Research Article

Predicting Real-Time Crash Risk for Urban Expressways in China

Miaomiao Liu[1,2] and Yongsheng Chen[1]

[1]*Research Institute of Highway, Ministry of Transport, 8 Xitucheng Road, Haidian District, Beijing 100088, China*
[2]*School of Transportation Science and Engineering, Beihang University, 37 Xueyuan Road, Haidian District, Beijing 100191, China*

Correspondence should be addressed to Miaomiao Liu; liumiao-0605@163.com

We developed a real-time crash risk prediction model for urban expressways in China in this study. About two-year crash data and their matching traffic sensor data from the Beijing section of Jingha expressway were utilized for this research. The traffic data in six 5-minute intervals between 0 and 30 minutes prior to crash occurrence was extracted, respectively. To obtain the appropriate data training period, the data (in each 5-minute interval) during six different periods was collected as training data, respectively, and the crash risk value under different data conditions was defined. Then we proposed a new real-time crash risk prediction model using decision tree method and adaptive neural network fuzzy inference system (ANFIS). By comparing several real-time crash risk prediction methods, it was found that our proposed method had higher precision than others. And the training error and testing error were minimum (0.280 and 0.291, resp.) when the data during 0 to 30 minutes prior to crash occurrence was collected and the decision tree-ANFIS method was applied to train and establish the real-time crash risk prediction model. The prediction accuracy of the crash occurrence could reach 65% when 0.60 was considered as the crash prediction threshold.

1. Introduction

Because of the rapid increase of traffic flow and frequent crash occurrence, traffic safety has become a severe problem for rural roads and urban expressways in China [1]. Modeling real-time crash risk prediction is an important approach to identifying traffic condition causing crash, which can be used in the active traffic management control to reduce traffic accidents and ensure traffic safety. In China, due to the lack of traffic flow detection devices in rural roadways, it is difficult to collect real-time traffic flow data and predict real-time crash risk for these roads. For most of urban expressways in China, traffic detection devices, such as loop detector, microwave sensor, and video detection system, have been well installed. This makes it easier to detect and extract the traffic flow data. Thus, in this study, we mainly focused on urban expressways in China and established the real-time crash risk prediction model for these roads.

Recently, many researchers have analyzed the interrelationship between crash and traffic flow variables using loop detector data or microwave sensor data, and almost all of them emphasized that certain traffic conditions could be associated with high crash likelihood.

In 2001, Oh et al. [2] established the first real-time crash prediction model where they divided traffic dynamics into two conditions: normal and disruptive. Then they applied Bayesian model to assess the likelihood of future traffic flow data falling into these two conditions. In 2005, Oh et al. [3] analyzed 52 crash data variables and corresponding traffic data from loop detectors and identified the real-time crash likelihood by using nonparametric Bayesian approach. These two studies of Oh et al. identified standard deviation of speed to be the most significant variable. In the later study [4], Oh et al. applied Probabilistic Neural Network (PNN) and employed *t*-test on the mean and deviation of three variables, occupancy, flow, and speed, to identify the crash indicator. The results showed that the standard deviation of speed as well as the average occupancy could be considered as the predictors. Then the new real-time crash prediction model was established by randomly selecting 30 crash data variables

from their sample and testing their outcome and repeating the process for 30 times. The threshold value and the accurate prediction rate were, respectively, 38.2% and 44.9%.

In 2002, Lee et al. [5] pointed out the potential of real-time crash prediction to be applied as a proactive road safety management system and used a log-linear model to estimate crash risks based on real-time traffic flow data collected from freeway loop detector stations. They introduced a new concept called crash precursors, which was defined as traffic conditions that exist before the occurrence of a crash. Then Lee et al. [6] basically reduced the number of assumptions they made in the first study to make it more acceptable. It was concluded that the coefficient of variation in speed, traffic density, and speed difference between upstream and downstream loop detector stations were significantly correlated with the crash risk. In the later study, Lee et al. selected speed variations along a lane, traffic queue, and traffic density at given road geometry, weather condition, and time of the day as predictors and applied aggregated first-order log-linear model to predict crash. The developed model was not validated with another dataset and the prediction success was represented with the overall model fit, statistical significance of the coefficients, and the consistency of the coefficients with the order of levels of crash precursors.

In 2004, Abdel-Aty and Pande [7] used a sample size of 148 crashes, of which 100 were used to generate the model and the remaining 48 were used for validation. They used the concept of logistic regression and odds ratio to develop a new index called Hazard ratio, which essentially represents the factor with which the risk of observing a crash in the vicinity of the station of the crash will increase with unit increase in the corresponding risk factor (here, the predictors of crash). Lastly, they used Probabilistic Neural Network (PNN) to distinguish between crash and noncrash situation. They found the coefficient of variation in the speed obtained from the station near the crash and two stations immediately preceding in the upstream direction prior to crash to be the most suitable predictors. Although their study produced by far the best results to predict crashes, the overall classification, that is, for both crash and noncrash situations together, was poor (62%). In a later study, Abdel-Aty and Abdalla [8] used Generalized Estimating Equation method where they included road geometry as variables as well. The study found that high variability in speed for a period of 15 minutes for a specific location increases the likelihood of crash and, also, low variability in volume over 15 minutes at a given location increases the crash likelihood in the downstream. In addition, Abdel-Aty et al. [9] used matched case-control logistic regression to analyze the relationship between crash likelihood and real-time traffic flow characteristics. The analysis results showed that the most significant factors influencing the likelihood of crash occurrence were average occupancy observed at the upstream station and coefficient of variation in speed at the downstream station. In 2005, Abdel-Aty and Pande [10] collected the multiple speed derivatives, including the logarithms of the coefficient of the variation in speed for both crash and noncrash conditions. Then they applied a Bayesian classifier based methodology, the Probabilistic Neural Network (PNN) model, to predict crash

occurrences on freeways and classify the collected data as belonging to either crashes or no-crashes. Pande and Abdel-Aty [11] collected the traffic surveillance data from a pair of dual loop detectors and developed a crash risk prediction model by using the classification tree and neural network. They found that, based on this model, the hazardous traffic conditions prone to lane-change related collisions could be identified.

