Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 4954035, 11 pages
https://doi.org/10.1155/2023/4954035

Research Article

Regular Vehicle Spatial Distribution Estimation Based on Machine Learning

Lin Liu, Bingbing Wang, Yongfu Li, and Nenglong Hu

School of Automation, Chongqing University of Posts and Telecommunications, Chongqing, China

Correspondence should be addressed to Lin Liu; liulin@cqupt.edu.cn

Received 5 May 2023; Revised 22 July 2023; Accepted 18 August 2023; Published 30 August 2023

Academic Editor: Neng Ye

For the mixed traffic flow, obtaining the distribution of connected vehicles (CVs) and regular vehicles (RVs) is of great significance for road network analysis and cooperative control in intelligent transportation systems (ITSs). However, whether it is based on fixed sensors or based on CVs and traffic mechanism to estimate the spatial distribution of RVs, the implementation complexity and low estimation accuracy are the points that need to be improved. This paper proposes a regular vehicle spatial distribution estimation method using adjacent connected vehicles as mobile sensors. First, to investigate the hidden relationship between the interaction information of adjacent CVs and the spatial distribution of RVs among CVs, the Gaussian mixture model-hidden Markov model (GMM-HMM) is selected as the identification method. Then, three sets of experiments were designed to study the influence of observed features on the identification capability of the model, generalization capability validation, and comparison with other methods, respectively. Finally, the proposed method is verified by the dataset generated by the car-following model. The experimental results show that selecting the relative position and time headway as observed features can effectively reflect the regular vehicle spatial distribution between adjacent CVs. The average accuracy of the proposed method to identify the regular vehicle spatial distribution is over 93.7%, which can provide valuable suggestions for the Internet of Vehicles application.

1. Introduction

ITSs have developed rapidly in recent years because of their great potential for improving traffic flow characteristics, such as solving existing traffic congestion problems, low safety, and low resource utilization. As an essential part of the deployment of ITS, CVs are known as vehicles that can share information (such as position, velocity, and acceleration) with other CVs by using the vehicle to everything (V2X) technology. Therefore, a safer, more efficient, and energy-saving road network is created. However, the penetration of CVs will not be entirely popularized in the short term. At this stage, CVs and RVs will coexist on the roads, and a mixed traffic flow will be emerging [1]. At the same time, many researchers have conducted in-depth research on the characteristics of the mixed traffic flow, such as the queue safety evaluation [2], capacity analysis [3], and vehicle to road cooperative control optimization [4].

These studies explain the critical role of CVs in the overall improvement of traffic flow characteristics. However, due to the existence of RVs, the application performance of CVs is inevitably limited. Therefore, estimating the distribution of RVs and CVs is indispensable for deploying ITS (e.g., analyzing road networks and achieving traffic optimization control). Traditional traffic flow state estimation methods widely use fixed sensors [5], such as loop detectors, cameras, and other equipment, to monitor the road traffic flow and vehicle state information. However, these methods have limitations, such as fixed monitoring positions and high installation and maintenance costs. Unlike fixed sensors, CVs can interact with roads, vehicles, and cloud platforms. With the advantages of high flexibility and low cost, it has become a reality for CVs to collect data as mobile sensors for traffic flow analysis [6]. Reference [7] analyzed the characteristics of a mixed traffic flow with the maximum platoon size of CAVs. The conclusion shows that the traffic benefit does not increase all the time when the vehicle

platoon reaches a certain level. References [8, 9] pointed out that there are three spatial distributions of CVs and RVs in the queue. In the first case, CVs are concentrated; the road traffic efficiency is the highest, and the safety is the best. In the second case, CVs and RVs are uniformly distributed on the road, the road traffic efficiency is the lowest, and the safety is the worst. CVs and RVs are randomly distributed in the third case, and the traffic flow characteristics are between the previous two cases. Reference [10] proposed a method for estimating the traffic state and market penetration of CVs on highways. CVs and roadside units are used as mobile and fixed sensors to form a hybrid sensor. A filtering approach is used to estimate the traffic state under the mixed traffic flow. Reference [11] used the Markov chain to prove that the orderly arrangement of CVs significantly increased the road capacity. At the same time, relative entropy was introduced to quantitatively describe the orderliness of the mixed traffic flow, and the root cause of the improvement of road traffic capacity by CVs was clarified. Reference [12] pointed out that the emergence of RVs inhibits the formation of CV queues, which is not conducive to realizing cooperative driving. The discrete hidden Markov method is used to estimate the number of RVs in adjacent CVs. However, the discretization process of this method is prone to problems such as quantization error and signal distortion, which weakens the resolution of the model, so the recognition accuracy needs to be improved.

Overall, there are two areas for improvement in the existing methods for estimating the state of mixed traffic flows. First, they rely on fixed sensors to detect the road flow and vehicle status, and it is difficult to identify the specific distribution of CVs and RVs. Second, some of the methods start from studying traffic flow mechanism characteristics, and the theoretical techniques and implementation are relatively complicated, and the accuracy needs to be improved.

Aiming to address the problem of the regular vehicle spatial distribution estimation without introducing complex statistical derivation processes or adding other monitoring equipment, this paper proposes a method to estimate the spatial distribution of RVs by using the interaction information of adjacent CVs. Based on the concept of data driven, by analyzing the internal mechanism of the spatial distribution of RVs and the information interaction of CVs, a GMM-HMM model is established by taking the relative position and the time headway of adjacent CVs as the input and the spatial distribution of RVs as the output. Theoretical derivation and numerical simulation verify the effectiveness of this method.

2. Modeling of Regular Vehicle Spatial Distribution Estimation

2.1. Mechanistic Analysis. CVs and RVs are most likely to be driven by humans at this stage. CVs can use V2X to share vehicle driving status information within a certain communication range. At the same time, it reduces the reaction delay time of the driver's decision making during driving so that the vehicle can drive on the road with minor headway and spacing. For the mixed traffic flow composed of CVs and RVs, the spatial distribution of RVs and the information of CVs are spatially and temporally correlated, as shown in Figure 1.

