A Data-Driven Car-Following Model Based on the Random Forest

Huili Shi[1], Tingli Wang[1,2], Fusheng Zhong[1], Hanqing Wang[1], Junyan Han[1], Xiaoyuan Wang[1,2*]

[1]College of Electromechanical Engineering, Qingdao University of Science & Technology, Qingdao, China
[2]Shandong Intelligent Green Manufacturing Technology and Equipment Collaborative Innovation Center, Qingdao, China
Email: *wangxiaoyuan@qust.edu.cn

Abstract

The car-following models are the research basis of traffic flow theory and microscopic traffic simulation. Among the previous work, the theory-driven models are dominant, while the data-driven ones are relatively rare. In recent years, the related technologies of Intelligent Transportation System (ITS) represented by the Vehicles to Everything (V2X) technology have been developing rapidly. Utilizing the related technologies of ITS, the large-scale vehicle microscopic trajectory data with high quality can be acquired, which provides the research foundation for modeling the car-following behavior based on the data-driven methods. According to this point, a data-driven car-following model based on the Random Forest (RF) method was constructed in this work, and the Next Generation Simulation (NGSIM) dataset was used to calibrate and train the constructed model. The Artificial Neural Network (ANN) model, GM model, and Full Velocity Difference (FVD) model are employed to comparatively verify the proposed model. The research results suggest that the model proposed in this work can accurately describe the car-following behavior with better performance under multiple performance indicators.

Keywords

Traffic Flow, Car-Following Model, Data-Driven Method, Random Forest, Intelligent Transportation System

1. Introduction

The traffic flow theory is the theoretical basis for analyzing the operation mechanism of traffic flow under different traffic conditions to effectively organize and manage the transportation system. The car-following behavior is the driving

behavior that the driver follows the preceding vehicle when he/she cannot change lane. As the most basic driving behavior, the modeling study on car-following behavior is one of the core research contents of traffic flow theory, and it has received extensive attention from researchers from multiple research fields [1] [2]. Compared with another common driving behavior model (*i.e.* the lane-changing model [3] [4]), the car-following model describes the longitudinal behavior of vehicles in the current lane, which is very common in the restricted overtaking section (such as ramp) and the continuous-flow facilities (such as the highway). Establishing an effective model is the premise of accurately describing the car following behavior. At present, the theory-driven models are dominant in the research on car-following behavior [5]. The theory-driven models represented by the GM model [6], Gipps model [7], OV model [8], FVD model [9], and ID model [10] as well as their extended models [11] [12] [13] [14] [15] have shown high performance in the respective research fields. However, the car-following behavior is a typical nonlinear and time-varying research object. For this type of research object, it is difficult to apply one single theoretical method to construct a model that can describe its characteristics with higher accuracy and strong generalization ability. Comparably, the data-driven method has shown unparalleled performance in describing non-linear and time-varying research objects. Different from the theory-driven methods, which have a clear model structure and are based on various premises as well as the strict mathematical derivation, the data-driven methods are based on data to establish a description method of the research object by exploring the internal connections of the data. Data-driven methods are not sensitive to prior knowledge and theoretical assumptions but are very sensitive to the quality of data. In other words, the availability of high-quality data directly determines whether an effective and accurate model can be constructed using data-driven methods. In recent years, the ITS-related technologies have been rapidly developing and popularizing, of which the core feature is informatization. In ITS, using the high-altitude or overhead image acquisition system, global positioning system, smartphone, vehicle-mounted sensors, roadside sensors, and other V2X equipment, traffic managers and researchers can obtain high-precision and large-scale vehicle trajectory data, which provides the basis of modeling the car-following behavior based on the data-driven methods. The existing car-following models based on the data-driven methods mainly focus on the fuzzy logic method [16] [17] [18], the ANN method [19] [20] [21] [22], and the combination of these two methods [23]. In the fuzzy logic method, it is difficult to construct the fuzzy sets and the corresponding membership functions. And in the ANN method, the structure is relatively complex and the train requires high-performance computing resources. In contrast, as a typical integrated machine learning method, the RF [24] has shown very high performance in many fields [25] [26] [27] [28].

Based on this, a car-following model based on the RF is constructed employing high-precision, high-refresh-rate, and large-scale vehicle trajectory data by

exploring the internal connections of the data in this work to achieve an accurate description of the car-following behavior. The main contents are: in Section 2, the model is proposed; in Section 3, the training and verification of the model are carried out; and in Section 4, the conclusion is given.