Recently, Hossain and Muromachi [12] divided expressways into several segments (basic freeway, upstream and downstream of exits, and entrance ramps) and developed separate crash risk prediction models for different segments based on advanced ensemble learning methods such as random forest and classification and regression trees. The results showed that the contributing factors to crash risk were quite different for different road segments. In 2012, Xu et al. [13] conducted a K-means clustering analysis to classify traffic flow into five different states. Then they developed conditional logistic regression models to analyze the relationship between crash risks and traffic states on freeways. The results demonstrated that each traffic state could be assigned with a certain safety level and the effects of traffic flow characteristics on crash risks were different for different traffic states.

The primary objective of this study is to divide freeway traffic flow into different states and to evaluate the safety performance associated with each state. Using traffic flow data and crash data collected from a northbound segment of the I-880 freeway in the state of California, United States, K-means clustering analysis was conducted to classify traffic flow into five different states. Conditional logistic regression models using case-controlled data were then developed to study the relationship between crash risks and traffic states. Traffic flow characteristics in each traffic state were compared to identify the underlying phenomena that made certain traffic states more hazardous than others. Crash risk models were also developed for different traffic states to identify how traffic flow characteristics such as speed and speed variance affected crash risks in different traffic states. The findings of this study demonstrate that the operations of freeway traffic can be divided into different states using traffic occupancy measured from nearby loop detector stations, and each traffic state can be assigned with a certain safety level. The impacts of traffic flow parameters on crash risks are different across different traffic flow states. A method based on discriminant analysis was further developed to identify traffic states given real-time freeway traffic flow data. Validation results showed that the method was of reasonably high accuracy for identifying freeway traffic states.

In 2013, Hosseinpour et al. [14] used adaptive neuro-fuzzy inference system (ANFIS) for modeling traffic accidents as a function of road and roadside characteristics. Then the ANFIS model was compared with the Poisson, negative binomial, and nonlinear exponential regression models. The results showed that road width, shoulder width, land use, and access points significantly affected accident frequencies and the proposed ANFIS model had higher prediction performance than the other three traditional models. Then, Xu et al. [15] applied random parameters logistic regression to

develop a real-time crash risk model and Bayesian inference based on Markov chain Monte Carlo simulations was used for model estimation. The parameters of traffic flow variables in the model were allowed to vary across different traffic states. Compared with the standard logistic regression model, the proposed model significantly improved the goodness-of-fit and predictive performance. In addition, Xu and Qu [16] also showed and analyzed some basic descriptive statistics of TTC (time to collision) samples, and used T-test to analyze the effect of road environments, traffic conditions, and vehicle types on TTC statistically. In 2015, Wang et al. [17] presented a multilevel Bayesian logistic regression model for crashes at expressway weaving segments using crash, geometric, Microwave Vehicle Detection System (MVDS), and weather data. The results showed that the mainline speed at the beginning of the weaving segments, the speed difference between the beginning and the end of weaving segment, and logarithm of volume had significant impacts on the crash risk of the following 5–10 minutes for weaving segments. Sun et al. [18] also utilized Bayesian belief net to build the real-time crash prediction model for the basic freeway segments, and predicted the formation probability of a hazardous traffic condition in 4–9 minutes in a 250-meter-long freeway road section. The analysis results indicated that the proposed method could be used for the urban freeway management departments to understand the risk factors and take immediate actions in advance to avoid traffic accidents on the freeway. In 2016, Shi et al. [19] developed a multi-level Bayesian framework to identify the crash contributing factors on an urban expressway in the Central Florida area. Multilevel and random parameters models were constructed and compared with the negative binomial model under the Bayesian inference framework. The results showed that the models with random parameters could achieve the best model fitting, and lower speed and higher speed variation could significantly increase the crash likelihood on the urban expressway.

As mentioned above, the previous studies have comprehensively analyzed traffic flow characteristics and crash data, established various real-time crash risk prediction models by using different methods, and have made considerable achievements. However, there were quite few studies to analyze real-time traffic flow data for urban expressways in China and establish real-time crash prediction model applicable to Chinese urban expressways. Thus, in this study, we attempted to address these issues by developing a real-time crash risk prediction model with readily available variables and realize real-time risk assessment for urban expressways in China.

Based on decision tree method and adaptive neural network fuzzy inference system (ANFIS), we proposed a new real-time crash risk prediction model. Then we compared several other real-time crash risk prediction methods, such as logistic regression, decision tree, and supported vector machine (SVM). The manuscript was organized into five sections. The introductory section has laid out the background and stated the purpose and objective of the study. Section 2 described the activities involving data extraction and processing. Section 3 defined real-time crash risk, presented

a self-containing introduction to modeling method, and evaluated the established model by comparing the results of several other real-time crash risk prediction methods. Section 4 discussed the model building and evaluation process and summarized the salient contributions and findings of the study along with identifying the limitations and subsequent future scopes.