In Figure 1, within a certain communication range, CVs interact with each other for driving information, such as position $x_n(t), x_{n-1}(t)$, velocity $v_n(t), v_{n-1}(t)$, and acceleration $a_n(t), a_{n-1}(t)$) at different times. This time-varying information is closely related to the spatial distribution of RVs (the number of RVs in the green dotted box in Figure 1) and can be collected for further processing. The relative position $\Delta x_n(t)$, relative velocity $\Delta v_n(t)$, relative acceleration $\Delta a_n(t)$, and time headway $T_h(t)$ of adjacent CVs are defined as follows:

$$\Delta x_n(t) = x_{n-1}(t) - x_n(t), \tag{1}$$

$$\Delta v_n(t) = v_{n-1}(t) - v_n(t), \tag{2}$$

$$\Delta a_n(t) = a_{n-1}(t) - a_n(t), \tag{3}$$

$$T_h(t) = \frac{x_{n-1}(t) - x_n(t)}{v_n(t)}, \tag{4}$$

where $x_{n-1}(t)$, $v_{n-1}(t)$, and $a_{n-1}(t)$ denote the position, velocity, and acceleration of the leading connected vehicle, respectively; $x_n(t)$, $v_n(t)$, and $a_n(t)$ denote the position, velocity, and acceleration of the following connected vehicle, respectively.

Using different combinations of the abovementioned four features as the observed features (the dimensionality is determined by the number of selected features), it is estimated that the regular vehicle spatial distribution can be implemented by using machine learning methods.

2.2. Model Construction. Before introducing the model of our paper, we present some key assumptions to facilitate the modeling process.

(1) We consider only the longitudinal behavior of all kinds of vehicles. That is, the behavior of vehicle changing the lane is not considered.

(2) The network information is reliable, and the transmission delay is ignored. All drivers fully obey the advanced driving assistance suggestions.

(3) All kinds of vehicles are driven by humans, regardless of the existence of automatic drive.

2.2.1. Methodology. The extracted features of the information of CVs are continuous and generally present a Gaussian distribution. To estimate the hidden regular vehicle spatial distribution from the visible observed features, this paper uses the GMM-HMM [13] as the identification method.

In this paper, the spatial distribution of RVs between adjacent CVs is regarded as the hidden state to be identified. More specifically, there are 0 RV, 1 RV, and at least 2 RVs

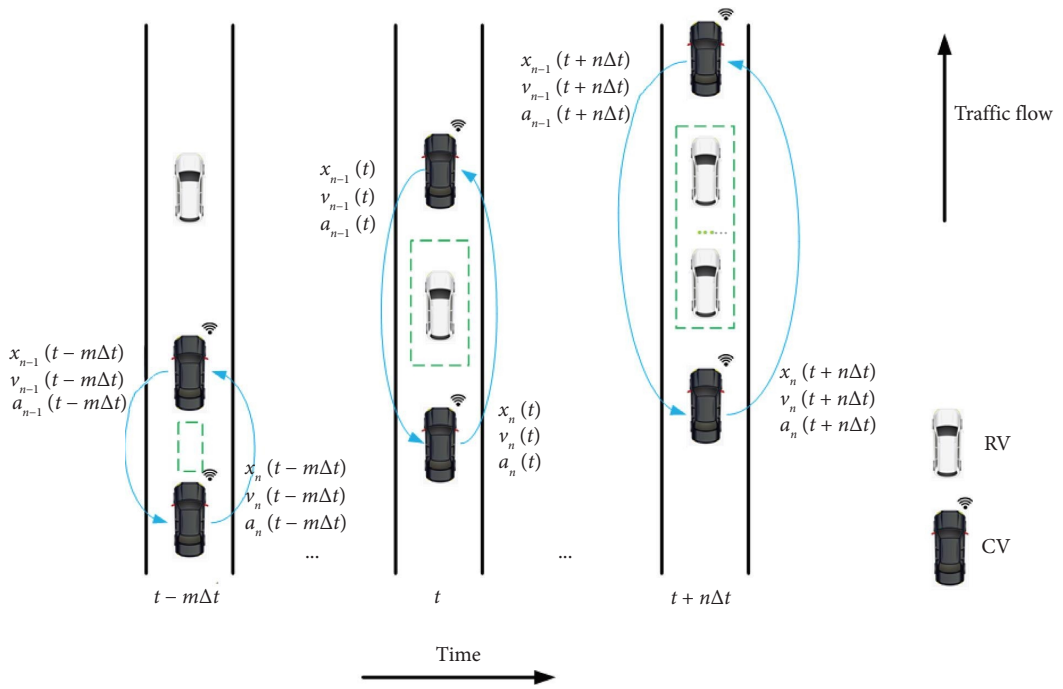


Figure 1: The information of CVs corresponds to RV spatial distribution at different times.

between adjacent CVs represented by hidden states q_3, respectively. The feature information of adjacent CVs (different combinations of relative position, relative velocity, relative acceleration, and time headway) is used as the observed features. The transition of the hidden state at different moments is described by the state transition probability, and the mapping relationship between the observed features and the hidden state is described by the output probability, as shown in Figure 2.

GMM-HMM is composed of the initial probability vector $\boldsymbol{\pi}$, the state transition matrix $\mathbf{A}$, and the output probability matrix $\mathbf{B}$, which is represented by a triple symbol $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$. When the hidden state sequence $\mathbf{I} = (i_1, i_2, \ldots, i_T)$ and the observation feature sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$ of the model are given, the initial probability vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ with N components, which satisfies the following equation:

$$\pi_j = P\left(i_1 = q_j\right), \quad j = 1, 2, \ldots, N. \tag{5}$$

The state transition matrix $\mathbf{A} = [a_{ij}]_{N\times N}$, which describes the probability of transferring from the hidden state q_i at time t to the hidden state q_j at time $t+1$, satisfies the following equation:

$$a_{ij} = P\left(i_{t+1} = q_j \,\middle|\, i_t = q_i\right), \quad i = 1, 2, \ldots, N. \tag{6}$$

The output probability matrix $\mathbf{B} = [b_j(k)]_{N\times k}$ refers to the mapping relationship between the hidden state value q_i and the observed feature $\mathbf{v_k}$ at any time which satisfies the following equation:

$$b_j(k) = P\left(\mathbf{o_t} = \mathbf{v_k} \,\middle|\, i_t = q_j\right). \tag{7}$$

Since the observed feature $\mathbf{v_k}$ is continuous, the multidimensional mixed Gaussian distribution describes the joint probability distribution between the hidden state value q_i and the observed feature $\mathbf{v_k}$. Equation (7) can be rewritten as follows:

$$b_j(k) = \sum_{m=1}^{M} c_{jm} N\left(o_t \,\middle|\, u_{jm}, \sum_{jm}\right),$$

$$\sum_{m=1}^{M} c_{jm} = 1,\ 0 \le c_{jm} \le 1, \tag{8}$$

$$N\left(\mathbf{o_t} \,|\, \mathbf{u_{jm}}, \boldsymbol{\Sigma}_{\mathbf{jm}}\right) = \frac{1}{(2\pi)^{(D/2)}\left|\boldsymbol{\Sigma}_{\mathbf{jm}}\right|^{(1/2)}} \cdot \exp\left[-\frac{1}{2}\left(\mathbf{o_t} - \mathbf{u_{jm}}\right)^T \boldsymbol{\Sigma}_{\mathbf{jm}}^{-1}\left(\mathbf{o_t} - \mathbf{u_{jm}}\right)\right],$$