2. Model

The RF is a parallel ensemble learning algorithm based on the Bagging ensemble learning theory [29] and the random subspace method [30], of which the basic learner is the Classification and Regression Tree (CART). The basic structure of RF is as shown in **Figure 1**.

From **Figure 1**, we can obtain that the core characteristics of the RF method are "random" and "parallel". The "random" gives the RF method the performance with high prediction accuracy and strong generalization ability, and the "parallel" gives the method the high training and working efficiency. The "random" of the RF method is reflected in two aspects: the randomness of the sample and the randomness of the attributes of the sample. The "parallel" of the RF method is embodied in that one can train all the T decision trees contained in the RF at the same time, thereby greatly improving the efficiency of training and working.

The training process of the RF method is as shown in **Figure 2**.

As shown in **Figure 2**, in the RF method, when the data set is input, it selects the input data set according to the Bagging theory, and randomly extracts the sample set. For the m data in the input data set, the probability P of each data not being selected is:

$$P = \left(1 - 1/m\right)^{m} \tag{1}$$

Taking the limit of Equation (1), one can obtain

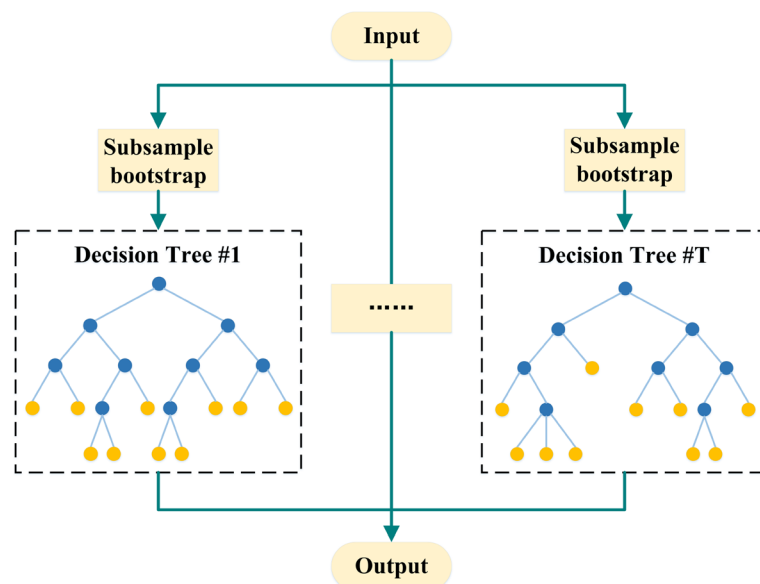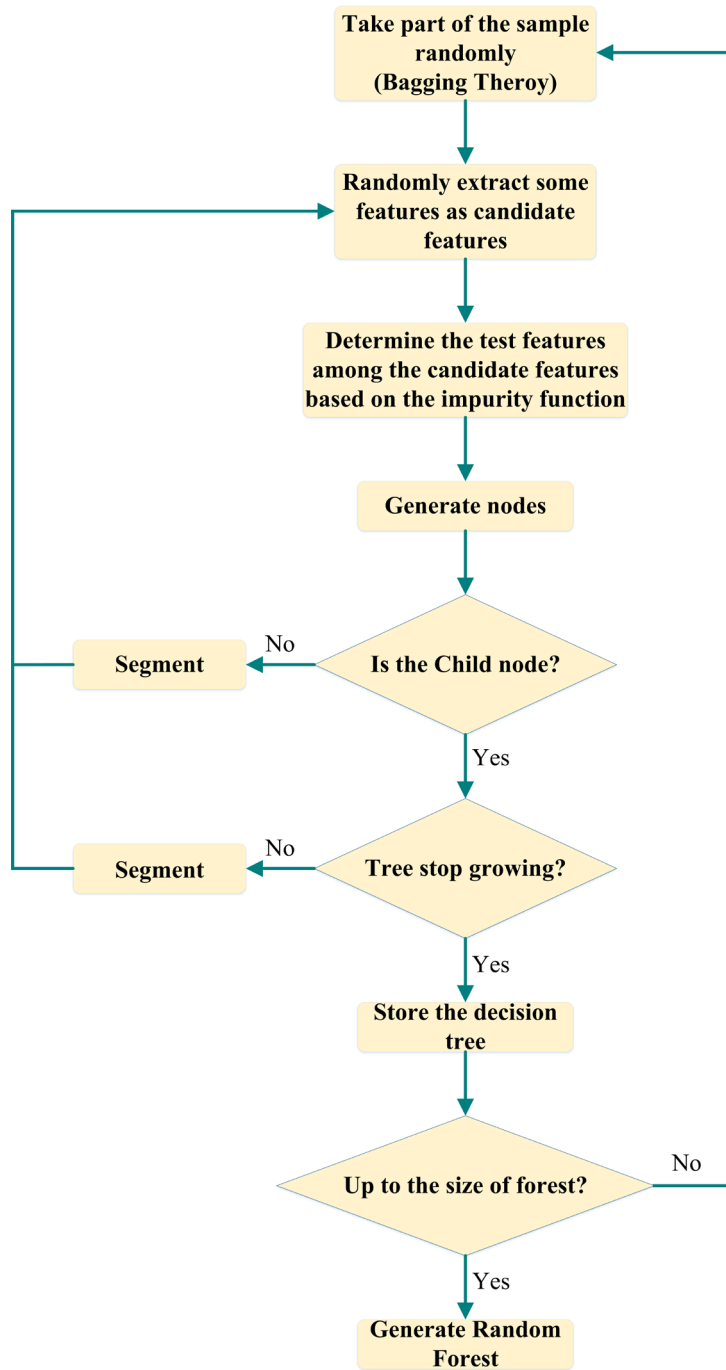**Figure 1.** Basic structure of the RF method.

