Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 9695433, 13 pages
https://doi.org/10.1155/2023/9695433

Research Article

Highway Traffic Crash Risk Prediction Method considering Temporal Correlation Characteristics

Liping Zhao,[1] **Feng Li**,[1] **Dongye Sun**,[2] **and Fei Dai**[1]

[1]*Institute of Systems Engineering, Academy of Military Sciences, No. 2 Fengti South Road, Fengtai District, Beijing 100166, China*
[2]*National Engineering Research Center for Transportation Safety and Emergency Informatics, Telecommunications & Information Center, No. 1 Anwai Waiguan Houshen, Beijing 100011, China*

Correspondence should be addressed to Liping Zhao; 825889797@qq.com and Dongye Sun; 14114218@bjtu.edu.cn

Received 22 April 2022; Revised 15 September 2022; Accepted 25 November 2022; Published 15 February 2023

Academic Editor: Fei Hui

Crash risk analysis and prediction are considered the premise of highway traffic safety control, which directly affects the accuracy and effectiveness of traffic safety decisions. A highway traffic crash risk prediction method considering temporal correlation characteristics is proposed in this research. Firstly, the case-control sample analysis method is used to extract 6 time series sample data composed of crash traffic flow data and corresponding non-crash traffic flow data for crash risk analysis and prediction. Secondly, the multiparameter fusion clustering analysis method is used to indicate that the sample data of different time series have different effects on the crash risk. Then, the random forest model is used to screen several traffic flow variables that affect the highway crash risk. Thereafter, the downstream mean speed (ASD1D2), the upstream mean occupancy (AOU1U2), and the speed difference (DSU1D1) on the nearest detector were determined as the explanatory variables of the crash risk prediction model. Finally, based on the three variables, the dynamic Bayesian network model for highway traffic crash risk prediction is proposed. The overall prediction accuracy of this model is 84.9%, the crash prediction accuracy is 60.8%, and the non-crash prediction accuracy is 92.3%. Also, the prediction results show that the dynamic Bayesian model has better prediction effect than the static Bayesian model for the same sample data.

1. Introduction

Research on road traffic safety has a long history. Early studies about road traffic safety focused on the crash cause mechanism and influencing factors analysis. Yang et al. explored and analyzed the influence of different geographical conditions and environmental factors on highway crash risk by using the improved association rule algorithm [1–4]. Wang et al. [5–7] analyzed the traffic crash causative factors under different traffic modes from the microscopic perspective. However, it is not realistic to consider all the factors (drivers, vehicles, roads, and environments) in the traffic crash cause modeling. The randomness of traffic crashes makes traffic crash-causing analysis fall into the bottleneck.

In recent years, the concept of active safety has gradually entered the vision of researchers. More and more studies [8–10] have found that traffic flow state has a strong correlation with the road traffic crash occurrence. For example, Lee et al. used the traffic flow data within 5 minutes before the crash to study the influence of traffic flow dynamic characteristics on the collision pattern of the crash [11]. Golob and Recker divided the traffic flow pattern before the crash into different traffic flow states and found that different traffic flow states likely led to different types of traffic crashes [12]. Golob et al.'s study firstly extracted six variables representing the traffic flow characteristics before the crash, then divided the traffic flow state by using these six variables as the clustering index, and finally analyzed the types of traffic crashes that are prone to occur in various traffic flow states [13, 14]. Golob et al.'s study indicates that in traffic flow running state, there is a certain relationship with the traffic crash [15]. However, these studies are all based on the traffic flow data before the crashes. It is difficult to reflect the

randomness of traffic crashes. Also, it is not conducive to accurately identify the prone crash traffic state from the normal traffic flow state.

The highway traffic system is a dynamic, undulating, and non-linear complex system. From the perspective of spatial dimension, the traffic flow variables closest to the crash site, that is, the traffic flow variables upstream and downstream of the crash site, are most correlated with crash risk. From the time dimension, the traffic flow state variables in a period of time before the crash are most likely to have a certain correlation with the crash occurrence [16–19]. At present, a large number of studies [20–22] have used traffic state parameters such as flow/speed/density as explanatory variables to analyze and predict the possibility of traffic crashes. However, the traffic flow data collected by coil, microwave radar, and floating vehicle have strong time-varying characteristics. Therefore, the temporal and spatial characteristics of traffic flow state variables have a certain degree of influence on crash risk modeling. Moreover, the essence of highway crash risk prediction and discrimination is the causal relationship between the running traffic flow state variables of upstream and downstream and the possibility of crash occurrence. Then, we establish linear or non-linear relationship model and determine whether there is a risk in the future traffic safety running state. At the same time, the traffic safety data have certain temporal correlation characteristics. The long scale cumulative traffic flow sequence generally contains multiple subsequences with different stage characteristics. The time correlation between the sequences will have a certain influence on the crash risk prediction model. Conventional real-time crash risk prediction models do not consider the temporal characteristics of traffic flow sequences, which may affect the prediction accuracy of the model.

In order to solve this problem, this paper proposes a highway crash risk prediction model considering time sequence correlation. The dynamic Bayesian network model is used to characterize the influence of time correlation characteristics on crash risk prediction model. Firstly, the random forest model is used to screen the traffic flow state variables that affect the highway crash risk. Then, the dynamic Bayesian network model is used to explain the influence of the temporal correlation of traffic flow state variables on the modeling of traffic safety crash risk prediction. The influence of time sequence characteristics between variables on highway crash risk modeling is illustrated by comparative analysis. The results show that the proposed method has better identification rate and lower error rate. It further shows that the dynamic Bayesian network model can better describe the dynamic time-varying characteristics of traffic flows before crashes.

The rest of the text is arranged as follows. Section 2 briefly introduces the study area, sample data sources, and temporal characteristics of crash traffic flow variables. Section 3 presents the research problems, solutions, and related model methods involved. Section 4 introduces the comparative analysis and discussion of the model operation results. The research results are summarized in Section 5.

2. Study Area and Data Survey

2.1. Data Collection and Processing. The main research field of this paper is to reveal the influence of traffic flow timing characteristics on highway crash risk prediction. The data involved include traffic crash data and corresponding upstream and downstream traffic flow data within a certain time range. The sample data used in this study are the traffic safety crash data and corresponding traffic flow state data on the 495.493–539.045 miles section of interstate highway I5 in California, USA. In order to control the influence of weather, road conditions, and other factors on crash risk prediction modeling, the case-control sample structure was used to match the sample data. Based on the location where each kind of crash data occurred, the traffic flow state data of the four detectors closest to the crash were extracted, and the two detectors upstream were named U2 and U1, and the two detectors downstream were named D2 and D1, as shown in Figure 1(a). In order to accurately identify the influence of time series characteristics of crash traffic flow variables on crash risk prediction model, the data of traffic flow variables within 30 minutes before the crash were extracted. They are divided into 6 time segments every 5 minutes, including time series 0 (i.e., 0–5 minutes before the crash), time series 1 (i.e., 5–10 minutes before the crash), time series 2 (i.e., 10–15 min before the crash), time series 3 (i.e., 15–20 min before the crash), time series 4 (i.e., 20–25 min before the crash), and time series 5 (i.e., 25–30 minutes before the crash), as shown in Figure 1(b). It should be noted that since time series 0 is after the crash, it is only suitable for crash detection, but not for crash risk estimation. In addition, traffic crash identification and taking corresponding measures need response time. Only the model established by using the traffic detection information in time series 2, series 3, and series 4 has practical value for active safety management.