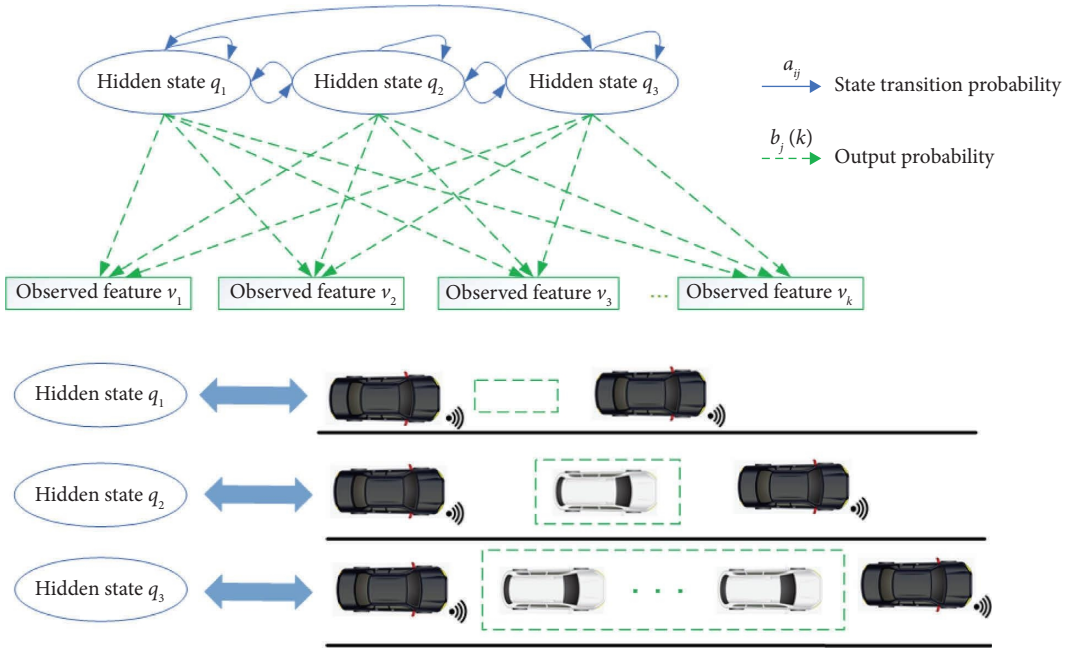


Figure 2: The structure of GMM-HMM.

where c_{jm} is the weight coefficient of the $m-$th Gaussian distribution in the GMM when it is in the hidden state q_j; M is the number of Gaussian components; D denotes the dimensionality of the observed random variable $\mathbf{o_t}$; and $\mathbf{u_{jm}} \in \mathbb{R}^{D\times 1}$ and $\mathbf{\Sigma_{jm}} \in \mathbb{R}^{D\times D}$ are the mean vector and the covariance matrix of Gaussian distribution $N(\mathbf{o_t} \mid \mathbf{u_{jm}}, \mathbf{\Sigma_{jm}})$, respectively.

2.2.2. Model Training and Testing Process. The number of hidden states N and Gaussian components M in the GMM-HMM is regarded as hyperparameters. The remaining model parameters need to be trained by EM (expectation-maximum) algorithm through massive historical data. The model training and testing process is shown in Figure 3.

The EM algorithm is used to estimate the model parameters, including initial state probability distribution $\hat{\boldsymbol{\pi}}_{\mathbf{j}}$, state transition probability $\hat{a}_{\mathrm{ij}}$, mixed Gaussian distribution weight coefficient $\hat{c}_{\mathrm{jm}}$, mean vector $\hat{\mathbf{u}}_{\mathbf{jm}}$, and covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{jm}}$. After the abovementioned training process, the model parameters $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ are obtained. Given the observation feature sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$, the Viterbi algorithm can be used to determine the hidden state sequence $\mathbf{I} = (i_1, i_2, \ldots, i_T)$.

3. Experimental Verification

3.1. Data Preparation and Processing. The basic dataset used in this paper comes from NGSIM (next generation simulation) [14]. From this dataset, 50 vehicle trajectories with recorded position, speed, and acceleration information were randomly extracted from the I-80 section. Due to some noise and errors in the raw data, the vehicle position, velocity, and acceleration are preprocessed using a moving average filter. The original data and filtered data are shown in Figure 4.

To obtain the characteristic data of the combination of CVs and RVs required for the study using the theory proposed in Reference [15], the car-following models of CVs and RVs are considered IDM [16] with different response delay times. The acceleration of the following vehicle n at time $t+\tau$ satisfies the following equations:

$$a_n(t+\tau) = a\left[1 - \left(\frac{v_n(t)}{v_f}\right)^{\sigma} - \left(\frac{s^*(v_n(t), \Delta v_n(t))}{s_n(t)}\right)^2\right],$$

$$s^*(v_n(t), \Delta v_n(t)) = s_0 + v_n(t)T - \frac{v_n(t)\Delta x_n(t)}{2\sqrt{ab}}, \tag{9}$$

where a_n and v_n denote the acceleration and velocity of the vehicle n, respectively; $s^*(\cdot)$ is the desired minimum gap; $\Delta v_n = v_{n-1} - v_n$ and $\Delta x_n = x_{n-1} - x_n$ denote the speed difference and position difference between the leading vehicle $n-1$ and the following vehicle n, respectively; a and b denote the maximum acceleration and expected deceleration of the vehicle n, respectively; v_f is the desired velocity; and s_0 and T denote the jam gap and safe time headway, respectively.

The abovementioned 50 vehicles are the leading vehicles, and the sampling time is set to 0.1 s. When the market penetration rate (MPR) of CVs is 0.3, 0.5, and 0.7, the feature datasets (the corresponding sample sizes are 36400 groups, 72800 groups, and 109200 groups, respectively) of 500 mixed CVs and RVs are generated, respectively. In the mixed traffic flow, the spatial distribution of CVs and RVs is random. It should be noted that only if both the following and the leading vehicle are CVs, the following vehicle can apply the car-following model of the CVs. Otherwise, the RV car-following model is used. The difference between the two is in the driver's reaction delay time, and the CV has a lower time delay than the RV. The parameters are given in Table 1.