2. Data Collection and Preparation

To accomplish the research objective, data were obtained from a 39.7-kilometer segment on the Jingha Expressway in Beijing, China. There are 20 microwave detectors and 16 video detectors stations in upstream and downstream directions along the selected expressway section with an average spacing of 1.10 kilometers. The collected traffic flow and crash data was recorded from January, 2013, to October, 2014. A total of 123 crashes were identified and used in the study.

The traffic data were obtained from Huabei Expressway Corporation, LTD. The average speed, volume, and occupancy in 30-second aggregation intervals were collected in each lane. The 30-second raw detector readings from the upstream station were aggregated into 5-minute intervals and converted into the 9 traffic flow variables presented in Notations. The variables in Notations consist of five-minute observations. To identify hazardous traffic condition and make preemptive measures possible [10, 12], we extracted traffic data from the upstream station in six 5-minute intervals between 0 and 30 minutes prior to crash occurrence. For example, if a crash occurred at 8:00 am, the traffic data were extracted from 7:30 to 8:00 am, and six five-minute intervals were 7:30–7:35 am, 7:35–7:40 am, 7:40–7:45 am, 7:45–7:50 am, 7:50–7:55 am, and 7:55–8:00 am, respectively. For each crash in the dataset, the researchers selected two 30-minute traffic data (six five-minute intervals) without crashes from the crash-free days during the same period. These intervals were supplemented with the 9 traffic data variables to form crash-free observations.

3. Methodology

3.1. Defining Real-Time Crash Risk. To obtain the appropriate data training period, the data (in each 5-minute interval) during six different periods (including 0 to 5 minutes, 0 to 10 minutes, 0 to 15 minutes, 0 to 20 minutes, 0 to 25 minutes, and 0 to 30 minutes prior to crash occurrence) was collected as training data, respectively, and the crash risk value under different data conditions was defined. In this study, we assumed that the closer to the crash occurrence, the higher the crash risk, and the crash risk value revealed a linear decreasing trend from the first 5-minute interval prior to crash occurrence to the last interval. In addition, we considered that the crash risk value in the first and the last 5-minute interval prior to crash occurrence was 1 and 0, respectively.

That is to say, if we extracted traffic data (in each 5-minute interval) during 0 to 5 minutes prior to crash occurrence (i.e., the first 5-minute interval prior to crash occurrence) as training data, the crash risk for a crash case in this period

could be considered as 1 and the crash risk for a noncrash case could be considered as 0.

If we extracted traffic data during 0 to 10 minutes prior to crash occurrence (i.e., the first and the second 5-minute intervals prior to crash occurrence) as training data, the crash risk for a crash case in this period could be considered as 1 and 0 during the first and the second 5-minute intervals prior to crash occurrence, respectively, and the crash risk for a noncrash case could be considered as "0" for the two 5-minute intervals.

If we extracted traffic data during 0 to 15 minutes prior to crash occurrence (i.e., the first and the third 5-minute intervals prior to crash occurrence) as training data, the crash risk for a crash case in this period could be considered as 1, 1/2, and 0 during the first to the third 5-minute intervals, respectively, and the crash risk for a noncrash case could be considered as 0 for the three 5-minute intervals.

If we extracted traffic data during 0 to 20 minutes prior to crash occurrence (i.e., the first and the fourth 5-minute intervals prior to crash occurrence) as training data, the crash risk for a crash case in this period could be considered as 1, 2/3, 1/3, and 0 during the first to the fourth 5-minute intervals, respectively, and the crash risk for a noncrash case could be considered as 0 for all the four 5-minute intervals.

If we extracted traffic data during 0 to 25 minutes prior to crash occurrence (i.e., the first and the fifth 5-minute intervals prior to crash occurrence) as training data, the crash risk for a crash case in this period could be considered as 1, 3/4, 1/2, 1/4, and 0 during the first to the fifth 5-minute intervals, respectively, and the crash risk for a noncrash case could be considered as 0 for all the five 5-minute intervals.

If we extracted traffic data during 0 to 30 minutes prior to crash occurrence (i.e., the first and the sixth 5-minute intervals prior to crash occurrence) as training data, the crash risk for a crash case in this period could be considered as 1, 4/5, 3/5, 2/5, 1/5, and 0 during the first to the sixth 5-minute intervals, respectively, and the crash risk for a noncrash case could be considered as 0 for all the six 5-minute intervals.

3.2. Modeling Method

3.2.1. Identifying Main Factors Influencing the Crash Risk Based on Decision Tree Method.
To identify the most important variables influencing real-time crash risk, decision tree method was used to analyze the relationship between traffic variables and real-time crash risk. Decision trees or classification trees are among the popular statistical tools that emerged from the field of machine learning and data mining. Classification trees classify observations by recursively partitioning the predictor space. The resultant model can be expressed as a hierarchical tree structure. Especially since the introduction of the classification and regression trees (CART) [20], decision trees have received wide use in a variety of fields because of their nonparametric nature and easy interpretation [21].

In the traffic field, the application of decision trees is also extensive. For instance, De Oña et al. [22] employed decision tree method to identify the key factors that affected bus transit quality of service and to compare the key attributes identified

TABLE 1: The main factors influencing the crash risk under different data training conditions.