Figure 2. Training process of the RF method.

$$\lim_{m \to \infty} P \to \frac{1}{e} \approx 0.368 \qquad (2)$$

From Equation (2), we can see that 63.2% of the data is randomly selected from the input data set for the training of one of the decision trees in each round of sampling, which is the sample randomness mentioned above. For ensemble learning, the stronger the independence of the basic learners it contains, the better the performance of the assembled learner. It is almost impossible to construct

completely independent basic learners, and the random extraction principle of the Bagging theory guarantees the relative independence of the basic learners to the greatest extent, referring to the Equation (1) and (2).

When applying the RF method, there are points that need to be determined:

1) Input of the model;

2) Number of attributes in the split attribute set;

3) Impurity function;

4) Size of the forest.

The process of training the RF method is the process of training the decision tree it contains. The core of this process is how to segment features. Given the relatively low number of features involved in this study, the exhaustive method is adopted, which traverses all the values of each feature to find the optimal segmentation. The impurity is used to evaluate the optimal degree of segmentation. For each child node, the calculation method of impurity [31] is

$$G\left(x_i, u_{ij}\right) = \frac{n_{left}}{N_s} H\left(X_{left}\right) + \frac{n_{right}}{N_s} H\left(X_{right}\right) \tag{3}$$

where x_i is the segmentation variable, y_{ij} is a segmentation value of the segmentation variable, n_{left} and n_{right} respectively are the number of training samples of the left and right child nodes after segmentation, N_s is the number of training samples of the current node, X_{left} and X_{right} respectively are the training sample sets of the left and right child nodes, and $H(X)$ is the impurity function. The commonly used impurity functions are shown in Table 1.

The first two impurity functions are suitable for the classification problem, while the latter two impurity functions are suitable for the regression problem.

Based on the characteristics of the RF method and considering the characteristics of the research object car-following behavior, the structure of the RF-based car-following model constructed in this research is as shown in Figure 3.

Table 1. Optional impurity function in the RF method.

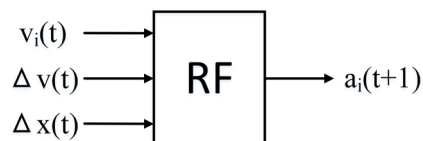| Impurity Function	Equations
Gini Impurity	$H\left(X_m\right) = \sum_k p_{mk}\left(1 - p_{mk}\right)$
Information Entropy	$H\left(X_m\right) = -\sum_k p_{mk} \log\left(p_{mk}\right)$
Mean Square Error	$H\left(X_m\right) = \frac{1}{N_m} \sum_{i \in N_m}\left(y - \bar{y}_m\right)^2$
Absolute Mean Error	$H\left(X_m\right) = \frac{1}{N_m} \sum_{i \in N_m}\left(y - \bar{y}_m\right)^2$

Figure 3. Structure RF-based car-following model.