2.2. Initial Variable Extraction. In order to facilitate the establishment of crash risk prediction model by using dynamic Bayesian network model, the original traffic flow variables on the four detectors and the mean and difference values of upstream and downstream traffic variables are used as the initial variables of the model. Relevant studies show that the traffic safety state upstream and downstream of the crash site can comprehensively reflect the influence of various factors on crash risk. In order to determine the mechanism of this influence, the original data collected by the detector are fused to further explore the impact of upstream and downstream traffic state on crash risk. Therefore, explanatory variables of the model can be divided into three types. The first type refers to the original traffic flow data extracted from four detections, the second type is the difference between traffic variables of upstream and downstream detectors, and the third type is the mean value of traffic variables of upstream and downstream detectors. The specific names of the three types of variables are shown in Table 1. The first letter of the variable name represents the type of the variable, O represents the original variable, D represents the difference of the upstream and downstream

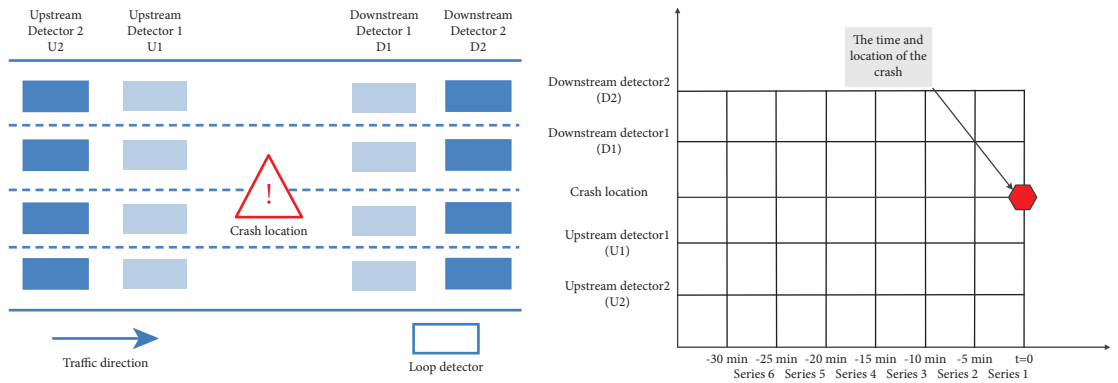

Figure 1: Schematic diagram of sample data extraction based on temporal and spatial features. (a) Spatial feature. (b) Temporal feature.

detector variables, and A represents the mean value of the upstream and downstream detector variables. The second letter represents the variable name, S for speed, V for flow, and O for occupancy. The part underlined in the variable represents the name of detector, U1 represents the first upstream detector, U2 represents the second upstream detector, D1 represents the first downstream detector, and D2 represents the second downstream detector. According to this coding rule, DOU1D1 is the average density difference between the first upstream detector and the first downstream detector, and other variables are named in the same way.

In this study, according to the data sample matching principle, 247 crash and 1096 non-crash traffic flow original data (i.e., volume, speed, and occupancy) of U1, U2, D1, and D2 30 minutes before the crash were extracted. Finally, a total of 1370 sets of data samples were obtained for highway crash risk modeling, in which the ratio of crash traffic flow to non-crash traffic flow was 1 : 4. According to the above method, it is divided into 6 time series, and each segment contains a total of 30 traffic flow variables, which are used as the basic sample data for crash risk prediction modeling. It should be pointed out that in the extracted data samples, the traffic flow variables in each time segment contain not only the three original variables of cumulative flow, average speed, and average occupancy rate on the four upstream and downstream detectors but also the difference and mean value of the traffic flow variables on the upstream and downstream detectors. That is, each of the 6 time series contains 30 variable values as explanatory variables of the model.

2.3. Analysis of Traffic Flow Temporal Characteristics. In order to more intuitively express the influence of the temporal correlation characteristics of traffic flow state variables on the modeling of highway crash risk, the traffic safety state is divided by the traffic flow state variable data in the six time segments mentioned above. According to the classification results, we can see the changing trend of traffic flow state in different time segments at the same place. In this paper, the average speed index of upstream and downstream in two adjacent time segments (time series 1 and series 2) is selected to draw scatter charts, so as to more intuitively analyze the changes of traffic flow state in different time segments. As shown in Figure 2, the horizontal axis is the average speed of upstream traffic flow, and the vertical axis is the average speed of downstream traffic flow.

It can be seen from the figure that the overall trend of traffic flow status did not change much in two consecutive periods before the crash, mainly because non-crash traffic flow accounted for a large proportion in the sample. Through further analysis, it is found that in the two 5-minute time intervals, the proportion of non-crash traffic flow state change is much less than the proportion of crash traffic flow state change. In all 274 crash traffic flow samples, 79 crash traffic flow samples (28.8%) exhibited state transition, while in all 1096 non-crash traffic flow samples, 36 non-crash traffic flow samples (3.2%) exhibited state change. From the proportion of state transition data, it can be seen that in the adjacent time series, the proportion of state transition in the accident traffic flow is much higher than that in the non-accident traffic flow. This indicates that under the same conditions, compared with non-accident traffic flow, time factor has a greater impact on accident traffic flow. In other words, it shows that the change of traffic flow state in different time series before the crash has a strong influence on the highway crash. Therefore, traffic flow characteristics of different time series have different impacts on road traffic risks. It is necessary to consider the state transition process of traffic data capture with multiple time intervals when building the crash risk prediction model.

3. Methodology

At present, there is no fixed procedure and process for constructing highway crash risk prediction model. In order to establish a reasonable crash risk prediction model, it usually has a strong relationship with the types of traffic safety data, the physical significance of explanatory variables, and the purpose of establishing the model. On the basis of existing studies, the general steps to be followed in highway crash risk prediction are proposed in this study, as shown in Figure 3.

Table 1: The statistics of initial variables.

Optional initial variable	Variable naming
Detector original variable	OSU1, OVU1, OOU1, OSU2, OVU2, OOU2, OSD1, OVD1, OOD1, OSD2, OVD2, OOD2
The difference between the variables of upstream and downstream detectors	DSU1D1, DVU1D1, DOU1D1, DSU2D1, DVU2D1, DOU2D1, DSU1D2, DVU1D2, DOU1D2, DSU2D2, DVU2D2, DOU2D2
Mean values of variables of upstream and downstream detectors	ASU1U2, AVU1U2, AOU1U2, ASD1D2, AVD1D2, AOD1D2