Data selection period (prior to crash occurrence)	The main factors influencing the crash risk
0 to 5 min	$F_{\text{avg}}, F_{\text{sd}}, F_{\text{vc}}, O_{\text{vc}}, O_{\text{sd}}$
0 to 10 min	$F_{\text{avg}}, F_{\text{sd}}, F_{\text{vc}}, O_{\text{vc}}, O_{\text{sd}}$
0 to 15 min	$F_{\text{avg}}, F_{\text{sd}}, F_{\text{vc}}, O_{\text{vc}}$
0 to 20 min	$F_{\text{avg}}, F_{\text{sd}}, F_{\text{vc}}, O_{\text{avg}}, O_{\text{sd}}, O_{\text{vc}}, V_{\text{vc}}$
0 to 25 min	$F_{\text{avg}}, F_{\text{sd}}, F_{\text{vc}}, O_{\text{avg}}, O_{\text{sd}}, O_{\text{vc}}$
0 to 30 min	$F_{\text{avg}}, F_{\text{sd}}, F_{\text{vc}}, O_{\text{vc}}$

before and after passengers reflect on the main aspects of the system. Using 2005 to 2006 truck-involved accident data from national freeways in Taiwan, Chang and Chien [23] developed a nonparametric decision tree model to establish the empirical relationship between injury severity outcomes and driver/vehicle characteristics, highway geometric variables, environmental characteristics, and accident variables.

In this study, we chose decision tree method to analyze the main factors that affected real-time crash risk. SPSS software package (version 13.0; SPSS Inc., Chicago, IL, USA) was used to conduct decision tree analyses. Then we considered all of the variables in Notations as input parameters and took the crash risk value (as defined in Section 3.1) as the output parameter. Because the CART method could avoid overfitting the model by "pruning the tree," all decision trees in this study were developed based on the CART approach. The Gini criterion was used as a measure of split criteria. All trees were trimmed automatically to the smallest subtree based on one standard error as the specified maximum difference in risk. Since the data size is not very large, the minimum number of cases for parent nodes was set as 10 and the minimum number of cases for child nodes was set as 3. By using SPSS, we could obtain hierarchical tree structures, as shown in Figure 1, and find out the main factors influencing the crash risk. Table 1 shows the main factors influencing the crash risk under different data training conditions (as shown in Table 1). For detailed structure of decision tree, see De Oña et al. [22].

3.2.2. Establishing the Real-Time Crash Risk Prediction Model Based on ANFIS.
Generally speaking, real-time crash risk exhibits nonlinear characteristics because of the effects of various factors. It is difficult to describe the real-time crash risk using one particular formula. Previous studies have demonstrated the general use of neural networks in nonlinear mapping, reasoning, and prediction [24]. However, a neural network has one disadvantage; that is, we cannot effectively obtain the implicit rules in a network structure. For a fuzzy logic system, it can be used to model human perception in an uncertain and imprecise environment. However, the fuzzy logic system is more complex; thus, it is difficult for the human brain to understand the causality existing in such system [25]. According to recent literature [26], an adaptive neuro-fuzzy inference system (ANFIS) is a combination of neural network and fuzzy logic approaches; hence, it inherently has the advantages of both, such as having a good
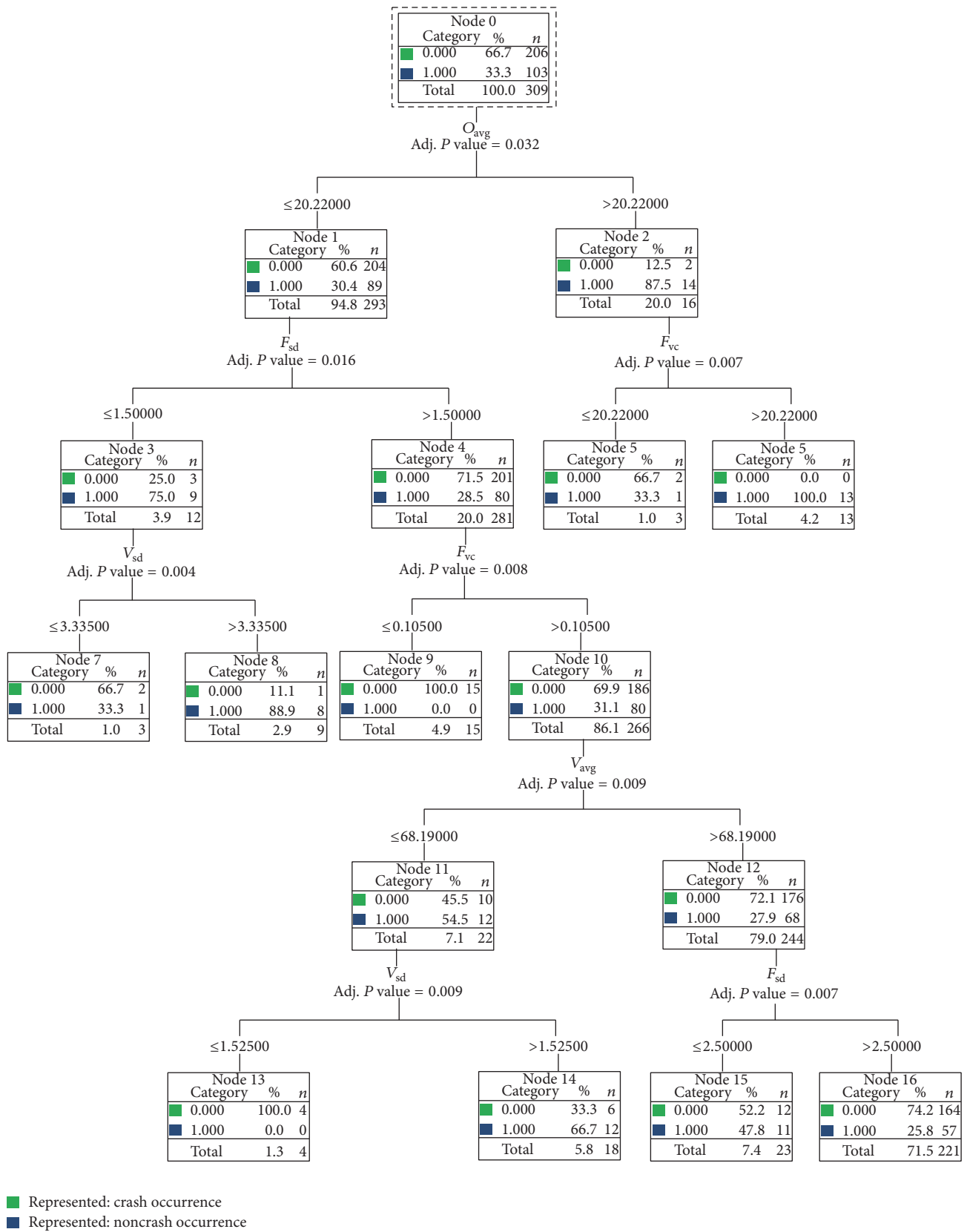
FIGURE 1: Decision tree result based on the data during 0 to 5 min prior to crash occurrence.
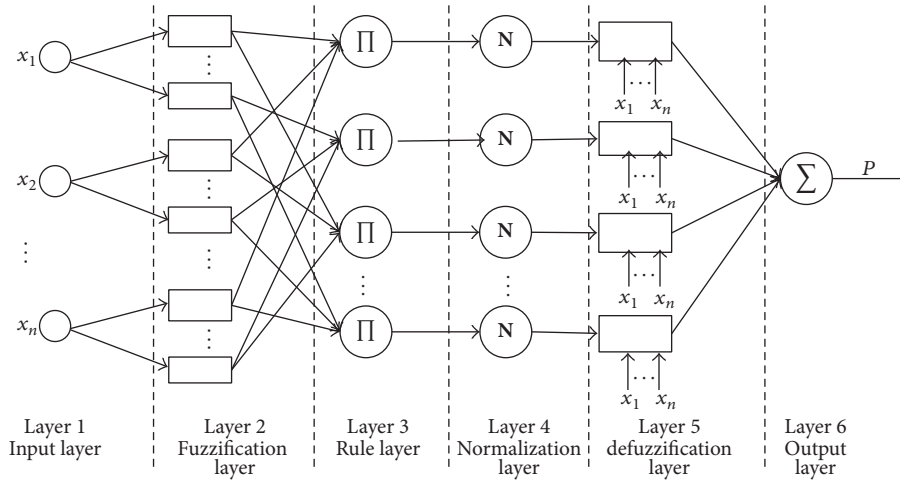
Figure 2: The structure of ANFIS.