In the model, the input is the velocity $v(t)$ of the object vehicle at the current moment t, the headway $\Delta x(t)$ between the object vehicle and its preceding vehicle at the current moment t, the relative velocity $\Delta v(t)$ between the object vehicle and its preceding vehicle at the current moment t, and the output is the acceleration $a(t+1)$ of the object vehicle at the next moment $t+1$. And the impurity function $H(X)$ employed in this work is

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y - \bar{y}_m) \tag{4}$$

Then the training process for a certain node in the RF is equivalent to the following optimization problem

$$(x^*, u^*) = \arg\min_{x,y} G(x_i, u_{ij}) \tag{5}$$

Substituting Equation (4) into Equation (5), we can obtain

$$G(x,u) = \frac{1}{N_s} \left(\sum_{y_i \in X_{left}} (y_i - \bar{y}_{left})^2 + \sum_{y_i \in X_{right}} (y_i - \bar{y}_{right})^2 \right) \tag{6}$$

Equation (6) is the solution method of each node in the RF-based car-following model constructed in this research.

The size of the forest is determined by the iterative method during the training process, and the detailed information about this is given in Section 3.

3. Calibration and Training

3.1. Data Preprocessing

The effectiveness and accuracy of the data-driven model depend on the quality of the training data. The US101 dataset provided by the NGSIM project initiated by the Federal Highway Administration is utilized to complete the training and verification of the proposed model. The NGSIM project aims to provide high-precision vehicle trajectory data required for research in the transportation field. It has the characteristics of abundant data, complete objects, high accuracy, and acquisition frequency of 0.1 s/time, and it is widely used in car-following behaviors and other research fields. The validity of this data set has been widely recognized. However, the total amount of the data set is abundant, and many data are not suitable for this research. The preprocess needs to be carried out.

The NGSIM project implemented vehicle trajectory data collection on different road sections in December 2003, April 2005, and June 2005. The US101 data set employed in this work was collected on Hollywood Expressway (No. US-101) in June 2005, and the lane setting of the data collection section is as shown in **Figure 4**.

The US101 data set contains microscopic car-following trajectory data such as the position, velocity, and acceleration of 6101 different types of vehicles. The specific data fields include Vehicle ID, Frame ID, Total Frames, Global Time, Local X, Local Y, Global X, Global Y, Vehicle Length, Vehicle Width, Vehicle Class, Vehicle Velocity, Vehicle Acceleration, Lane Identification, Preceding Ve-

hicle, Following Vehicle, Spacing, and Headway. For the usage in this work, the above data fields contain redundant ones, specifically: Local X, Local Y, Global X, Global Y, Vehicle Length, Vehicle Width, Vehicle Class, Vehicle Acceleration, and Following Vehicle.

Although the US101 data set contains large-scale car-following trajectory data up to 6101, it cannot be directly used for the l training and verification of the constructed model. The preprocess needs to be carried out, and the detailed process is as shown in Figure 5.
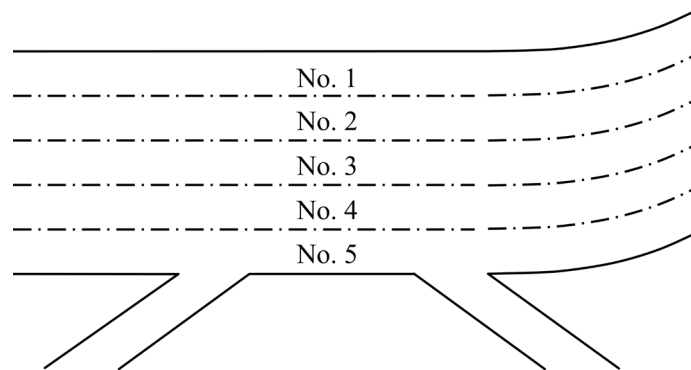