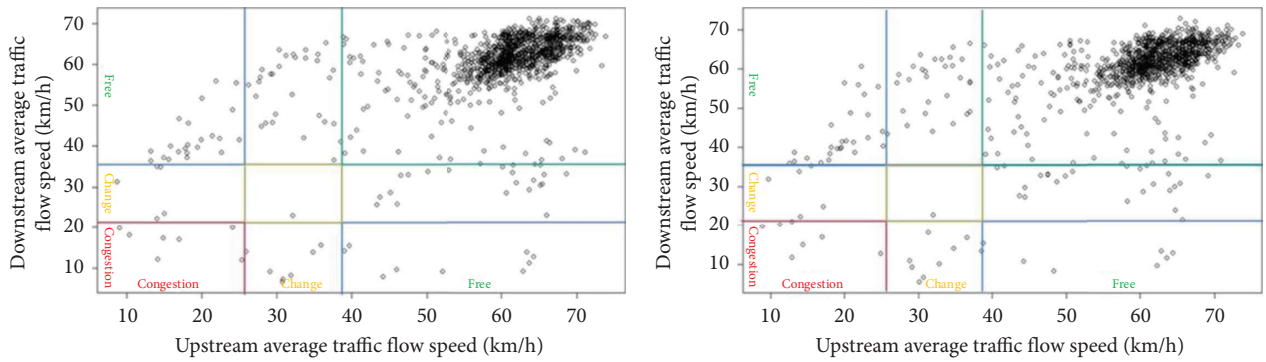

Figure 2: Traffic flow state change in different time segments. (a) Time series 1 (5–10 min before crash). (b) Time series 2 (10–15 min before crash).

Model preparation stage: the preparatory stage of highway crash risk modeling mainly includes defining modeling objectives, understanding the advantages and disadvantages of existing prediction models, and collecting relevant basic data according to the requirements of the models.

Model establishment stage: according to the purpose of modeling, analyze variable requirements and basic data, select appropriate models, assign corresponding physical meanings to model variables, and make corresponding model assumptions for problems that cannot be considered.

Model application stage: according to the established model, with the help of necessary mathematical software and computer technology to solve the model for parameter estimation, model results are discussed and analyzed, such as error analysis, sensitivity analysis, and prediction accuracy analysis.

In addition, the following problems need to be considered and solved during the of highway crash risk prediction modeling. Firstly, there are many traffic flow variables that affect the highway crash risk. How to select the variables with the highest correlation with the crash risk as the explanatory variable of the crash risk prediction model? Secondly, how long are the traffic flow sequence data before the crash used to predict the highway traffic crash risk? Thirdly, what methods or models are used to characterize the crash and non-crash traffic flow data samples with temporal correlation characteristics, so as to establish an effective highway crash risk prediction model?

Aiming at the problems in the modeling process mentioned above, we put forward the modeling steps of highway crash risk assessment model in this paper. First of all, data matching was carried out by using matched case-control sample structure to eliminate the influence of other factors on crash risk modeling to the maximum extent. Secondly, the random forest model is used to explore the correlation between the initial variables of the model and the traffic crash risk, and the variable with the largest correlation coefficient is extracted as the input variable of the crash risk prediction model. Then, the dynamic Bayesian network model is used to quantify the influence of traffic flow timing sequence correlation characteristics on highway crash risk, and the highway crash risk prediction model based on dynamic Bayesian network is established. Finally, the effectiveness of the proposed model is verified by comparison with the prediction results of the static Bayesian network model. The steps of model analysis are as follows.

According to the above crash risk modeling process, the random forest model and dynamic Bayesian network model are used for variable selection and model structure construction. In order to facilitate understanding, the random forest model and dynamic Bayesian network model used in this paper are introduced, respectively.

3.1. Random Forest. Random forest algorithm is an ensemble learning algorithm proposed by Cutler et al. in 2001 [23]. A major advantage of the model is that it is easy to measure the relative importance of each characteristic variable to the prediction [24]. Random forest algorithm is a supervised machine learning algorithm that builds multiple decision trees and merges them together to obtain more accurate and stable classification or regression results [25]. In the random forest algorithm, there are two indexes used to evaluate the importance of variables, which are, respectively, the evaluation indexes of the importance of variables based on Gini value and OOB (out-of-bag) error rate. The influence of different traffic variables on highway crash risk was explored by using OOB error rate evaluation index in this research [26]. Relevant studies have shown that random forest algorithm can solve multicollinearity problem without separate cross validation, and this method can effectively deal with data samples with multiple variables. Therefore, random forest algorithm is used to extract the traffic flow variables highly related to the highway crash risk as the explanatory variables of the prediction model.

3.2. Dynamic Bayesian Networks. Based on probability network, dynamic Bayesian network combines original static network structure with time information to form

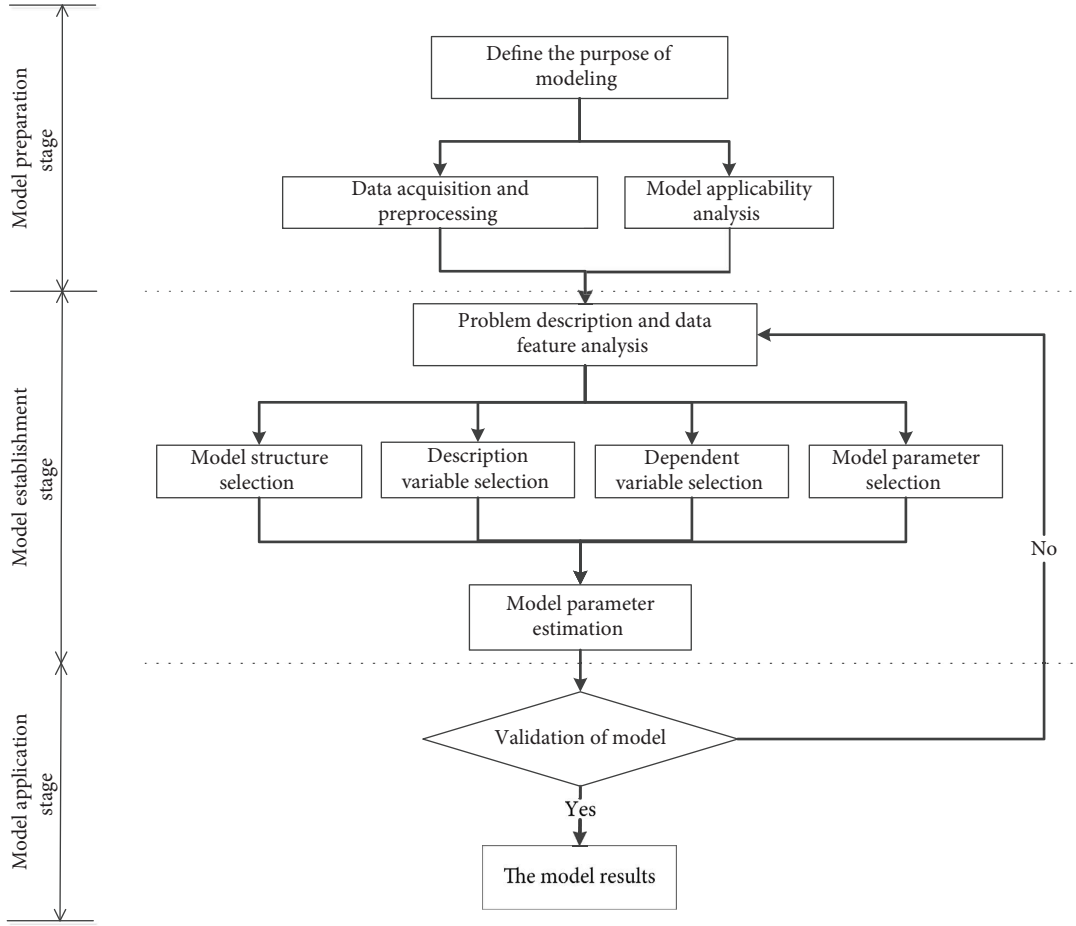

Figure 3: Framework and process of highway crash risk forecasting model.

a new stochastic model with time sequence data. With the introduction of time factor, the data formed at different moments reflect the change and development law of descriptive variables. For Bayesian networks, the key problem is to make probabilistic inference about the hidden states of a group of random variables, and the random variables representing the hidden states in dynamic Bayesian networks have the characteristics of time series. These observed samples can be represented in terms of decomposition or distribution. In addition, because dynamic Bayesian network is a typical directed acyclic graph model, the conditional probability distribution of each node in it can be estimated independently. Therefore, the dynamic Bayesian network model is easier to explain and learn [27].