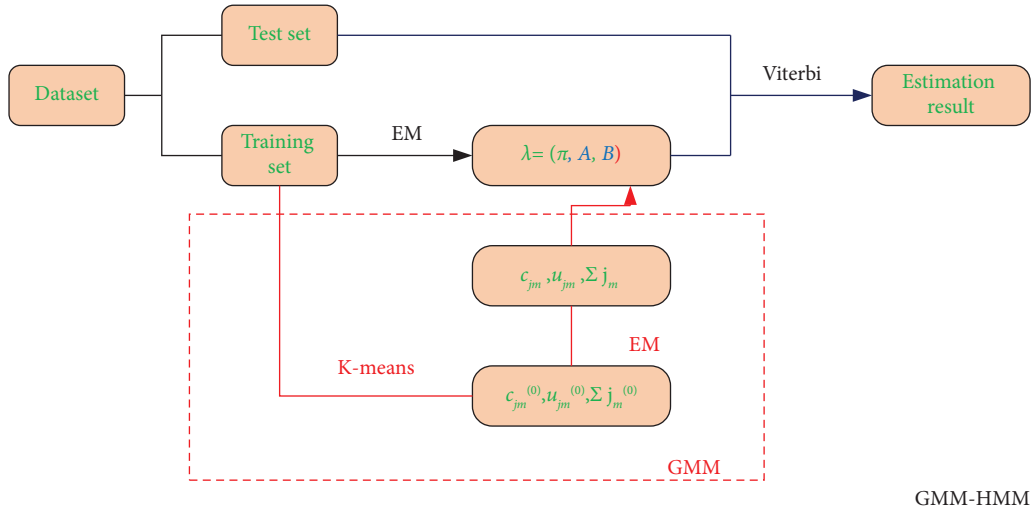


Figure 3: The training and testing process of GMM-HMM.

The database of the observed features required for model training and testing is obtained by recording the information of CVs at each simulation step and then processing the information using equations (1)–(4).

3.2. Experiment Setting. The model parameters need to be initialized before using the GMM-HMM. Among hyperparameters, the number of Gaussian components $M = 2$ and the number of hidden states $N = 3$. The initial probability vector $\boldsymbol{\pi}$ is uniformly distributed. The mean vector and the covariance matrix in the state transition matrix **A** and the output probability matrix **B** are generated by the initialization of the k-means algorithm. The number of iterations is set to 50, and the convergence threshold is set to 1×10^{-4}. The classification model evaluation method proposed in reference [17] was used to evaluate the model performance. Accuracy (ACC), macroaverage precision (MAP), macroaverage recall (MAR), class balance accuracy (CBA), and the F1 score (F1), which integrates accuracy and recall, were used as indicators to evaluate the model performance. At the same time, to make the evaluation results of the model convincing and effectively avoid overfitting and underfitting, k-cross validation is adopted to use all the data for training and testing.

In experiment 1, to determine the effect of different observed features on the recognition ability of the GMM-HMM, five groups of observed features are used as the model's input to obtain the state estimation results of the GMM-HMM. The five different observed features are as follows:

(1) Relative position+relative velocity ($\Delta x_n + \Delta v_n$)
(2) Relative position+time headway ($\Delta x_n + T_h$)
(3) Relative position+relative acceleration ($\Delta x_n + \Delta a_n$)
(4) Relative position+relative velocity+relative acceleration ($\Delta x_n + \Delta v_n + \Delta a_n$)
(5) Relative position+relative velocity+relative acceleration+time headway ($\Delta x_n + \Delta v_n + \Delta a_n + T_h$)

In experiment 2, to verify the adaptability of the model under different MPR environments, the mixed traffic flow dataset generated at a certain MPR is taken as the training set. The data generated from the remaining MPR are used as the test set (e.g., the features data obtained at an MPR of 0.5 for the CVs are used as the training set, and the feature data at MPRs of 0.3 and 0.7 are used as the test set) to verify the generalization ability of the model.

In experiment 3, GMM-HMM is briefly compared with other machine learning methods, such as the support vector machine (SVM) and artificial neural network (ANN) for RV spatial distribution estimation. The same dataset is used for training and testing to make the experimental results relatively fair. The parameters of both SVM and ANN are selected by constantly adjusting the grid search method.

3.3. Results and Discussion

3.3.1. Determine the Optimal Observed Features. In this paper, different combinations of relative position Δx_n, relative velocity Δv_n, relative acceleration Δa_n, and time headway T_h of adjacent CVs are extracted as observed features. In order to estimate the spatial distribution of RVs, it is necessary to combine the four features in the process of establishing the GMM-HMM. The distributions of Δx_n, Δv_n, Δa_n, and T_h under different hidden states are shown in Figure 5.

It can be directly seen from Figure 5 that the spatial distribution of RVs between adjacent CVs (the hidden state) can be reflected by the information of CVs. There are differences in the probability distribution of information under different hidden states. The characteristic information of network connection varies greatly in different hidden states. This point is shown in the figure as "thin" and "fat" degrees are different. The fundamental reason is that the mean value and the standard deviation of the network characteristic information are different. Statistical methods are used to describe the distribution of information, as shown in Table 2.