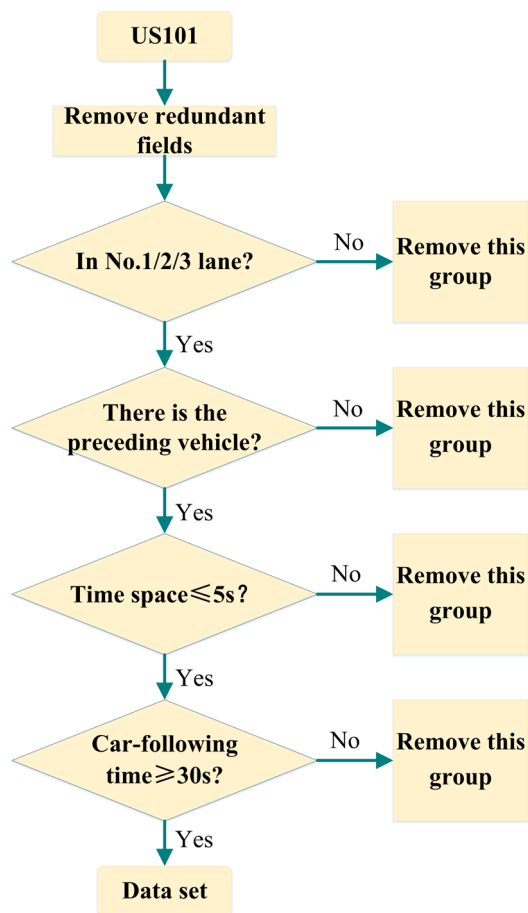**Figure 4.** Lane setting of the US101 road section.

Figure 5. Preprocess process of the data set.

By traversing and processing all the items included in the US101 data set one by one according to the process shown in **Figure 5**, a data set containing 2152 groups of car-following trajectory data suitable for this study is obtained. 70% of them (1506 groups in total) are randomly selected as the training set, and the remaining 30% (646 groups in total) is used as the validation set.

3.2. Model Calibration and Training

Input the training set into the RF-based car-following model, and the model is trained based on this. In the training of the model, the size of the trees contained in the RF has a significant impact on the training quality. Among the previous research and application, there are no certain rules for setting the size of the trees. The common processing method is to rely on expert experience to set the initial value and repeat the process of testing, adjusting parameters and retesting, and finally get the optimal setting value, which is the so-called iterative method. Based on the iterative method, the size of the trees is set as a value in a given interval, and the optimal value of the parameter is determined by examining the prediction error of the model under the corresponding value. Considering the scale of the data set and the features used in the research, the interval is set as $[10, 210]$, and the size of the trees is set as S_{tree}. Then we can obtain $S_{tree} = 10, 11, 12, \cdots, 209, 210$.

The corresponding error under different values of S_{tree} is as shown in **Figure 6**.

From **Figure 6**, one can obtain that with the increase of S_{tree}, the error is considerably decreased when S_{tree} is less than 100. The error decreased slightly with the increase of S_{tree}, and there are also some fluctuates of the error, when S_{tree} is more than 110. Considering that more consumption of computing resources with the increase of S_{tree} and the amount of error reduction caused by
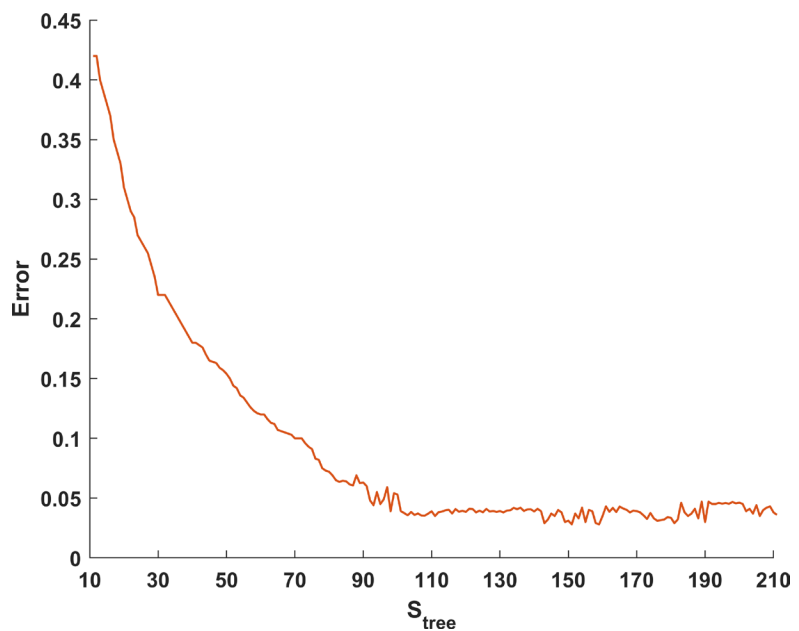

Figure 6. Errors under different values of S_{tree}.

this increase in the unit is becoming less and less significant, the S_{tree} in the proposed model is set as 110. Based on this, the training of the model is conducted.