As a more general spatial-temporal state analysis model, the dynamic Bayesian network model actually extends the static Bayesian network to stochastic process model with time factor. Such an extension would make the distribution of random variables very complicated and difficult to solve. In order to facilitate modeling and solving, it is generally necessary to make simplified treatment and necessary condition assumption [28]. First, the conditional probability process is assumed to be uniformly stable for all T in finite time. Second, the dynamic probabilistic process is assumed to be Markov. That is, future satisfaction $\mathrm{P}(X[t+1] \mid X[1], X[2], \ldots, X[t]) = \mathrm{P}(X[t+1] \mid X[t])$. Finally, the conditional probabilistic process of adjacent time is assumed to be stationary. That is, $\mathrm{P}(X[t+1] \mid X[t])$ has nothing to do with the time t.

Based on the above assumptions, the dynamic Bayesian network model is defined as a random process of joint probability distribution on time trajectory. It consists of a pair of states $(B_0, B_{\rightarrow})$. B_0 is defined as a priori network, which is used to describe the joint probability $P(X_1)$ on the initial state. $B_{\rightarrow}$ is defined as the transfer network, which is used to describe the variable transfer probability $P(X_{t+1} \mid X_t)$. The network graphical expression is shown in Figure 4.

When there are only two time series, $(B_0, B_{\rightarrow})$ is a Bayesian network with two time series. It includes the functions of both transition probability and observation probability models. The node probability of this Bayesian network containing two time series can be calculated by

$$P(X_t \mid X_{t-1}) = \prod_{i=1}^{N} P\left(X_t^{(i)} \mid Pa\left(X_t^{(i)}\right)\right), \tag{1}$$

where $X_t^{(i)}$ is the ith node (including hidden node and observation node, $N = N_h + N_o$) in time series T. $Pa(X_t^{(i)})$ is the parent node of node $X_t^{(i)}$. The parent node contains not only variables at time T but also variables at time $t-1$. In

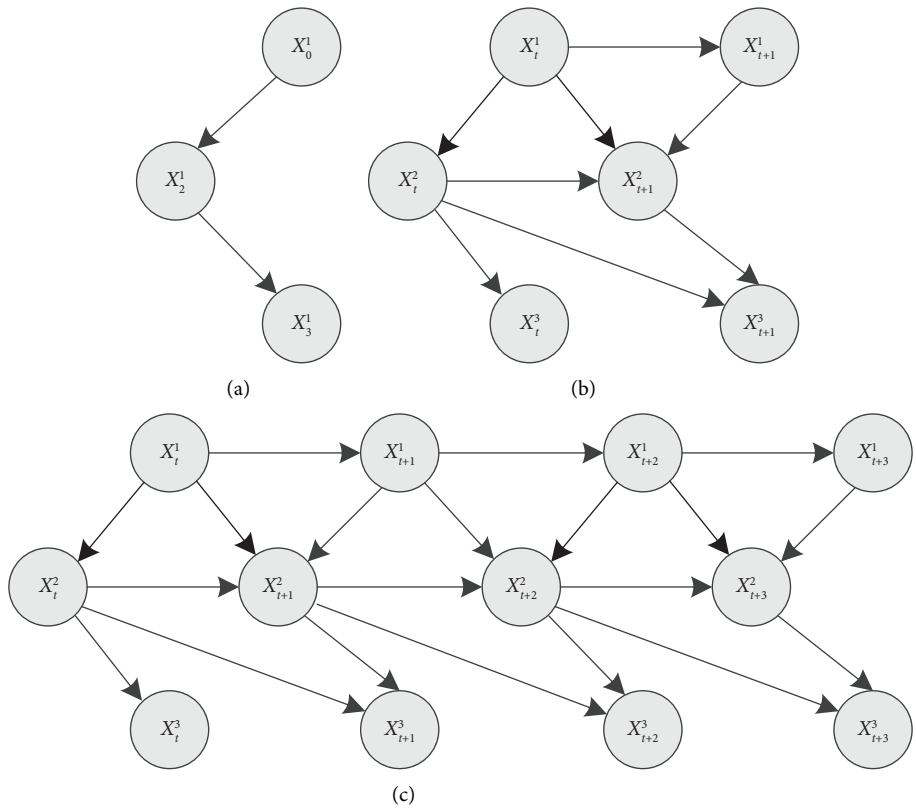

Figure 4: Graphic representation of dynamic Bayesian network.

a Bayesian network containing two time series, the first time series has no parameters, and only the nodes in the second time series have node probability parameters. Similarly, for dynamic Bayesian networks containing T time series, the node probability distribution can be calculated by

$$P\left(X_{1:T}^{(1:N)}\right) = \prod_{i=1}^{N} P_{B_0}\left(X_1^{(i)} \mid Pa\left(X_1^{(i)}\right)\right) \times \prod_{t=2}^{T} \prod_{i=1}^{N} P_{B_{\rightarrow}}\left(X_t^{(i)} \mid Pa\left(Z_t^{(i)}\right)\right). \tag{2}$$

In dynamic Bayesian networks, the hidden state of time series T is represented by a series of random variables, denoted as $H_t^{(i)}, i \in \{1, \ldots, N_h\}$. Meanwhile, the observation state can also be represented by a series of random variables, denoted as $E_t^{(j)}, j \in \{1, \ldots, N_o\}$. Each hidden and observed state variable can be a discrete or continuous variable. In spatial-temporal state analysis models such as hidden Markov model and state space model, there is usually a transition probability $P(H_t \mid H_{t-1})$, an observation probability $P(E_t \mid H_t)$, and an original state distribution $P(H_1)$. However, this kind of model cannot accurately describe the causal relationship between variables, while the dynamic Bayesian network model can consider both the causal relationship between variables and the dynamic change of the causal relationship caused by time factors, which are more suitable for analyzing the highway crash risk considering the space-time characteristics. Therefore, the structure diagram of dynamic Bayesian network model can be given as shown in Figure 5.