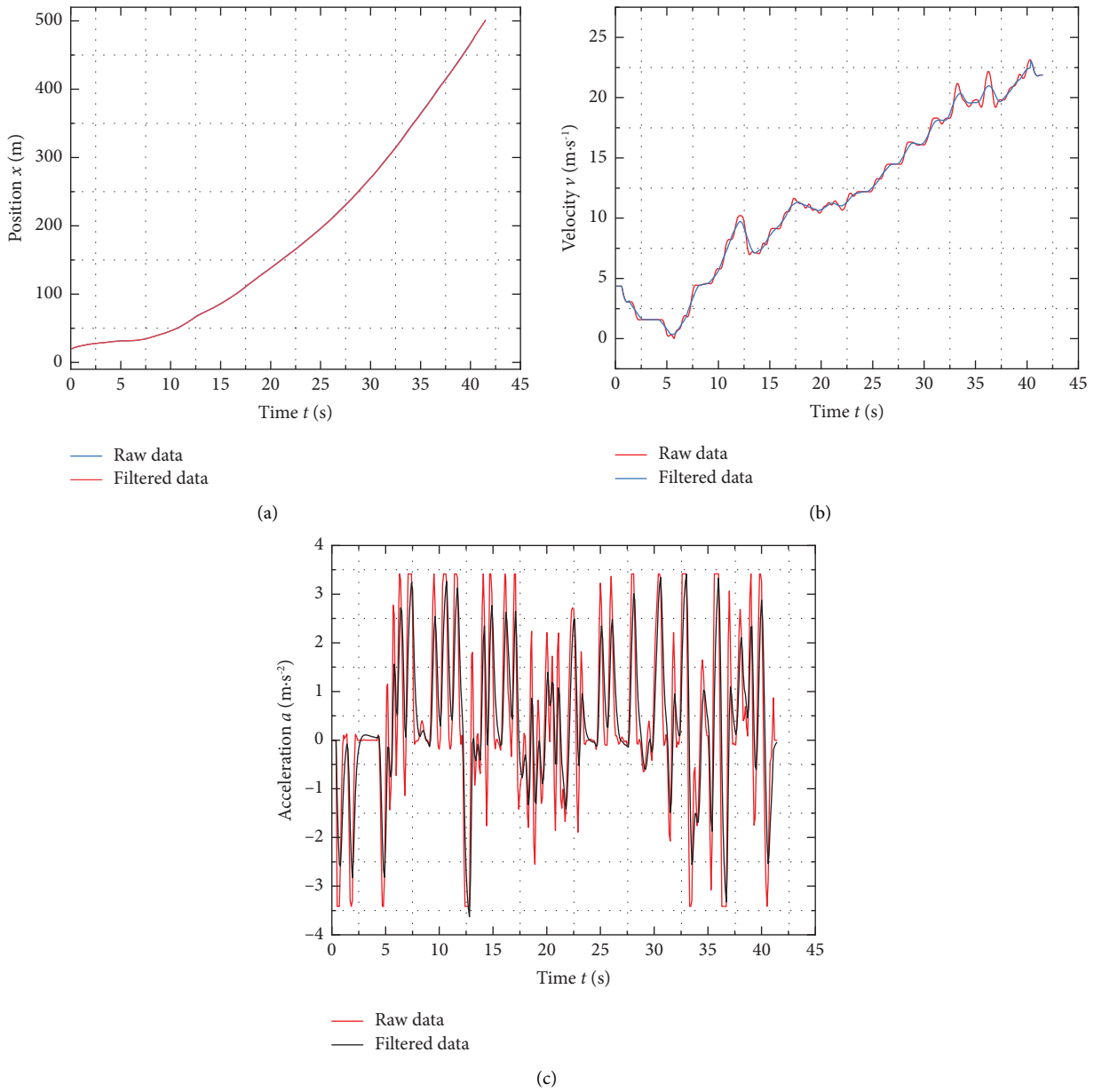


Figure 4: Comparison of the original trajectory data and the filtered trajectory data of the vehicle. (a) Vehicle position data. (b) Vehicle velocity data. (c) Vehicle acceleration data.

Table 1: Parameters setting of the car-following model for CVs and RVs [15].

Parameters	CVs	RVs
Desired velocity v_f (m/s)	33	33
Jam gap s_0 (m)	2	2
Safe time headway T (m)	1.4	1.4
Maximum acceleration a (m/s^2)	4	4
Acceleration exponent σ	2	2
Expected deceleration b (m/s^2)	2	2
Time delay τ (s)	0	0.4

In this section, five different observed features were selected to train and test the model. The five evaluation indices (the evaluation indices take values ranging from 0 to 1, with values closer to 1 indicating better model performance) derived from the confusion matrix are used to further analyze the model's performance. The results are shown in Figure 6.

Figure 6 shows the effect of selecting different observed features on the model performance. It can be observed that the relative position Δx_n and time headway T_h of adjacent CVs are selected as the input of the model, which can make GMM-HMM have the best effect on the spatial distribution estimation of RVs. More specifically, the combination of the relative position Δx_n and time headway T_h as the observed features can effectively reflect the spatial distribution of RVs. In the cross-validation experiment, the accuracy reached 0.972 in the best case, 0.891 in the worst