learning mechanism and reasoning capability. Accordingly, we can adopt ANFIS to model the real-time crash prediction for urban expressway in China.

In general, ANFIS has a six-layer feedforward neural network structure. Figure 2 shows the ANFIS structure for a model with n inputs: $x_1, x_2, \ldots, x_n$ and one output: P. Explanation of each layer is as follows.

Layer 1 is the input layer. Neurons in layer 1 pass the input signals to layer 2. Layer 2 is the fuzzification layer. In this layer, the input variables can be divided into linguistic variables. Layer 3 is called the rule layer. A corresponding Sugeno-type fuzzy rule exists in each neuron in this layer. Layer 4 is the normalization layer. Each neuron in layer 4 accepts inputs from all neurons in layer 3 and figures out the normalized firing strength (i.e., the ratio of the firing strength of a given rule to the sum of firing strengths of all rules). Layer 5 is called the defuzzification layer. The weighted consequent value is determined for a given rule in the processing of the defuzzification. Layer 6 is the output layer. The overall output P can be determined by summing all the outputs from layer 5. For detailed structure of ANFIS, see Jang et al. [27].

To obtain the more accurate real-time crash risk prediction model, we used the main factors influencing the crash risk as the input variables (as shown in Table 1) and the crash risk as the output variable to train ANFIS of real-time crash risk in this study. All the input variables were fuzzy variables, which should be described and measured using linguistic rather than precise numerical values [26]. In this study, each input variable was divided into the following linguistic variables: negative big (NB), negative small (NS), zero (ZO), positive small (PS), and positive big (PB). The membership functions of all the input variables were initially supplied exogenously. The output variable was the value of real-time crash risk.

Then we used the fuzzy logic toolbox in MATLAB to develop and test the real-time crash risk prediction models under five different data training conditions. The specific steps were described as follows.

Step 1. Generate and input training and testing data. We chose 103 crash cases and 206 noncrash cases as training samples for this study. The parameter data obtained from the other 20 crash cases and 40 noncrash cases were used as testing data.

Step 2. Identify the type of membership functions. In this study, a Gaussian membership function was selected to fuzzify all the input variables.

Step 3. Use the "genfis1" function to generate the original fuzzy inference structure (FIS).

Step 4. Set the training parameters of ANFIS.

Step 5. Use the "anfis" function to train ANFIS.

Step 6. Use the "evalfis" function to test the obtained performance of FIS.

Step 7. Determine whether the model simulation results meet the requirements. If yes, the established model is the real-time crash risk prediction model; if no, it is needed to adjust the parameters of the membership function until the model simulation to achieve the ideal effect.

3.3. Model Result. According to the method described in Section Modeling method, we could obtain the input-output curves and the training step-error curves of the real-time crash risk prediction model (based on the data during 0 to 5 minutes prior to crash occurrence), as shown in Figures 3 and 4. And the training error could also be calculated from the MATLAB program.

