdataset 2-1 (morning). (b) Subdataset 2-2 (evening). (c) Subdataset 2-3 (off-peak).

stable), is greatly different from other styles during this rainy morning on Monday. After comparing this type with the overall driving styles in the morning peak, it can be indicated that drivers with this type of driving style tend to drive faster and more aggressively, which increases traffic instability and risk since most drivers drive cautiously because of the traffic congestion in the morning peak. For this type of driver, they should first pay more attention to the road situations in everyday driving. Second, it would be better to slow down driving speed and avoid rapid acceleration in peak hours to improve traffic safety.

Another aberrant type of driving style is conservative and cautious during the off-peak period, such as driving style 555 in Figure 13(c), where driving states #5 (purple, low-to-medium speed, and unstable) takes up to 81% of the driving style composition and state #4 (gray, low speed, and medium stable) accounts for the rest of the proportion. After comparing this type with driving style 553 in the same time slot, it is apparent that its speed decreases significantly. Although this type of driver behaves more steadily than other drivers, it would reduce traffic efficiency to a certain extend. Therefore, for the second abnormal type of driver, it would be wise to practice driving more and improving proficiency and carefulness.

By recognizing aberrant driving styles, traffic managers can reeducate these drivers with high-risky driving styles. In addition, these drivers should be given more attention in daily management. Individualized traffic proactive management measures, such as warnings, can be activated to inform drivers once they are identified to have risky driving styles during their trip.

Different from other studies that obtained driving styles by directly clustering driving behavior data [17, 19, 63, 64], the proposed method in this study shows that each driver's style is made up of multiple driving states. Human driving styles are not always the same and can probably change due to the driving environment [23, 61]. However, the approach used in our study can reflect the style of each driver throughout the driving process. Besides, this method can extract the abnormal driving styles which are different from the overall styles.

Furthermore, according to the existing literature, some researchers have also used LDA models to study driving behavior [23, 55, 65]. However, unlike these research studies, our study explores driving behavior and styles under different traffic conditions, while they basically only conducted the experiments during sunny daytime. Besides, more explicit meanings of driving styles are explored in our study than previous studies. The driving style is a person's habitual behavior in a given situation [8, 11, 20, 37]. Therefore, exploring driving styles in different traffic conditions helps to understand the influence of external factors such as traffic conditions, weather, and rush hour on driving styles. In addition, data used in this study have the advantage of wide coverage and scope, reflecting a large amount of diverse driving styles, making our results more universal.

6. Conclusion

This study identified and classified driving styles by extracting latent driving states using the LDA topic model. The k-means clustering algorithm was employed to acquire driving behavior. Then, the driving states were extracted and the driving styles of general situations and individuals under different conditions were compared. The main contributions of this research are summarized as follows:

(1) Apply the modified LDA to recognize driving styles. The LDA was employed to recognize driving styles by extracting latent driving states hidden in the large-scale GPS data. This unsupervised algorithm is an effective tool for recognizing driving styles, and five driving states were identified as well as diverse types of driving styles in this study.

(2) Compare the driving styles under different conditions. Differences between driving styles in different situations were analyzed. We first analyzed the difference occurring in the morning peak, evening peak, and off-peak hours, finding that styles are mostly conservative and cautious in the morning, free and discrete in the evening, and diverse in the off-peak hours. Then, the difference between weekdays and weekend was compared. The results showed that the driving styles tend to be more cautious and conservative on weekdays but freer on weekends. Weather factors were also examined and the results indicated that rainy days would increase the resistance of driving so that most drivers become cautious and conservative.

(3) Screen out the aberrant individual driving styles. Among all the driving styles, two aberrant styles that are negative for traffic efficiency and safety were screened out. The characteristic of the first aberrant style is its aggressiveness and instability in peak hours, which would easily lead to crashes and affect traffic safety. Another aberrant style is driving exceedingly conservatively in the off-peak period, which would have a negative impact on traffic efficiency.

There are also some limitations of this research. First, in this study, only four kinetic features were given based on previous studies. If more kinds of features, such as time series, can be taken into account and features are selected automatically by using a specific algorithm, the solving process would be more effective [66]. Besides, there could be more data with more drivers included in the analysis so that the results also make sense in more general situations. The authors recommend these two issues as follow-up research directions.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was sponsored by the National Natural Science Foundation of China (71901223 and 52102405) and the Natural Science Foundation of Hunan Province (2021JJ40746 and 2021JJ40603).

References

[1] P. Jing, G. Xu, Y. Chen, Y. Shi, and F. Zhan, "The determinants behind the acceptance of autonomous vehicles: a systematic review," *Sustainability*, vol. 12, no. 5, p. 1719, 2020.