As shown in Figure 5, nodes are used to represent hidden state variables or observed state variables. The hidden state uses a discrete random variable to represent the probability of picking every possible value. The hidden state variable H_t is a binary variable in the highway crash risk model. The crash and non-crash traffic flow variables are considered as the observed variable E_t. It should be pointed out that when only a time segment is considered, it is static Bayesian network structure. When multiple time segments are considered, it is a dynamic Bayesian network structure. The figure takes two time series as an example. Its network structure is time series dimension extension based on static network structure. The connecting lines between nodes are also divided into two types in DBN. One is the connection line within the same time segment, which represents the instantaneous correlation between variables and is represented by a solid line. The other is the connection line between different time segments, which is used to describe

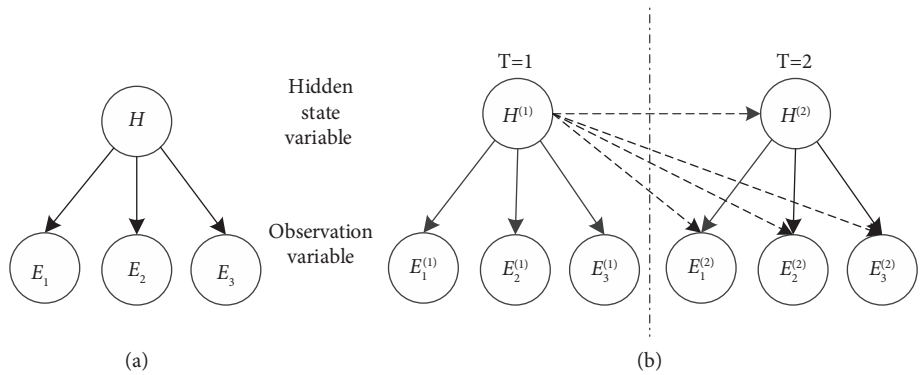

Figure 5: Structure diagram of dynamic Bayesian network model. (a) SBN. (b) DBN.

the transition between the crash states of two time series, and is represented by dashed lines.

Theoretically speaking, the more the input variables of the model, the higher the model accuracy. However, the complexity of the corresponding model is higher, and the solving time and computing resources required for the model will also increase. Relevant research results show that, when the number of modeling variables exceeds a certain number, the improvement range of model accuracy is very small, which is far less than the negative effect caused by the increase of model solving time. Through several modeling experiments, it is found that three variable indexes are considered the optimal scheme when establishing dynamic Bayesian network model. It can not only ensure the accuracy of the model but also achieve the highest computational efficiency. Therefore, the three variables with the highest correlation with road accident risk were selected as modeling indicators in this study.

According to the different characteristics of specific research fields, the application of Bayesian network modeling technology is mainly designed from the following three aspects.

The first step is describing the variables of the research problem and their value range. In this study, hidden variables (including crash state and non-crash state) and observation variables (including traffic flow, speed, and occupancy on upstream and downstream detectors) are mainly included.

The second step is structural learning that represents the dependencies between variables. For dynamic Bayesian networks, structural learning should not only consider the causal relationship between variables in the same time segment but also consider the causal relationship between variables in different time segments. Therefore, static Bayesian network structure learning algorithms such as mountain climbing algorithm, simulated annealing algorithm, and genetic algorithm cannot be directly used in dynamic Bayesian network structure learning [29]. According to the above model variable selection results, it can be seen that there are not many traffic observation variables affecting crash risk. Also, there is a clear causal relationship between variables, so the network structure can be given directly. In addition to observation variables and state variables, the number of time segments is a relatively important influencing factor in the establishment of dynamic Bayesian network model.

The third step is to learn the parameter estimation of conditional probability distribution between observed variables. Dynamic Bayesian network parameter learning is similar to static Bayesian parameter learning algorithm. In order to avoid the error of parameter estimation caused by missing traffic flow observation variables, the expectation-maximization algorithm is used to estimate the maximum likelihood of parameters [30].

4. Case Study

4.1. Results

4.1.1. Variable Selection Results Based on Random Forest Model. There is usually a few minutes delay between the recorded time and the actual time of the crash occurrence [31]. Therefore, traffic flow data 5–10 minutes before the crash are used as the basic variables of the random forest model. This paper uses R language computing platform to realize the random forest model program. The 30 traffic flow variables were taken as the initial variables, and the traffic flow data 5–10 minutes before the crash were taken as the sample data. The random forest algorithm was used to calculate the importance of each variable. The calculation results are shown in Figure 6, where the horizontal axis is the name of the variable, and the vertical axis represents the average model accuracy reduced by the variable, namely, the importance of the variable.

It can be seen from the figure that the index with the largest average variation of model classification accuracy is the downstream average speed. The second is the average occupancy rate of upstream detector, and the average variation of accuracy is between 0.03 and 0.035. The effect of the difference index of upstream and downstream detectors on the model classification accuracy is obviously weaker than the detector original index and mean index. At the same time, it can be seen that for the same type of indicators, the impact of speed and density indicators on model accuracy is significantly higher than that of flow indicators. The reason for this result may be that the data used for modeling are

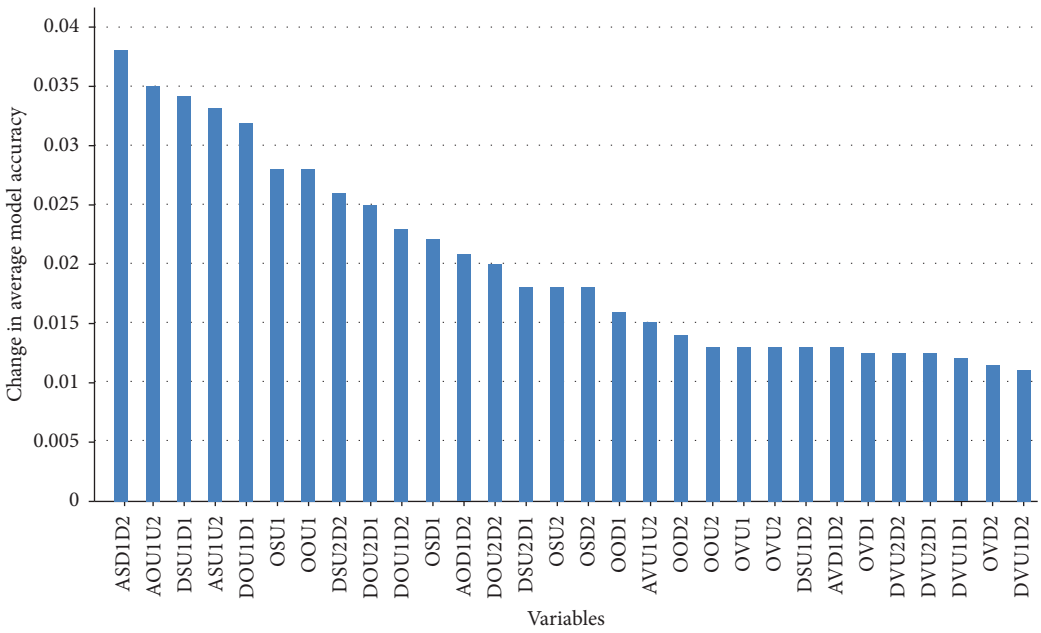

Figure 6: Ranking of the importance of variables.

5 min accumulated flow data, while the speed and density indicators are 5 min average data. This makes the data significantly different at the dimensional level, which may cause variables to have different effects on the model. According to the ranking results of the importance of variables, the variables with high importance should be selected as the input variables of the crash risk model. Meanwhile, in order to reduce the complexity of the model, the number of input variables should not be too much. Therefore, the three variables with the highest importance are selected as the input variables of the prediction model, namely, the mean speed variable ASD1D2 of downstream detector, the mean occupancy variable AOU1U2 of upstream detector, and the speed difference variable DSU1D1 of upstream and downstream nearest detector.