4. Verification and Discussion

To verify the validity and accuracy of the RF-based car-following model constructed in this research, the performance of the model is evaluated utilizing the verification set. Then, the representative data-driven model (the one based on the ANN) and theory-driven models (the GM model and the FVD model) are employed to compare with and verify the proposed model with the same data set (the verification set). Before the verification, the training or the calibration of the above models are carried out according to the previous research: the aforementioned $v(t)$, $\Delta x(t)$ as well as $\Delta v(t)$ are set as the input of the ANN model and the Genetic Algorithm is employed to calibrate parameters in the GM model and the FVD model. After the training or the calibration of these models, the verification set is used to evaluate the performance of the proposed model in this work. The Mean Error (ME), Mean Absolute Error (MAE), Mean Absolute Relative Error (MARE) and Root Mean Squared Error (RMSE) are employed as the evaluation indicators, and the equations of these indicators are

$$\text{ME} = \frac{1}{N}\sum_{i=1}^{N}(d_{ci} - d_{mi}) \tag{7}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|d_{ci} - d_{mi}| \tag{8}$$

$$\text{MARE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_{ci} - d_{mi}|}{d_{ci}} \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_{ci} - d_{mi})^2} \tag{10}$$

where N is the total amount of data, d_{ci} is the output value of the *i*-th object vehicle, and d_{mi} is the measured value of the *i*-th object vehicle.

The above evaluation indicators are used to evaluate the performance of the proposed model in this work and the models employed to compare with the proposed model. The evaluation results are as shown in Table 2.

The ME refers to the arithmetic mean of the errors of all output values relative

Table 2. Evaluation results of the models.

Model / Evaluation Indicators	ME	MAE	MARE	RMSE
RF-Based Model	0.1673	1.5746	0.4356	1.0753
ANN Model	0.1952	1.7853	1.9174	1.1346
GM Model	1.1712	3.2715	2.6371	1.2174
FVD Model	1.1548	3.0187	2.5986	1.1949

to the measured ones, which reflects the average deviation between the output value and the measured value. The MAE further introduces the absolute value to avoid the problem of inaccurate evaluation caused by the offset of the positive and negative ones. The RMSE index is very sensitive to extra large and small values, and thus it can reflect the obvious degree of the deviation between the output value and the measured value. ME, MAE, RMSE reflect the degree of the error, while MARE reflects the proportion of the error in the samples. From Table 2, we can see that the four models show considerably different performances with the same data set. Among these models, the performance of the model proposed in this work is better than the others. According to the evaluation indicators, the performance improvement range of the model proposed in this work is up to 85.716% and can maintain 5.227% at the lowest level. In addition, the fit degree of the two data-driven models to the measured data is significantly better than that of the two theory-driven models. Among the data-driven models, compared with the ANN model commonly used in previous research, the performance improvement of the model proposed in this work can reach up to 77.282%. Among the theory-driven models, the FVD model, in which more factors are considered, shows better performance than the GM model. This is consistent with the research consensus in the field of modeling car-following behavior, which verifies the validity and reliability of the employed evaluation system. Compared with the FVD model, the performance improvement of the model proposed in this work is up to 85.513%, and the value can reach up to 85.716% when compared with the GM model. Even considering the lowest improvement range, the value is 11.672% when compared with the FVD model, and that is 10.009% when compared with the GM model.

5. Conclusion

The theory-driven car-following behavior model still has shortcomings in terms of prediction accuracy and generalization ability. The application of the ITS facilitates the collection of large-scale, high-quality vehicle trajectory data, which is the research foundation for the car-following models based on data-driven methods. In this work, a data-driven car-following model was constructed based on the RF method, and the NGSIM data set was used to train and verify the model. The results show that compared with the data-driven model and theory-driven models that are widely used in the previous research, the model proposed in this work has better performance represented by four typical evaluation indicators, which verified the validity and accuracy of the model. Compared with typical data-driven methods, such as the ANN method, the RF method employed in this work not only has better prediction accuracy, but also has the advantages of low computational power consumption and extensive trial range. It is not required for the RF method to achieve excellent training performance with a high-performance GPU. With the appropriate data set, the RF method can theoretically be suitable for solving a considerable part of scientific issues, including regression and clas-

sification issues. This maps to the car-following behavior and the lane-change behavior, when talking about the traffic flow theory. The application efficiency of random forest method in other traffic flow theories, other than car-following behavior and even broader fields is worthy of further exploration.