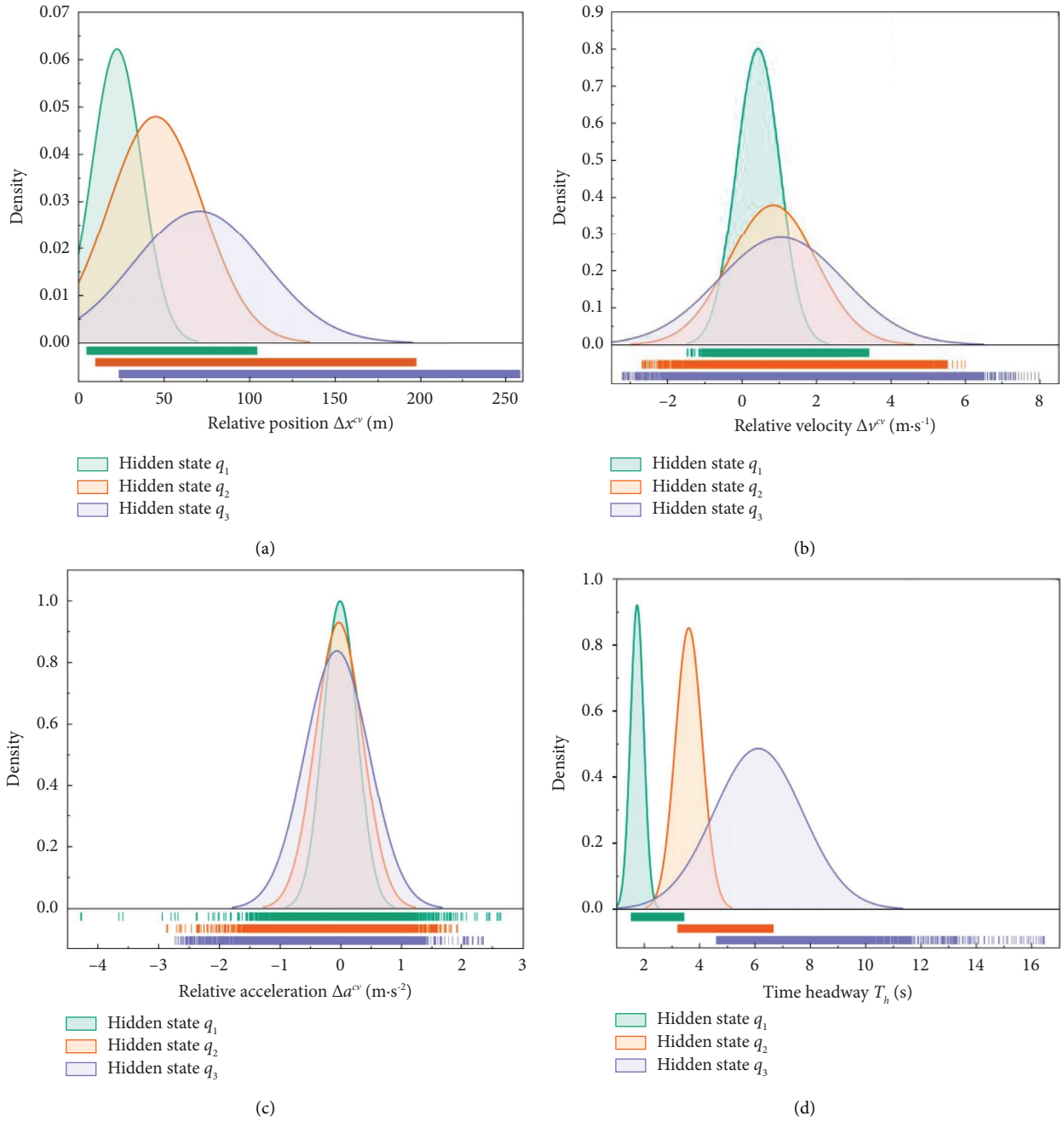


FIGURE 5: The distribution of information. (a) The distribution of relative position. (b) The distribution of relative velocity. (c) The distribution of relative acceleration. (d) The distribution of time headway.

case, and the average accuracy was about 0.937. Considering the combination of relative position Δx_n, relative velocity Δv_n, relative acceleration Δa_n, and time headway T_h as the observed features, the accuracy of the former is slightly lower than that of the former. The accuracy of recognition using the remaining three observed features is much lower than that of the previous two. Similarly, from the perspective of recall, precision, class balance accuracy, and the F1 score, the combination of relative position Δx_n and time headway T_h is selected as the observed features, which makes the model performance better than the other four.

The model's performance using different observed features is shown in Table 3. The results show that the combination of the relative position Δx_n and time headway T_h is selected as the observed features of the GMM-HMM, which can effectively realize the spatial distribution estimation of RVs.

3.3.2. Generalization Capability Validation. To verify the adaptability of the proposed method in different environments, the observed features dataset extracted under a certain MPR is used for training, and the test dataset is from

Table 2: The statistics of information.

Information of CVs	Means			Standard deviations		
	Hidden state q_1	Hidden state q_2	Hidden state q_3	Hidden state q_1	Hidden state q_2	Hidden state q_3
Relative position Δx_n (m)	22.424	45.070	70.505	3.810	5.258	6.185
Relative velocity Δv_n (m/s)	0.416	0.818	1.052	0.762	1.081	1.296
Relative acceleration Δa_n (m/s^2)	−0.007	−0.025	−0.059	0.526	0.620	0.727
Time headway T_h (s)	1.737	3.605	6.118	0.489	0.694	1.265

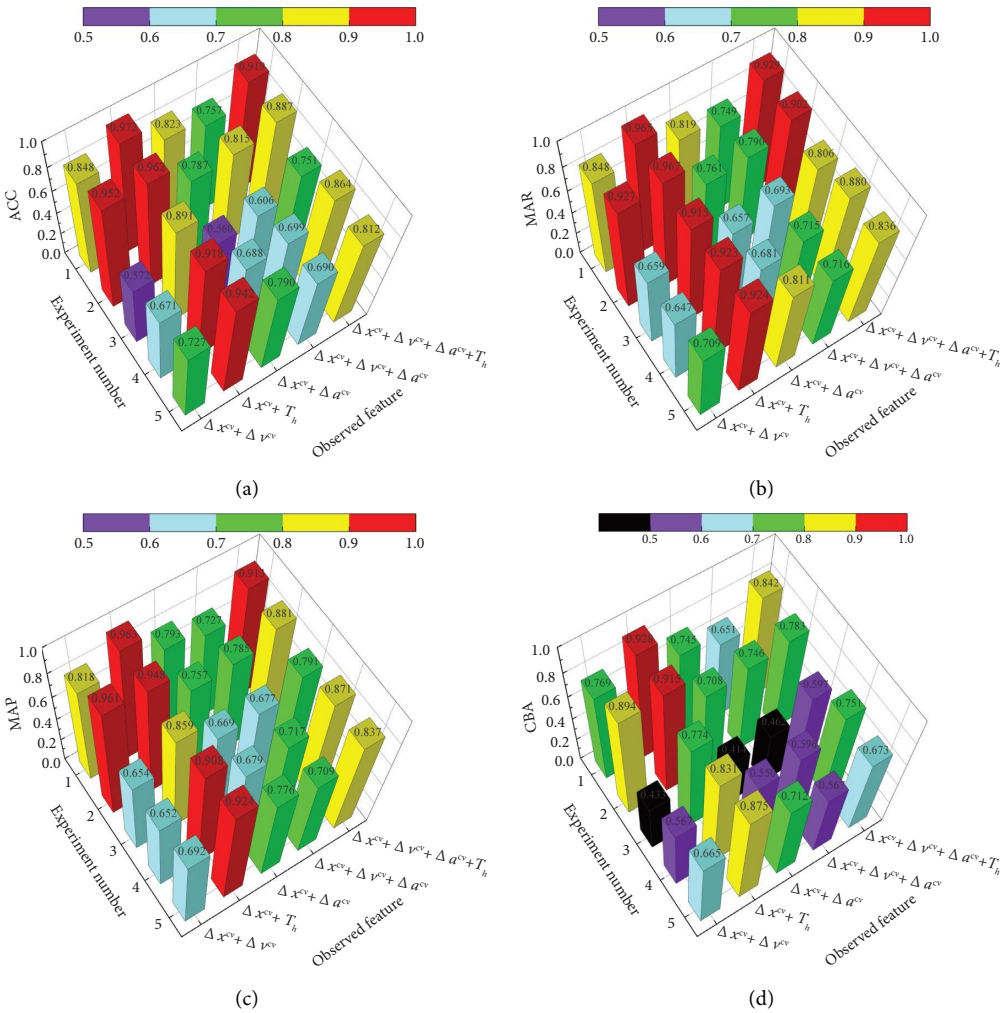
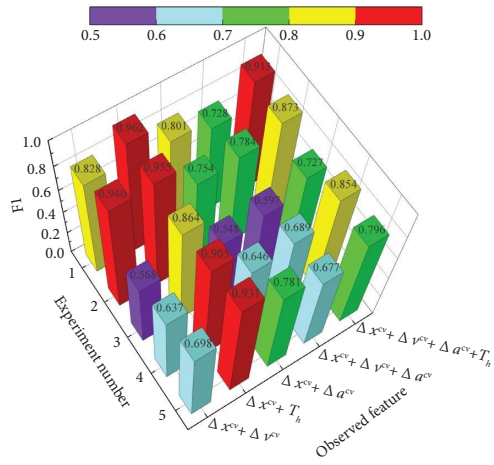


Figure 6: Continued.