4.1.2. Highway Crash Risk Prediction Results Based on Dynamic Bayesian Network. In order to ensure the generalization ability of the model, the sample data were randomly divided into training dataset and validation dataset. The proportion of crash and non-crash sample data in the two datasets remains unchanged, namely, the sample ratio of crash traffic flow to non-crash traffic flow is 1 : 4. The proportion of crash sample data in the dataset is much smaller than that of non-crash sample data. For the classification prediction model based on such unbalanced data, only the overall classification accuracy cannot completely explain the quality of the model. It is necessary to construct a confusion matrix based on the prediction results to illustrate the prediction accuracy of such classification models based on unbalanced samples, and the confusion matrix is shown in Table 2.

The overall prediction accuracy, crash prediction accuracy and non-crash prediction accuracy, and the *F*-value measure of crash prediction can be calculated as the evaluation index of the model prediction validity. Its calculation formula is as follows:

Table 2: The confusion matrix of prediction result.

Actual results	Predicted results	
	Crash	Non-crash
Crash	True crash (T_{crash})	False non-crash ($F_{\text{non-crash}}$)
Non-crash	False crash (F_{crash})	True non-crash ($T_{\text{non-crash}}$)

$$\text{Overall prediction accuracy} = (T_{\text{crash}} + T_{\text{non-crash}})/(T_{\text{crash}} + F_{\text{crash}} + F_{\text{non-crash}} + T_{\text{non-crash}}) * 100\%.$$

$$\text{Non-crash prediction accuracy} = T_{\text{non-crash}}/(F_{\text{non-crash}} + T_{\text{non-crash}}) * 100\%.$$

$$\text{Crash prediction accuracy} = T_{\text{crash}}/(T_{\text{crash}} + F_{\text{crash}}) * 100\%.$$

$$G\text{-value} = \sqrt{\text{crash prediction accuracy} * \text{non} - \text{crash prediction accuracy}}.$$

$$\text{Crash accuracy rate} = T_{\text{crash}}/(T_{\text{crash}} + F_{\text{crash}}) * 100\%.$$

$$\text{Crash recall rate} = T_{\text{crash}}/(T_{\text{crash}} + F_{\text{non-crash}}) * 100\%.$$

$$F\text{-value} = 2 * \text{crash accuracy rate} * \text{crash recall rate}/(\text{crash accuracy rate} + \text{crash recall rate}).$$

From the 1370 groups of sample data collected above (274 groups of crash sample data and 1096 groups of non-crash sample data), 870 groups of data (174 groups of crash sample data and 696 groups of non-crash sample data) were randomly selected as training sample data. Also, the remaining 500 groups of sample data (100 groups of crash samples and 400 groups of non-crash samples) were validation samples. The structure learning and parameter learning process of dynamic Bayesian network is realized by using R language mathematical statistics analysis platform. The validation dataset is input into the trained model to analyze the validity of the predictive evaluation model. The

confusion matrix of single prediction results is shown in Table 3.

Overall prediction accuracy = (72 + 361)/(72 + 39 + 361 + 28) * 100% = 86.6%.

Non-crash prediction accuracy = 361/(28 + 361) * 100% = 92.8%.

Crash prediction accuracy = 72/(72 + 39) * 100% = 64.9%.

G-value = $\sqrt{0.928 * 0.649} = 0.776$.

Crash accuracy rate = 72/(72 + 39) * 100% = 64.9%.

Crash recall rate = 72/(72 + 28) * 100% = 72.0%.

F-value = 2 * 64.9% * 72%/(64.9% + 72%) = 0.682.

In order to reduce the error caused by the randomness of sample data extraction, 10 training datasets and validation datasets were randomly divided from the original dataset. The mean value of multiple model estimates is taken as the final estimate value of the model, and the mean value of multiple validation results is taken as the predicted value of the model to illustrate the validity of the model. The prediction results are shown in Table 4. From the perspective of prediction accuracy, the overall prediction accuracy of dynamic Bayesian network model is above 80%, the prediction accuracy of crash is between 55% and 65%, and the prediction accuracy of non-crash is about 90%. This is mainly because the proportion of crash traffic flow to non-crash traffic flow in the sample data is 1 : 4, and the proportion of non-crash traffic flow sample is significantly higher than that of crash traffic flow sample. Therefore, there will be a phenomenon that the accuracy of non-crash prediction is significantly higher than that of crash prediction. Nevertheless, from the perspective of prediction accuracy index, the predictive ability of dynamic Bayesian network model for traffic crash risk has reached a high level.

4.2. Discussion

4.2.1. Comparative Analysis of Prediction Results. In order to further illustrate the effectiveness of Bayesian network model considering temporal correlation characteristics, the prediction results of dynamic Bayesian network model and static Bayesian network model are compared and analyzed in this study. The data used for the static Bayesian network model include time segment 1, that is, the traffic flow variable sample in 5–10 minutes before the crash. The dynamic Bayesian network uses two time series data, time segment 1 and time segment 2, for model training and verification. The structure learning and parameter learning of static and dynamic Bayesian network models are realized by the R language package BNLearn. The final prediction results of the two models are shown in Figure 7.

As can be seen from Figure 7, the overall prediction accuracy of the dynamic Bayesian model is 1.8% higher than that of the static Bayesian model. The prediction accuracy of the dynamic Bayesian model is 2.1% higher than that of the static Bayesian model. The non-crash prediction accuracy of the dynamic Bayesian model is 1.9% higher than that of the static Bayesian model. The F-value of the dynamic Bayesian model is 0.023 higher than that of the static Bayesian model. The G-value of the dynamic Bayesian model is 0.017 higher than that of the static Bayesian model. Therefore, the prediction results show that the dynamic Bayesian model has better prediction effect than the static Bayesian model for the same sample data.

Table 3: The confusion matrix of single prediction result.

Actual results	Predicted results	
	Crash	Non-crash
Crash	72 (T_{crash})	28 ($F_{\text{non-crash}}$)
Non-crash	39 (F_{crash})	361 ($T_{\text{non-crash}}$)

4.2.2. Influence Analysis of Time Segment Number. As mentioned above, when establishing the dynamic Bayesian network model, the selection of time segment is an important influencing factor in the process of establishing the dynamic Bayesian network model. The 5 min cumulative traffic flow state index is generally accepted as an explanatory variable of highway crash risk. Therefore, the unit time length of each time segment is set to 5 min. In addition, another important variable affecting model performance is the number of time fragments, that is, each dynamic network model is composed of several variables of time fragments. In this paper, traffic flow index data of upstream and downstream detectors were extracted in 30 min before the crash (a total of six time segments). Due to the error of 3–5 minutes between the recording time of the crash sample and the actual time of the crash, time segment 0, that is, the crash sample data accumulated 0–5 minutes before the crash occurred, has a certain error, so it is omitted. If five time series are considered into the model, the dynamic Bayesian network structure will be composed of five static Bayesian networks. Also, there is correlation between adjacent fragment variables, so the computation and complexity of model parameter estimation are very high. If only two or three adjacent time segments are considered, the prediction results cannot fully reflect the influence of temporal correlation between traffic flow observation variables on crash risk.