[2] Q. Shen, Y. Ni, H. Cao, W. Qian, and G. Li, "How do vehicles make decisions during implementation period of discretionary lane change? A data-driven research," *Journal of Advanced Transportation*, vol. 2023, Article ID 2586372, 15 pages, 2023.

[3] J. Elander, R. West, and D. French, "Behavioral correlates of individual differences in road-traffic crash risk: an examination of methods and findings," *Psychological Bulletin*, vol. 113, no. 2, pp. 279–294, 1993.

[4] F. Sagberg, G. F. B. Piccinini, G. F. Bianchi Piccinini, and J. Engström, "A review of research on driving styles and road safety," *Human Factors*, vol. 57, no. 7, pp. 1248–1275, 2015.

[5] H. Jeong, W. Park, J. Lee, S. Park, and I. Yun, "Influence of public bus driver's driving behaviors on passenger fall incidents: an analysis using digital tachograph data," *Journal of Advanced Transportation*, vol. 2022, Article ID 2941327, 10 pages, 2022.

[6] Y. Li, S. Zhang, Y. Pan, B. Zhou, and Y. Peng, "Exploring the stability and capacity characteristics of mixed traffic flow with autonomous and human-driven vehicles considering aggressive driving," *Journal of Advanced Transportation*, vol. 2023, Article ID 2578690, 21 pages, 2023.

[7] L. Zhao, F. Li, D. Sun, and F. Dai, "Highway traffic crash risk prediction method considering temporal correlation characteristics," *Journal of Advanced Transportation*, vol. 2023, Article ID 9695433, 13 pages, 2023.

[8] D. J. French, R. J. West, J. Elander, and J. M. Wilding, "Decision-making style, driving style, and self-reported involvement in road traffic accidents," *Ergonomics*, vol. 36, no. 6, pp. 627–644, 1993.

[9] J. Reason, A. Manstead, S. Stradling, J. Baxter, and K. Campbell, "Errors and violations on the roads: a real distinction?" *Ergonomics*, vol. 33, no. 10-11, pp. 1315–1332, 1990.

[10] J. E. Dotse and R. Rowe, "Modelling Ghanaian road crash risk using the Manchester driver behaviour Questionnaire," *Safety Science*, vol. 139, Article ID 105213, 2021.

[11] Z. Deng, D. Chu, C. Wu, Y. He, and J. Cui, "Curve safe speed model considering driving style based on driver behaviour questionnaire," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 65, pp. 536–547, 2019.

[12] S. Jomnonkwao, S. Uttra, and V. Ratanavaraha, "Analysis of a driving behavior measurement model using a modified driver behavior questionnaire encompassing texting, social media use, and drug and alcohol consumption," *Transportation Research Interdisciplinary Perspectives*, vol. 9, Article ID 100302, 2021.

[13] S. A. Useche, B. Cendales, I. Lijarcio, and F. J. Llamazares, "Validation of the F-DBQ: a short (and accurate) risky driving behavior questionnaire for long-haul professional drivers," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 82, pp. 190–201, 2021.

[14] F. Wu, Z. Zhang, and Z. Han, "How do cognitive interventions impact driver aggressiveness in China?—a driving simulator study," *Journal of Advanced Transportation*, vol. 2023, Article ID 7300548, 10 pages, 2023.

[15] M. Treiber and A. Kesting, "An open-source microscopic traffic simulator," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 3, pp. 6–13, 2010.

[16] C. Laugier, I. E. Paromtchik, M. Perrollaz et al., "Probabilistic analysis of dynamic scenes and collision risks assessment to improve driving safety," *IEEE Intelligent Transportation Systems Magazine*, vol. 3, pp. 4–19, 2011.

[17] W. Wang and J. Xi, "A rapid pattern-recognition method for driving styles using clustering-based support vector machines," in *Proceedings of the 2016 American Control Conference (Acc)*, pp. 5270–5275, Boston, MA, USA, July 2016.

[18] W. Wang, J. Xi, A. Chong, and L. Li, "Driving style classification using a semi supervised support vector machine," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 5, pp. 650–660, 2017.

[19] J. Warren, J. Lipkowitz, and V. Sokolov, "Clusters of driving behavior from observational smartphone data," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 3, pp. 171–180, 2019.

[20] A. Aljaafreh, N. Alshabatat, and M. S. Najim Al-Din, "Driving style recognition using fuzzy logic," in *Proceedings of the 2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012)*, pp. 460–463, Istanbul, Turkey, July 2012.

[21] Y. Ma, W. Li, K. Tang, Z. Zhang, and S. Chen, "Driving style recognition and comparisons among driving tasks based on driver behavior in the online car-hailing industry," *Accident Analysis and Prevention*, vol. 154, Article ID 106096, 2021.