(e)

Figure 6: Model performance is affected by the observed feature value. (a) ACC. (b) MAR. (c) MAP. (d) CBA. (e) F1.

Table 3: The average effect of different observed features on model performance.

Observed features	ACC	MAR	MAP	CBA	F1
Relative position + relative velocity $(\Delta x_n + \Delta v_n)$	0.754	0.758	0.756	0.666	0.734
Relative position + time headway $(\Delta x_n + T_h)$	0.937	0.944	0.921	0.865	0.924
Relative position + relative acceleration $(\Delta x_n + \Delta a_n)$	0.730	0.746	0.735	0.626	0.706
Relative position + relative velocity + relative acceleration $(\Delta x_n + \Delta v_n + \Delta a_n)$	0.714	0.732	0.723	0.605	0.695
Relative position + relative velocity + relative acceleration + time headway $(\Delta x_n + \Delta v_n + \Delta a_n + T_h)$	0.847	0.871	0.859	0.729	0.832

Table 4: The generalization ability in different environments.

Training environment	Testing environment	Hidden state q_1	Hidden state q_2	Hidden state q_3	Average accuracy
MPR = 0.3	MPR = 0.5	0.938	0.887	1.0	0.933
	MPR = 0.7	0.930	0.891	1.0	0.924
MPR = 0.5	MPR = 0.3	0.965	0.882	1.0	0.963
	MPR = 0.7	0.978	0.834	1.0	0.942
MPR = 0.7	MPR = 0.3	0.887	0.878	1.0	0.939
	MPR = 0.5	0.965	0.819	1.0	0.918

other MPR environments. In this section, the combination of relative position Δx_n and time headway T_h is used as the input of the model and the accuracy of the model to identify each hidden state is obtained. The results are shown in Table 4.

The results show that the overall accuracy of the proposed method in different MPR environments is above 0.918. It shows that the model also has good adaptability in different MPR environments.

3.3.3. Comparison with Other Methods. After using GMM-HMM to obtain the regular vehicle spatial distribution estimation results, the same dataset is also used to train and test SVM and ANN in the comparison experiment. The estimation results are shown in Figure 7.

Figure 7 shows the significant advantages of using GMM-HMM for regular vehicle spatial distribution estimation. Figure 7(a) shows the true labels of the 3 hidden states. It can be seen that the estimated values of GMM-HMM in Figure 7(b) are very close to the true labels. The regular vehicle spatial distribution estimation results using SVM and ANN are shown in Figures 7(c) and 7(d), respectively. However, these two methods' estimated values differ from the real labels. The results of evaluating the spatial distribution of RVs using different methods are shown in Table 5.

The experimental results show that GMM-HMM has the highest accuracy in estimating the spatial distribution of CVs, with a value of about 0.937. When using SVM and ANN, the accuracy is 0.934 and 0.879, respectively. As for the stability of the model, the class balance accuracy and F1

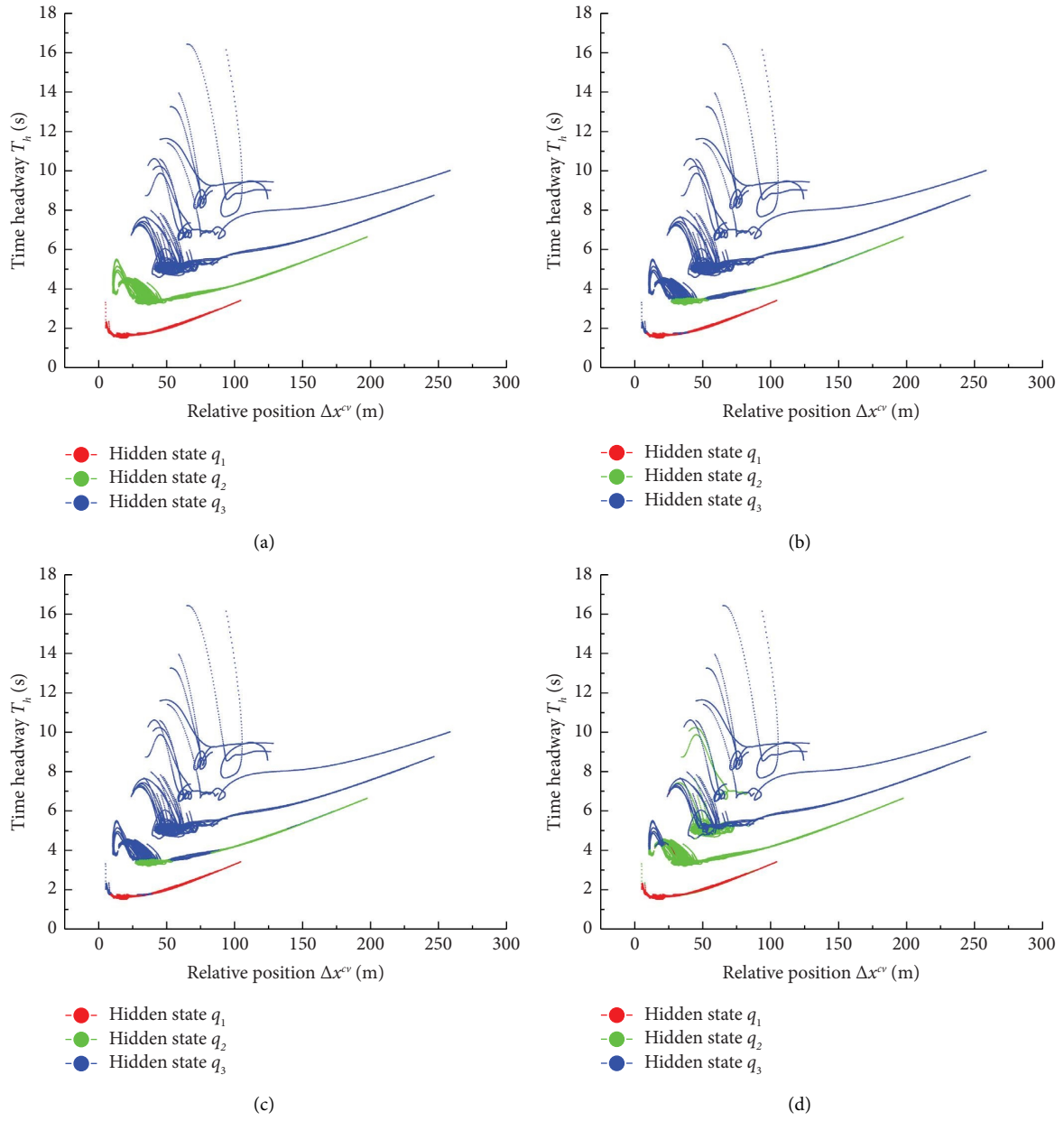


Figure 7: Estimated effect of different methods of regular vehicle spatial distribution. (a) Real labels. (b) Estimated effect of using GMM-HMM. (c) Estimated effect of using SVM. (d) Estimated effect of using ANN.