In order to ensure the operation efficiency and prediction accuracy of the model, this paper establishes 10 dynamic Bayesian network models with 2, 3, 4, and 5 time series, respectively. Among them, there are four combinations of models including two time series, which are, respectively, training models with sample data of series 1 and series 2, series 2 and series 3, series 3 and series 4, or series 4 and series 5. The model containing three time series has three combinations, which are modeled with sample data of series 1, series 2, and series 3, series 2, series 3, and series 4, or series 3, series 4, and series 5, respectively. The model containing four time series has two combinations, which are modeled with sample data in series 1, series 2, series 3, and series 4 or series 2, series 3, series 4, and series 5, respectively. There is only one combination method for the model containing 5 time series, that is, modeling with sample data in series 1, series 2, series 3, series 4, and series 5. Therefore,

Table 4: The confusion matrix of multiple prediction results.

The serial number	Predicted results				
	Overall prediction accuracy (%)	Accuracy of crash prediction (%)	Accuracy of non-crash prediction (%)	*G*-value	*F*-value
1	86.6	64.9	92.8	0.776	0.682
2	82.2	54.8	90.4	0.704	0.586
3	83.4	56.4	93.2	0.725	0.644
4	84.2	59.1	91.7	0.736	0.633
5	86.8	68.5	90.0	0.789	0.656
6	86.2	62.6	93.9	0.767	0.691
7	84.8	60.1	92.0	0.746	0.645
8	85.6	63.7	91.2	0.762	0.644
9	82.6	55.5	91.1	0.711	0.603
10	86.4	61.8	95.6	0.768	0.712
Mean	84.9	60.8	92.3	0.748	0.649

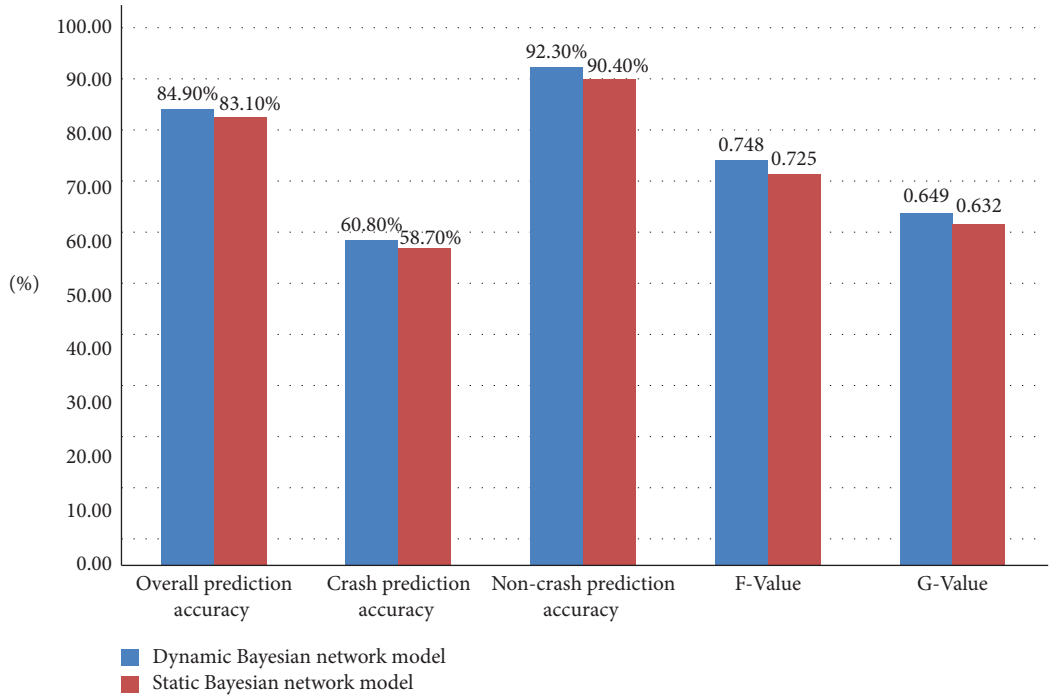

Figure 7: Prediction result comparison between dynamic Bayesian network model and static Bayesian network model.

the influence of the number of time series on the crash risk model can be illustrated by comparing the crash prediction accuracy of 10 dynamic Bayesian network models. The prediction results are shown in Table 5.

As can be seen in Table 5, in terms of prediction accuracy, the prediction results obtained by modeling with different number of time periods are different. The model with the highest accuracy is the dynamic Bayesian network model with the sample data in the time series 1, series 2, and series 3. Meanwhile, it is noted that the prediction accuracy of the model does not increase with the increase of the number of time series. This indicates that in the modeling process, it is not the more the time segments considered, the better the prediction accuracy of the model is. In addition, it is found that the prediction accuracy of the model containing time series 1 is higher than that of the model without time series 1. Also, the prediction accuracy of the model containing both series 1 and series 2 is higher than that of the model without these two time series. This indicates that the variables closer to the crash occurrence time have a greater impact on the crash risk prediction model, and taking them as model training and validation data can effectively improve the prediction accuracy of the model.

Table 5: Prediction results of dynamic Bayesian network model considering different time series.

The model number	Accuracy of prediction results					Contains time series
	Overall prediction accuracy (%)	Accuracy of crash prediction (%)	Accuracy of non-crash prediction (%)	*G*-value	*F*-value	
1	84.9	60.8	92.3	0.748	0.649	12
2	83.8	58.8	93.4	0.741	0.640	23
3	83.5	57.7	91.8	0.728	0.634	34
4	81.7	50.7	88.4	0.670	0.589	45
5	85.9	61.1	92.9	0.753	0.653	123
6	83.5	58.7	91.4	0.733	0.640	234
7	82.5	54.7	90.4	0.703	0.615	345
8	84.8	59.7	92.8	0.744	0.645	1234
9	83.8	56.7	92.1	0.723	0.627	2345
10	85.1	59.8	93.2	0.746	0.646	12345

5. Conclusion

Crash risk analysis and prediction are considered the premise of highway traffic safety control, which directly affects the accuracy and effectiveness of traffic safety decisions. This paper analyzes the influence of temporal correlation characteristics of traffic flow state variables on highway crash risk. A highway crash risk prediction method considering time series correlation feature is proposed. Firstly, the "case-control" analysis method is used to extract crash traffic flow data and corresponding non-crash traffic flow data as sample data for crash risk prediction modeling. Meanwhile, the sample data of half an hour were divided into 6 time segments every 5 minutes as sample data for model training and verification. Then, the random forest model is used to select the traffic flow variables highly correlated with the risk of highway crashes from 30 initial variables. The downstream mean speed variable ASD1D2, the upstream mean occupancy variable AOU1U2, and the speed difference variable DSU1D1 on the upstream and downstream nearest detector are determined as the explanatory variables of the crash risk prediction model. Finally, based on the dynamic Bayesian network modeling method, the highway traffic crash risk prediction model considering the temporal correlation feature is proposed. The validity of the model is illustrated by comparing the prediction accuracy of the model with that of the static Bayesian network model in the test dataset. The results of case study show that the prediction accuracy of the crash risk prediction model considering the temporal correlation features is higher than that of the static Bayesian network method. Also, the prediction model using the first three time series has the best effect.