[22] G. Qi, Y. Du, J. Wu, and M. Xu, "Leveraging longitudinal driving behaviour data with data mining techniques for driving style analysis," *IET Intelligent Transport Systems*, vol. 9, no. 8, pp. 792–801, 2015.

[23] G. Qi, J. Wu, Y. Zhou et al., "Recognizing driving styles based on topic models," *Transportation Research Part D: Transport and Environment*, vol. 66, pp. 13–22, 2019.

[24] X. Dong, M. Zhang, S. Zhang, X. Shen, and B. Hu, "The analysis of urban taxi operation efficiency based on GPS trajectory big data," *Physica A: Statistical Mechanics and Its Applications*, vol. 528, Article ID 121456, 2019.

[25] B. Hu, X. Xia, X. Shen, and X. Dong, "Analyses of the imbalance of urban taxis' high-quality customers based on didi trajectory data," *Journal of Advanced Transportation*, vol. 2019, Article ID 3689389, 14 pages, 2019.

[26] D. Sun, K. Zhang, and S. Shen, "Analyzing spatiotemporal traffic line source emissions based on massive didi online car-hailing service data," *Transportation Research Part D: Transport and Environment*, vol. 62, pp. 699–714, 2018.

[27] D. Zhang, F. Xiao, G. Kou, J. Luo, and F. Yang, "Learning spatial-temporal features of ride-hailing services with fusion convolutional networks," *Journal of Advanced Transportation*, vol. 2023, Article ID 4427638, 12 pages, 2023.

[28] Q. Shi, M. Abdel-Aty, and J. Lee, "A Bayesian ridge regression analysis of congestion's impact on urban expressway safety," *Accident Analysis and Prevention*, vol. 88, pp. 124–137, 2016.

[29] C. N. Alam, K. Manaf, A. R. Atmadja, and D. K. Aurum, "Implementation of haversine formula for counting event visitor in the radius based on Android application," in *Proceedings of the 2016 4th International Conference on Cyber and IT Service Management*, pp. 1–6, Bandung, Indonesia, April 2016.

[30] G. Zylius, "Investigation of route-independent aggressive and safe driving features obtained from accelerometer signals," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 2, pp. 103–113, 2017.

[31] K. Wang, Y. Yang, S. Wang, and Z. Shi, "Research on car-following model considering driving style," *Mathematical Problems in Engineering*, vol. 2022, Article ID 7215697, 9 pages, 2022.

[32] X. Hu, Z. Zheng, D. Chen, X. Zhang, and J. Sun, "Processing, assessing, and enhancing the Waymo autonomous vehicle open dataset for driving behavior research," *Transportation Research Part C: Emerging Technologies*, vol. 134, Article ID 103490, 2022.

[33] M. Montanino and V. Punzo, "Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns," *Transportation Research Part B: Methodological*, vol. 80, pp. 82–106, 2015.

[34] V. C. Magana and M. Munoz-Organero, "GAFU: using a gamification tool to save fuel," *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 2, pp. 58–70, 2015.

[35] S. Yang, W. Wang, C. Lu, J. Gong, and J. Xi, "A time-efficient approach for decision-making style recognition in lane-changing behavior," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 579–588, 2019.

[36] L. Lu, S. Xiong, and Y. Chen, "A bi-level distribution mixture framework for unsupervised driving performance evaluation from naturalistic truck driving data," *Engineering Applications of Artificial Intelligence*, vol. 104, Article ID 104349, 2021.

[37] M. V. N. de Zepeda, F. Meng, J. Su, X.-J. Zeng, and Q. Wang, "Dynamic clustering analysis for driving styles identification," *Engineering Applications of Artificial Intelligence*, vol. 97, Article ID 104096, 2021.

[38] E. Zhang, H. Li, Y. Huang, S. Hong, L. Zhao, and C. Ji, "Practical multi-party private collaborative k-means clustering," *Neurocomputing*, vol. 467, pp. 256–265, 2022.

[39] L. Wang, Y. Chen, Y. Wang et al., "Identification and classification of bus and subway passenger travel patterns in Beijing using transit smart card data," *Journal of Advanced Transportation*, vol. 2023, Article ID 6529819, 15 pages, 2023.

[40] A. Likas, N. Vlassis, and J. J Verbeek, “The global K-means clustering algorithm,” *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[41] R. Lleti, M. C. Ortiz, L. A. Sarabia, and M. S. Sanchez, “Selecting variables for k-means cluster analysis by using a genetic algorithm that optimizes the silhouettes,” *Analytica Chimica Acta*, vol. 515, pp. 87–100, 2004.