Table 5: Comparison of different identification methods.

Methods	ACC	MAR	MAP	CBA	F1
GMM-HMM	0.937	0.944	0.921	0.865	0.924
SVM	0.934	0.889	0.937	0.840	0.899
ANN	0.879	0.816	0.900	0.741	0.812

score of GMM-HMM reach 0.865 and 0.924, respectively, which are higher than those of SVM (the class balance accuracy and F1 score are 0.840 and 0.899, respectively) and ANN (the class balance accuracy and F1 score are 0.741 and 0.812, respectively).

4. Conclusion

In this paper, GMM-HMM is used as the identification method to solve the problem of regular vehicle spatial distribution estimation in the mixed traffic flow.

The spatial distribution of traditional vehicles is regarded as a hidden state, and the different combinations of the relative position Δx_n, relative velocity Δv_n, relative acceleration Δa_n, and time headway T_h of adjacent CVs are used as observed features. From the perspective of sample distribution, the relative position Δx_n and time headway T_h in different hidden states are quite different. It shows that using these two kinds of information can better establish the relationship between the hidden state and the observed features, and this view is verified in the experiments. The experimental results show that using the combination of the relative position Δx_n and time headway T_h as the observed features, the average accuracy is 0.937, the class balance accuracy is 0.865, and the F1 score is 0.924, which is higher than the other four observed features. To verify the model's generalization ability, this paper uses data in different MPR environments for training and testing. The experimental results show that the overall recognition accuracy is above 0.918, and the model can adapt to different MPR environments. In addition, GMM-HMM is briefly compared with SVM and ANN, and the experimental results show the effectiveness of the proposed method.

Data Availability

The datasets used to support the finding of this study are publicly available and can be downloaded from the following website: https://ops.fhwa.dot.gov/trafficanalysi.stools/ngsim.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. Bansal and K. M. Kockelman, "Forecasting Americans' long-term adoption of connected and autonomous vehicle technologies," *Transportation Research Part A: Policy and Practice*, vol. 95, no. 1, pp. 49–63, 2017.

[2] D. Tian, J. Zhou, Y. Wang, Z. Sheng, H. Xia, and Z. Yi, "Modeling chain collisions in vehicular networks with variable penetration rates," *Transportation Research Part C: Emerging Technologies*, vol. 69, no. 1, pp. 36–59, 2016.

[3] Y. Qin, Y. Zhu, L. Zhu, and H. Tang, "Capacity analysis method of mixed flow with connected and automated truck platooning," *Journal of Transportation Systems Engineering and Information Technology*, vol. 22, no. 4, pp. 275–282, 2022.

[4] T. Li, Y. Luo, Y. Meng et al., "Sexual activity and related factors of older women in hunan, China: a cross-sectional study," *The Journal of Sexual Medicine*, vol. 19, no. 2, pp. 302–310, 2022.

[5] M. Sarvi and M. Kuwahara, "Using ITS to improve the capacity of freeway merging sections by transferring freight vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 4, pp. 580–588, 2008.

[6] A. Abadi, T. Rajabioun, and P. A. Ioannou, "Traffic flow prediction for road transportation networks with limited traffic data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 1–10, 2014.

[7] Z. Yao, Y. Wu, Y. Wang, B. Zhao, and Y. Jiang, "Analysis of the impact of maximum platoon size of CAVs on mixed traffic flow: an analytical and simulation method," *Transportation Research Part C: Emerging Technologies*, vol. 147, pp. 103989–104039, 2023.

[8] A. Sharma, Z. Zheng, J. Kim, A. Bhaskar, and M. Haque, "Estimating and comparing response times in traditional and connected environments," *Transportation Research Record*, vol. 2673, no. 4, pp. 674–684, 2019.

[9] A. Sharma, Z. Zheng, J. Kim, A. Bhaskar, and M. Mazharul Haque, "Assessing traffic disturbance, efficiency, and safety of the mixed traffic flow of connected vehicles and traditional vehicles by considering human factors," *Transportation Research Part C: Emerging Technologies*, vol. 124, no. 1, Article ID 102934, 2021.

[10] M. Zhao, C. Roncoli, Y. Wang, N. Bekiaris-Liberis, J. Guo, and S. Cheng, "Generic approaches to estimating freeway traffic state and percentage of connected vehicles with fixed and mobile sensing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13155–13177, 2022.

[11] D. H. Wu, R. Peng, and X. L. Ling, "Hybrid characteristics of heterogeneous traffic flow in intelligent network," *Journal of Southwest Jiaotong University*, vol. 57, no. 4, pp. 761–768, 2022.

[12] A. Y. Zhou and S. Peeta, "Cooperative driving in mixed-flow traffic of connected vehicles and human-driven vehicles: a state estimation approach," in *Proceedings of the 100th Annual Meeting of the Transportation Research Board*, Washington, DC, USA, January 2021.

[13] P. Swietojanski, A. Ghoshal, and S. Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6744–6748, IEEE, Vancouver, Canada, May 2013.

[14] U.S. Federal Highway Administration, "I-80 Dataset," 2006, https://ops.fhwa.dot.gov/trafficanalysi.stools/ngsim.html.

[15] D. F. Xie, X. M. Zhao, and Z. B. He, "Heterogeneous traffic mixing regular and connected vehicles: modeling and stabilization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2060–2071, 2019.

[16] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical Review A*, vol. 62, no. 2, pp. 1805–1824, 2000.

[17] J. Shreve, H. Schneider, and O. Soysal, "A methodology for comparing classification methods through the assessment of model stability and validity in variable selection," *Decision Support Systems*, vol. 52, no. 1, pp. 247–257, 2011.