Acknowledgements

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Han, J., Zhang, J., Wang, X., Liu, Y., Wang, Q. and Zhong, F. (2020) An Extended Car-Following Model Considering Generalized Preceding Vehicles in V2X Environment. *Future Internet*, **12**, Article No. 216. https://doi.org/10.3390/fi12120216

[2] Wang, X., Han, J., Bai, C.-L., Shi, H., Zhang, J. and Wang, G. (2021) Research on the Impacts of Generalized Preceding Vehicle Information on Traffic Flow in V2X Environment. *Future Internet*, **13**, Article No. 88.
https://doi.org/10.3390/fi13040088

[3] Wang, X., Wang, F., Kong, D., Liu, Y., Liu, L. and Chen, C. (2018) Driver's Lane Selection Model Based on Phase-Field Coupling and Multiplayer Dynamic Game with Incomplete Information. *Journal of Advanced Transportation*, **2018**, Article ID: 2145207.
https://doi.org/10.1155/2018/2145207

[4] Wang, X., Liu, Y., Wang, F., Liu, Z., Zhao, H. and Xin, J.-F. (2019) Driver's Lane Selection Model Based on Multi-Player Dynamic Game. *Advances in Mechanical Engineering*, **11**, Article ID: 168781401881990.
https://doi.org/10.1177%2F1687814018819903

[5] Brackstone, M. and McDonald, M. (1999) Car-Following: A Historical Review. *Transportation Research Part F: Traffic Psychology and Behaviour*, **2**, 181-196.
https://doi.org/10.1016/S1369-8478(00)00005-X

[6] Ozaki, H. (1993) Reaction and Anticipation in the Car-Following Behavior. *Proceedings of the* 12*th International Symposium on the Theory of Traffic Flow and Transportation*, Berkeley, 21-23 July 1993, 349-366.

[7] Gipps, P.G. (1981) A Behavioural Car-Following Model for Computer Simulation. *Transportation Research Part B: Methodological*, **15**, 105-111.
https://doi.org/10.1016/0191-2615(81)90037-0

[8] Bando, M., Hasebe, K., Nakayama, A., Shibata, A. and Sugiyama, Y. (1995) Dynamic Model of Traffic Congestion and Numerical Simulation. *Physical Review E*, **51**, 1035-1042. https://doi.org/10.1103/PhysRevE.51.1035

[9] Jiang, R., Wu, Q. and Zhu, Z. (2001) Full Velocity Difference Model for Car-Follow-

ing Theory. *Physical Review E*, **64**, Article ID: 017101.
https://doi.org/10.1103/PhysRevE.64.017101

[10] Treiber, M. and Helbing, D. (2003) Memory Effects in Microscopic Traffic Models and Wide Scattering in Flow-Density Data. *Physical Review E*, **68**, Article ID: 046119. https://doi.org/10.1103/PhysRevE.68.046119

[11] Farhi, N. (2012) Piecewise Linear Car-Following Modeling. *Transportation Research Part C: Emerging Technologies*, **25**, 100-112. https://doi.org/10.1016/j.trc.2012.05.005

[12] Yu, S., Zhao, X., Xu, Z. and Zhang, L. (2016) The Effects of Velocity Difference Changes with Memory on the Dynamics Characteristics and Fuel Economy of Traffic Flow. *Physica A: Statistical Mechanics and Its Applications*, **461**, 613-628. https://doi.org/10.1016/j.physa.2016.06.060

[13] Tang, T.-Q., Zhang, J., Chen, L. and Shang, H.-Y. (2017) Analysis of Vehicle's Safety Envelope under Car-Following Model. *Physica A: Statistical Mechanics and Its Applications*, **474**, 127-133. https://doi.org/10.1016/j.physa.2017.01.076