Similarly, we could obtain all the input-output curves and the training step-error curves of the real-time crash risk prediction model under other data selection conditions.

3.4. Model Test

3.4.1. Comparison Analysis. To determine the validity and accuracy of our model, we selected several other real-time crash risk prediction methods, including logistic regression,

(a) The values of input variables

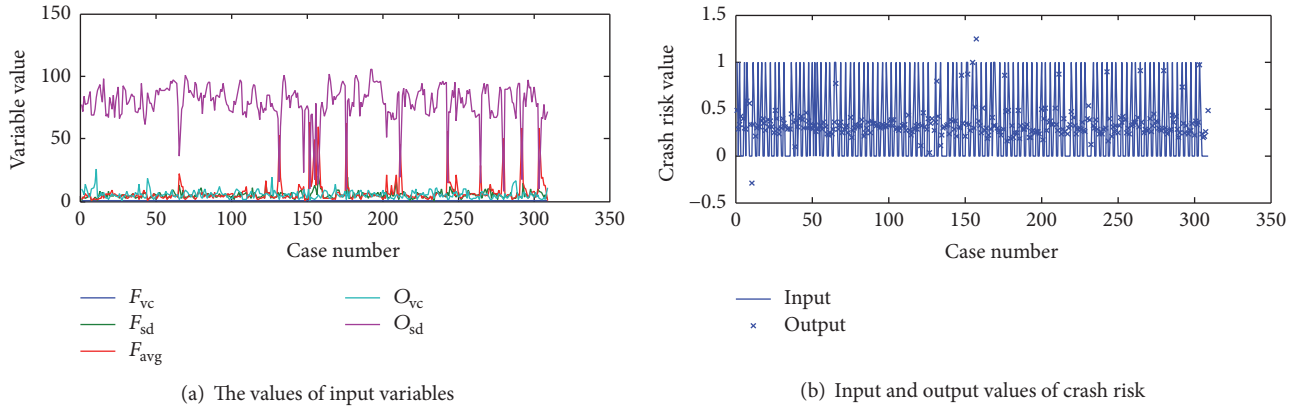(b) Input and output values of crash risk

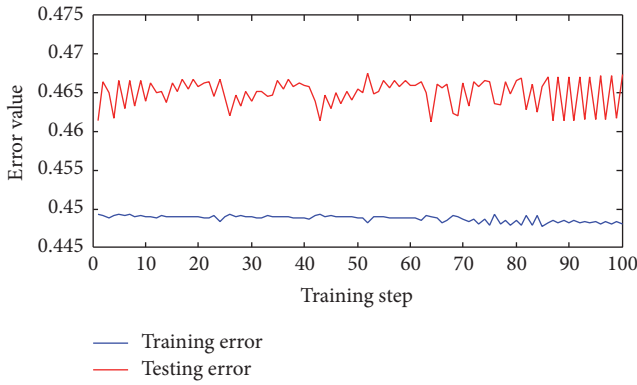FIGURE 3: Input-output curves of the real-time crash risk prediction model.

FIGURE 4: Training step-error curves of the real-time crash risk prediction model.

decision tree, and SVM, to establish the real-time crash risk prediction models and compared the model results. Similarly, for different methods, the same 103 crash cases and 206 noncrash cases were used as training samples, and the same 20 crash cases and 40 noncrash cases were used as testing data.

Modeling with Logistic Regression. As we had already separated the collected data into two categories (crash data and noncrash data), we could build a binary logistic regression model based on that. The crash indicator was 1 if a crash occurred. If noncrash occurred, the crash indictor was 0. The probability of a crash (i.e., the crash risk in this study) could be estimated according to R-package. For detailed instruction of this method, see [28].

Modeling with Decision Tree. As mentioned above, decision tree method had been used to analyze the main factors that affected real-time crash risk in this study. In addition, decision tree method could also be utilized to establish real-time crash risk prediction model. Similarly, we considered all of the variables in Notations as input parameters and took the crash risk value as the output parameter. And the CART approach was used to develop decision trees and the Gini criterion was used as a measure of split criteria. By using SPSS,

we could obtain hierarchical tree structures and save model parameters and model rules under different data training conditions. Then model results could be calculated by calling parameters and rules of the model. For detailed information about decision tree, see De Oña et al. [22].

Modeling with SVM. Support vector machine (SVM) was originally designed based on statistical learning theory and the structural risk minimization. With SVM, we could find a separating hyperplane by minimizing the distance of misclassified points to the decision boundary. The linear kernel was considered in modeling, and then the SVM models for predicting real-time crash risk could be established by using R-package. Detailed instruction of SVM could be seen in [29].

According to the methods above, we could obtain different real-time crash risk models. Table 2 shows the errors of different models under different data training conditions.

Table 2 showed that, in most cases, both training error and testing error based on our proposed method (decision tree-ANFIS) were smaller comparing with other methods. And the training error and testing error were minimum (0.280 and 0.291, resp.) when the data during 0 to 30 minutes prior to crash occurrence was collected and the decision tree-ANFIS method was applied to train and establish the real-time crash risk prediction model. In other words, our proposed method had higher precision than others in most cases, which might be more appropriate to predict real-time crash risk for urban expressways of China.

3.4.2. Prediction Effect. On the basis of our proposed decision tree-ANFIS model, the predicted crash risk value for the 20 crash and 40 noncrash testing cases could be obtained. Figure 5 shows the observed and predicted crash risk value for parts of the testing cases.