[42] X. Niu, J. Zhu, C. Q. Wu, and S. Wang, “On a clustering-based mining approach for spatially and temporally integrated traffic sub-area division,” *Engineering Applications of Artificial Intelligence*, vol. 96, Article ID 103932, 2020.

[43] W. Cui and Y. Yang, “Quantum simultaneous measurement of non-commuting observables based on K-means clustering,” *Physica A: Statistical Mechanics and Its Applications*, vol. 588, Article ID 126559, 2022.

[44] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[45] M. Lienou, H. Maitre, and M. Datcu, “Semantic annotation of satellite images using latent dirichlet allocation,” *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 28–32, 2010.

[46] H. Li, S. Cryer, J. Raymond, and L. Acharya, “Interpreting atomization of agricultural spray image patterns using latent Dirichlet allocation techniques,” *Artificial Intelligence in Agriculture*, vol. 4, pp. 253–261, 2020a.

[47] X. Li, Z. Ma, P. Peng et al., “Corrigendum to Supervised latent Dirichlet allocation with a mixture of sparse softmax,” *Neurocomputing*, vol. 318, p. 306, 2018.

[48] T. Mavridis and A. L. Symeonidis, “Semantic analysis of web documents for the generation of optimal content,” *Engineering Applications of Artificial Intelligence*, vol. 35, pp. 114–130, 2014.

[49] Y. Du, Y. Yi, X. Li, X. Chen, Y. Fan, and F. Su, “Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet Allocation,” *Engineering Applications of Artificial Intelligence*, vol. 87, Article ID 103279, 2020.

[50] A. S. Bakhtiari and N. Bouguila, “A variational Bayes model for count data learning and classification,” *Engineering Applications of Artificial Intelligence*, vol. 35, pp. 176–186, 2014.

[51] A. Berquand, Y. Moshfeghi, and A. Riccardi, “SpaceLDA: topic distributions aggregation from a heterogeneous corpus for space systems,” *Engineering Applications of Artificial Intelligence*, vol. 102, Article ID 104273, 2021.

[52] W. Wang, Y. Feng, and W. Dai, “Topic analysis of online reviews for two competitive products using latent Dirichlet allocation,” *Electronic Commerce Research and Applications*, vol. 29, pp. 142–156, 2018.

[53] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: a review for statisticians,” *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[54] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[55] J. Bao, P. Liu, X. Qin, and H. Zhou, “Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data,” *Accident Analysis and Prevention*, vol. 120, pp. 281–294, 2018.

[56] M. Buhin Pandur, J. Dobsa, and L. Kronegger, “Topic modelling in social sciences: case study of web of science,” in *Proceedings of the Central European Conference on Intelligent and Information Systems*, Varazdin, Croatia, October 2020.

[57] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, February 2015.

[58] D. Yu and B. Xiang, “Discovering topics and trends in the field of Artificial Intelligence: using LDA topic modeling,” *Expert Systems with Applications*, vol. 225, Article ID 120114, 2023.

[59] R. K. Gupta, R. Agarwalla, B. H. Naik, J. R. Evuri, A. Thapa, and T. D. Singh, “Prediction of research trends using LDA based topic modeling,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 298–304, 2022.

[60] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, “Optimizing semantic coherence in topic models,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, July 2011.

[61] Y. Chen, K. Wang, and J. J. Lu, “Feature selection for driving style and skill clustering using naturalistic driving data and driving behavior questionnaire,” *Accident Analysis and Prevention*, vol. 185, Article ID 107022, 2023.

[62] S. Haghzare, J. L. Campos, K. Bak, and A. Mihailidis, “Older adults’ acceptance of fully automated vehicles: effects of exposure, driving style, age, and driving conditions,” *Accident Analysis and Prevention*, vol. 150, Article ID 105919, 2021.

[63] R. Yao and X. Du, “Modelling lane changing behaviors for bus exiting at bus bay stops considering driving styles: a game theoretical approach,” *Travel Behaviour and Society*, vol. 29, pp. 319–329, 2022.

[64] Y. Zhang, Y. Chen, X. Gu, N. N. Sze, and J. Huang, “A proactive crash risk prediction framework for lane-changing behavior incorporating individual driving styles,” *Accident Analysis and Prevention*, vol. 188, Article ID 107072, 2023b.

[65] Z. Chen, Y. Zhang, C. Wu, and B. Ran, “Understanding individualization driving states via latent dirichlet allocation model,” *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 2, pp. 41–53, 2019.

[66] L. Li, J. Zhu, H. Zhang, H. Tan, B. Du, and B. Ran, “Coupled application of generative adversarial networks and conventional neural networks for travel mode detection using GPS data,” *Transportation Research Part A: Policy and Practice*, vol. 136, pp. 282–292, 2020b.