[14] Kuang, H., Xu, Z.-P., Li, X.-L. and Lo, S.-M. (2017) An Extended Car-Following Model Accounting for the Average Headway Effect in Intelligent Transportation System. *Physica A: Statistical Mechanics and its Applications*, **471**, 778-787. https://doi.org/10.1016/j.physa.2016.12.022

[15] Huang, Y.-X., Jiang, R., Zhang, H., Hu, M.-B., Tian, J.-F., Jia, B. and Gao, Z.-Y. (2018) Experimental Study and Modeling of Car-Following Behavior under High Speed Situation. *Transportation Research Part C: Emerging Technologies*, **97**, 194-215. https://doi.org/10.1016/j.trc.2018.10.022

[16] Kikuchi, S. and Chakroborty, P. (1992) Car-Following Model Based on Fuzzy Inference System. *Transportation Research Record*, No. 1365, 82-91.

[17] Rekersbrink, A. (1995) Mikroskopische Verkehrssimulation Mit Hilfe Der Fuzzy Logik. *Strass enverkehrstechnik*, **2**, 68-74.

[18] Mcdonald, M., Wu, J. and Brackstone, M. (1997) Development of a Fuzzy Logic Based Microscopic Motorway Simulation Model. *Proceedings of Conference on Intelligent Transportation Systems*, Boston, 12 November 1997, 82-87. https://doi.org/10.1109/ITSC.1997.660454

[19] Kehtarnavaz, N., Groswold, N., Miller, K. and Lascoe, P. (1998) A Transportable Neural-Network Approach to Autonomous Vehicle Following. Vehicular Technology. *IEEE Transactions on Vehicular Technology*, **47**, 694-702. https://doi.org/10.1109/25.669106

[20] Zhang, Y., Lin, Q., Wang, J. and Verwer, S. (2017) Car-Following Behavior Model Learning Using Timed Automata. *IFAC-Papers OnLine*, **50**, 2353-2358. https://doi.org/10.1016/j.ifacol.2017.08.423

[21] Zhu, M., Wang, X. and Wang, Y. (2018) Human-Like Autonomous Car-Following Model with Deep Reinforcement Learning. *Transportation Research Part C: Emerging Technologies*, **97**, 348-368. https://doi.org/10.1016/j.trc.2018.10.024

[22] Wang, X., Jiang, R., Li, L., Lin, Y.-L. and Wang, F.-Y. (2019) Long Memory Is Important: A Test Study on Deep-Learning Based Car-Following Model. *Physica A: Statistical Mechanics and its Applications*, **514**, 786-795. https://doi.org/10.1016/j.physa.2018.09.136

[23] Ma, X. (2006) A Neural-Fuzzy Framework for Modeling Car-Following Behavior. 2006 *IEEE International Conference on Systems, Man and Cybernetics*, Taipei, 8-11 October 2006, 1178-1183. https://doi.org/10.1109/ICSMC.2006.384560

[24] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.

https://doi.org/10.1023/A:1010933404324

[25]   Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A. and Blake, A. (2011) Real-Time Human Pose Recognition in Parts from Single Depth Images. 2011 *Conference on Computer Vision and Pattern Recognition*, Colorado, 20-25 June 2011, 1297-1304. https://doi.org/10.1109/CVPR.2011.5995316

[26]   Baumann, T. (2014) Decision Tree Usage for Incremental Parametric Speech Synthesis. 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Florence, 4-9 May 2014, 3819-3823. https://doi.org/10.1109/ICASSP.2014.6854316

[27]   Lindner, C., Bromiley, P., Ionita, M. and Cootes, T. (2014) Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 1862-1874. https://doi.org/10.1109/TPAMI.2014.2382106

[28]   Acharjee, A., Kloosterman, B., Visser, R. and Maliepaard, C. (2016) Integration of Multi-Omics Data for Prediction of Phenotypic Traits Using Random Forest. *BMC Bioinformatics*, **17**, Article No. 180. https://doi.org/10.1186/s12859-016-1043-4

[29]   Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1007/BF00058655

[30]   Ho, T. (1998) The Random Subspace Method for Constructing Decision Forests. Pattern Analysis and Machine Intelligence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 832-844. https://doi.org/10.1109/34.709601

[31]   Breiman, L., Friedman, J., Olshen, R. and Stone, C. (2017) Classification and Regression Trees. Routledge, Boca Raton. https://doi.org/10.1201/9781315139470