It was seen from Figure 5 that, according to our proposed model, the predicted crash risk value could reflect the change of the actual crash risk very well, and when the crash risk value was predicted to reach 0.60, one crash usually occurred. Thus, we could define the threshold of the real-time crash risk prediction as 0.60. Once the predicted crash risk was higher than 0.60, we could consider that a crash would happen. For the 20 crash testing cases, the testing results showed that,

TABLE 2: The errors of different models under different data training conditions.

| Data selection period (prior to crash occurrence)	Logistic regression		Decision tree		SVM		Decision tree-ANFIS	
	Training error	Testing error	Training error	Testing error	Training error	Testing error	Training error	Testing error
0 to 5 min	0.457	0.462	0.456	0.541	0.450	0.427	0.448	0.466
0 to 10 min	0.462	0.465	0.361	0.485	0.364	0.359	0.361	0.383
0 to 15 min	0.463	0.478	0.320	0.334	0.325	0.322	0.322	0.337
0 to 20 min	0.465	0.459	0.303	0.328	0.314	0.314	0.310	0.313
0 to 25 min	0.466	0.490	0.299	0.307	0.303	0.312	0.303	0.309
0 to 30 min	0.464	0.458	0.292	0.303	0.302	0.302	0.280	0.291
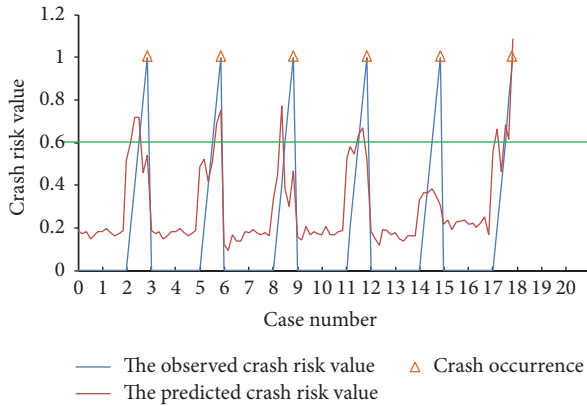
FIGURE 5: The observed and predicted crash risk value for parts of testing cases.

if we used 0.60 as the crash prediction threshold, 13 crash cases could be predicted. The prediction accuracy of the crash occurrence was 65.0%. In addition, for the 40 noncrash cases, only 3 cases of them were predicted as "crash." That is, the false alarm rate for predicting crash occurrence was 7.5%. However, the higher crash prediction accuracy and operation efficiency of the proposed model indicated that we could utilize decision tree-ANFIS method to assess the real-time crash risk for urban expressways in China.

4. Conclusion

In this study, we aimed to predict real-time crash risk for urban expressways in China and identify traffic condition causing crash. Based on decision tree method and ANFIS, we proposed a new real-time crash risk prediction model. Decision tree method was used to identify the most important variables influencing real-time crash risk. ANFIS was applied to establish the real-time crash risk prediction model.

To obtain the appropriate data training period, the data (in each 5-minute interval) during six different periods was collected as training data, respectively, and the crash risk value under different data conditions was defined. Then we used decision tree-ANFIS method to establish the real-time crash risk prediction models under different data training conditions. By comparing the results of other three real-time crash risk prediction methods (including logistic regression, decision tree, and SVM), we validated that

(1) in most cases, our proposed model had smaller training error and testing error than other models; it indicated that the model we established had higher precision, which might be more suitable to predict the real-time crash risk on urban expressways in China;

(2) the model error was minimum when the data during 0 to 30 minutes prior to crash occurrence was chosen and our proposed model was used to establish the real-time crash risk prediction model;

(3) according to our proposed method, the prediction accuracy of the crash occurrence could reach 65.0%, and the false alarm rate was 7.5%;

(4) this study can be applied to monitor real-time traffic risk on urban expressways in China, forecast the crash occurrence promptly, and assist traffic control decisions such as variable speed limit and warning messages through variable message signs to enhance safety.

Nevertheless, this study exhibited several limitations. We did not analyze the influence of geometric design and weather condition on the real-time crash risk prediction. Furthermore, time related variables were not considered in modeling. In our future research, we will examine more urban expressways to analyze the effects of various factors on real-time crash risk prediction.

Notations

Variables Considered for the Models

F_{avg}: Average 30-second vehicle counts at the upstream station (veh/30 s)

F_{sd}: Standard deviation of 30-second vehicle counts at the upstream station (veh/30 s)

F_{vc}:   Variation coefficient of 30-second vehicle counts at the upstream station (%)

V_{avg}:   Average 30-second speed at the upstream station (km/h)

V_{sd}:   Standard deviation of 30-second speed at the upstream station (km/h)

V_{vc}:   Variation coefficient of 30-second speed at the upstream station (%)

O_{avg}:   Average 30-second occupancy at the upstream station (%)

O_{sd}:   Standard deviation of 30-second occupancy at the upstream station (%)

O_{vc}:   Variation coefficient of 30-second occupancy at the upstream station (%).

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

References

[1] A. Montella, F. Galante, F. Mauriello, and L. Pariota, "Effects of traffic control devices on rural curve driving behavior," *Transportation Research Record*, vol. 2492, pp. 10–22, 2015.

[2] C. Oh, J. Oh, S. Ritchie, and M. Chang, "Real time estimation of freeway accident likelihood," in *Proceedings of the 80th Annual Meeting of Transportation Research Board*, Washington, DC, USA, 2001.

[3] J.-S. Oh, C. Oh, S. G. Ritchie, and M. Chang, "Real-time estimation of accident likelihood for safety enhancement," *Journal of Transportation Engineering*, vol. 131, no. 5, pp. 358–363, 2005.