Although the crash risk prediction model proposed in this study improves the accuracy of crash risk prediction to a certain extent, there is still a lot of room for improvement. Firstly, the influence of time series factors on crash risk prediction model is mainly considered in this model. Besides time factor, space factor is also one of the important factors affecting the accuracy of the model. How to consider the influence of both time and space factors in the process of accident risk modeling is the direction of future research. In addition, the essence of the prediction model proposed in this research is a dichotomous prediction of crash and non-crash traffic flow, so only two kinds of prediction results with or without risk can be obtained. In order to meet the requirements of highway safety risk management practice, more detailed crash risk classification level is needed. In the future, the crash severity index can be considered as the dependent variable of the prediction model. The dichotomy problem can be extended to multiclassification problem to achieve the classification and prediction of highway crash risk level.

Data Availability

All datasets were collected from the Performance Measurement System which can be freely downloaded from https://pems.dot.ca.gov/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Key R&D Program of China (2017YFC0803900).

References

[1] Y. Yang, Z. Yuan, J. Chen, and M. Guo, "Assessment of osculating value method based on entropy weight to transportation energy conservation and emission reduction," *Environmental Engineering & Management Journal*.vol. 16, no. 10, pp. 2413–2424, 2017.

[2] Y. Yang, Z. Yuan, and R. Meng, "Exploring traffic crash occurrence mechanism towards cross-area freeways via an improved data mining approach," *Journal of Transportation Engineering Part A Systems*, vol. 148, no. 9, Article ID 04022052, 2022.

[3] Y. Yang, K. He, Y. P. Wang, Z. Z. Yuan, Y. H. Yin, and M. Z. Guo, "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods," *Physica A: Statistical Mechanics and Its Applications*, vol. 595, Article ID 127083, 2022.

[4] Y. Yang, K. Wang, Z. Yuan, and D. Liu, "Predicting freeway traffic crash severity using XGBoost-bayesian network model with consideration of features interaction," *Journal of Advanced Transportation*, Article ID 4257865, 2022.

[5] W. Wang, Z. Yuan, Y. Yang, X. Yang, and Y. Liu, "Factors influencing traffic accident frequencies on urban roads: a spatial panel time-fixed effects error model," *PLoS One*, vol. 14, no. 4, Article ID e0214539, 2019.

[6] S. Yu, Y. Jia, and D. Sun, "Identify factors that influence the patterns of road-crashes by using association rules: a study case from Wisconsin, United States," *Sustainability*, vol. 11, 2019.

[7] W. Wang, Z. Yuan, Y. Liu, X. Yang, and Y. Yang, "A random parameter logit model of immediate red-light running behavior of pedestrians and cyclists at major-major intersections," *Journal of Advanced Transportation*, vol. 2019, Article ID 2345903, 13 pages, 2019.

[8] Y. Yang, N. Tian, Y. Wang, and Z. Yuan, "A parallel FP-growth mining algorithm with load balancing constraints for traffic crash data," *International Journal of Computers, Communications & Control*, vol. 17, no. 4, p. 4806, 2022.

[9] D. Sun, Y. Ai, Y. Sun, and L. Zhao, "A highway crash risk assessment method based on traffic safety state division," *PLoS One*, vol. 15, no. 1, Article ID e0227609, 2020.

[10] D. Sun, Y. Ai, and L. Wang, "Freeway traffic safety state classification method based on multi-parameter fusion clustering," *Modern Physics Letters B*, vol. 36, no. 20, Article ID 2250088, 2022.

[11] A. H. Lee, K. Wang, J. A. Scott, K. K. Yau, and G. J. McLachlan, "Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros," *Statistical Methods in Medical Research*, vol. 15, no. 1, pp. 47–61, 2006.

[12] T. F. Golob and W. W. Recker, "An analysis of truck-involved freeway accidents using log-linear modeling," *Journal of Safety Research*, vol. 18, no. 3, pp. 121–136, 1987.

[13] T. F. Golob and W. W. Recker, "Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions," *Journal of Transportation Engineering*, vol. 129, no. 4, pp. 342–353, 2003.

[14] T. F. Golob, W. W. Recker, and V. M. Alvarez, "Freeway safety as a function of traffic flow," *Accident Analysis and Prevention*, vol. 36, no. 6, pp. 933–946, 2004.

[15] T. F. Golob and W. W. Recker, "A method for relating type of crash to traffic flow characteristics on urban freeways," *Transportation Research, Part A (Policy and Practice)*, vol. 38, no. 1, 80 pages, 2004.

[16] L. Li, X. Sheng, B. Du, Y. Wang, and B. Ran, "A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction," *Engineering Applications of Artificial Intelligence*, vol. 93, Article ID 103686, 2020.

[17] Y. Lin, L. Li, H. Jing, B. Ran, and D. Sun, "Automated traffic incident detection with a smaller dataset based on generative adversarial networks," *Accident Analysis & Prevention*, vol. 144, Article ID 105628, 2020.

[18] L. Li, C. G. Prato, and Y. Wang, "Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: a sequential approach of multivariate imputation by chained equations and random forest classifier," *Accident Analysis & Prevention*, vol. 146, Article ID 105744, 2020.

[19] L. Li, Y. Lin, B. Du, F. Yang, and B Ran, "Real-time traffic incident detection based on a hybrid deep learning model," *Transportmetrica: Transport Science*, vol. 18, no. 1, pp. 78–98, 2022.

[20] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 5, pp. 291–305, 2010.

[21] S. Roshandel, Z. Zheng, and S. Washington, "Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis," *Accident Analysis & Prevention*, vol. 79, pp. 198–211, 2015.

[22] J. Sun and J. Sun, "Proactive assessment of real-time traffic flow accident risk on urban expressway," *Journal of Tongji University*, vol. 42, no. 6, pp. 873–879, 2014.

[23] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 157–176, 2004.

[24] R. Harb, X. Yan, E. Radwan, and X. Su, "Exploring precrash maneuvers using classification trees and random forests," *Accident Analysis and Prevention*, vol. 41, no. 1, pp. 98–107, 2009.

[25] C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[26] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics and Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[27] C. G. Enright, M. G. Madden, and N. Madden, "Bayesian networks for mathematical models: techniques for automatic construction and efficient inference," *International Journal of Approximate Reasoning*, vol. 54, no. 2, pp. 323–342, 2013.

[28] Q. Xiao and S. Gao, *Application of Bayesian Network in Intelligent Information Processing*, National Defense Industry Press, Beijing, China, 2012.

[29] I. N. Junejo, "Using dynamic Bayesian network for scene modeling and anomaly detection," *Signal, Image and Video Processing*, vol. 4, no. 1, pp. 1–10, 2010.

[30] K. P. Murphy, "The Bayes net toolbox for matlab," *Computational Statistics*, vol. 33, no. 2, pp. 1024–1034, 2001.

[31] H. Hassan and M. A. Abdelaty, "Exploring visibility-related crashes on freeways based on real-time traffic flow data," *Transportation Research Board Meeting*, vol. 11, 2011.