[4] C. Oh, J.-S. Oh, and S. G. Ritchie, "Real-time hazardous traffic condition warning system: framework and evaluation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 3, pp. 265–272, 2005.

[5] C. Lee, F. Saccomanno, and B. Hellinga, "Analysis of crash precursors on instrumented freeways," *Transportation Research Record*, no. 1784, pp. 1–8, 2002.

[6] C. Lee, F. Saccomanno, and B. Hellinga, "Real-time crash prediction model for the application to crash prevention in freeway traffic," *Transportation Research Record*, vol. 1840, pp. 67–77, 2003.

[7] M. Abdel-Aty and A. Pande, "Classification of real-time traffic speed patterns to prediction crashes on freeways," in *Proceedings of the 83th Annual Meeting of Transportation Research Board*, Washington, DC, USA, 2004.

[8] M. Abdel-Aty and M. F. Abdalla, "Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes: generalized estimating equations for correlated data," *Transportation Research Record*, no. 1897, pp. 106–115, 2004.

[9] M. Abdel-Aty, N. Uddin, A. Pande, M. F. Abdalla, and L. Hsia, "Predicting freeway crashes from loop detector data by matched case-control logistic regression," *Transportation Research Record*, no. 1897, pp. 88–95, 2004.

[10] M. Abdel-Aty and A. Pande, "Identifying crash propensity using specific traffic speed conditions," *Journal of Safety Research*, vol. 36, no. 1, pp. 97–108, 2005.

[11] A. Pande and M. Abdel-Aty, "Assessment of freeway traffic parameters leading to lane-change related collisions," *Accident Analysis and Prevention*, vol. 38, no. 5, pp. 936–948, 2006.

[12] M. Hossain and Y. Muromachi, "Understanding crash mechanisms and selecting interventions to mitigate real-time hazards on urban expressways," *Transportation Research Record*, vol. 2213, pp. 53–62, 2011.

[13] C. Xu, P. Liu, W. Wang, and Z. Li, "Evaluation of the impacts of traffic states on crash risks on freeways," *Accident Analysis and Prevention*, vol. 47, pp. 162–171, 2012.

[14] M. Hosseinpour, A. S. Yahaya, S. M. Ghadiri, and J. Prasetijo, "Application of adaptive neuro-fuzzy inference system for road accident prediction," *KSCE Journal of Civil Engineering*, vol. 17, no. 7, pp. 1761–1772, 2013.

[15] C. Xu, W. Wang, P. Liu, and F. Zhang, "Development of a real-time crash risk prediction model incorporating the various crash mechanisms across different traffic states," *Traffic Injury Prevention*, vol. 16, no. 1, pp. 28–35, 2015.

[16] C. Xu and Z. W. Qu, "Empirical analysis on time to collision at urban expressway," *Applied Mechanics & Materials*, vol. 505-506, pp. 1127–1132, 2014.

[17] L. Wang, M. Abdel-Aty, Q. Shi, and J. Park, "Real-time crash prediction for expressway weaving segments," *Transportation Research Part C: Emerging Technologies*, vol. 61, pp. 1–10, 2015.

[18] B. Sun, D. Dong, and S. Liu, "Bayesian belief net model-based traffic safety analysis on the freeway environment," in *Proceedings of the 5th International Conference on Transportation Engineering (ICTE '15)*, pp. 2754–2760, Dalian, China, September 2015.

[19] Q. Shi, M. Abdel-Aty, and R. Yu, "Multi-level bayesian safety analysis with unprocessed automatic vehicle identification data for an urban expressway," *Accident Analysis and Prevention*, vol. 88, pp. 68–76, 2016.

[20] M. G. Washington, M. G. Karlaftis, and F. L. Mannering, *Statistical and Econometric Methods for Transportation Data Analysis*, Chapman & Hall, London, UK, 2003.

[21] P. Kitsantas, M. Hollander, and L. Li, "Using classification trees to assess low birth weight outcomes," *Artificial Intelligence in Medicine*, vol. 38, no. 3, pp. 275–289, 2006.

[22] J. De Oña, R. De Oña, and F. J. Calvo, "A classification tree approach to identify key factors of transit service quality," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11164–11171, 2012.

[23] L.-Y. Chang and J.-T. Chien, "Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model," *Safety Science*, vol. 51, no. 1, pp. 17–22, 2013.

[24] Y.-H. Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley, Reading, Mass, USA, 1989.

[25] H. Xiao, H. Sun, and B. Ran, "Special factor adjustment model using fuzzy-neural network in traffic prediction," *Transportation Research Record*, no. 1879, pp. 17–23, 2004.

[26] J. P. Sangole and G. R. Patil, "Adaptive neuro-fuzzy interface system for gap acceptance behavior of right-turning vehicles at

partially controlled T-intersections," *Journal of Modern Transportation*, vol. 22, no. 4, pp. 235–243, 2014.

[27] J. S. R. Jang, C. T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Prentice Hall, Englewood Cliffs, NJ, USA, 1997.

[28] C. Xu, A. P. Tarko, W. Wang, P. Liu, L. R. Bai, and M. Chang, "Predicting freeway crash likelihood and severity with real-time loop detector data," in *Proceedings of the 92th Annual Meeting of Transportation Research Board*, Washington, DC, USA, 2013.

[29] R. Yu and M. Abdel-Aty, "Utilizing support vector machine in real-time crash risk evaluation," *Accident Analysis and Prevention*, vol. 51, pp. 252–259, 2013.