A Genetic Programming Model for Real-Time Crash Prediction on Freeways

Chengcheng Xu, Wei Wang, and Pan Liu

***Abstract*—This paper aimed at evaluating the application of the genetic programming (GP) model for real-time crash prediction on freeways. Traffic, weather, and crash data used in this paper were obtained from the I-880N freeway in California, United States. The random forest (RF) technique was conducted to select the variables that affect crash risk under uncongested and congested traffic conditions. The GP model was developed for each traffic state based on the candidate variables that were selected by the RF technique. The traffic flow characteristics that contribute to crash risk were found to be quite different between congested and uncongested traffic conditions. This paper applied the receiver operating characteristic (ROC) curve to evaluate the prediction performance of the developed GP model for each traffic state. The validation results showed that the prediction performance of the GP models were satisfactory. The binary logit model was also developed for each traffic state using the same training data set. The authors compared the ROC curve of the GP model and the binary logit model for each traffic state. The GP model produced better prediction performance than did the binary logit model for each traffic state. The GP model was found to increase the crash prediction accuracy under uncongested traffic conditions by an average of 8.2% and to increase the crash prediction accuracy under congested traffic conditions by an average of 4.9%.**

***Index Terms*—Binary logit model, freeway, genetic programming (GP), real-time crash prediction, traffic safety.**

I. Introduction

The development of real-time crash prediction models for freeways has recently received much attention from transportation professionals. One of the important practical applications of real-time crash prediction models is to identify hazardous traffic conditions that lead to crash occurrences in advanced traffic management systems (ATMSs) on freeways. In real-time crash prediction models, the likelihood of crash occurrences was related to freeway geometric characteristics and various real-time traffic flow variables such as vehicle speed, traffic occupancy, and the coefficient of the variation of vehicle speed. Real-time crash prediction models can be used to develop proactive traffic management strategies in ATMSs to improve traffic safety on freeways [1]–[3].

Previous studies that documented the development and application of real-time crash prediction models have usually focused on the statistical regression techniques, including the conditional logit model [4]–[11], the log-linear model [12], [13], the nonparametric Bayesian model [14], logistic regression [15]–[19], discriminate analysis [20], and the multivariate probit model [21]. Among these models, the conditional logit model and their variations are probably the most commonly used modeling techniques. In these studies, real-time crash prediction models were developed based on case-controlled data, in which traffic data before crash occurrence were used as cases, whereas the matched traffic data under crash-free conditions were used as controls. For example, Abdel-Aty *et al.* applied the conditional logit model to develop a real-time crash prediction model based on the matched-case-controlled data, in which each crash case was matched with a number of noncrash cases [4]. The results demonstrated that the likelihood of crash occurrence was affected by the average occupancy at the upstream station and by speed variance at the downstream station. Zheng *et al.* used the conditional logit model to evaluate the impacts of the standard deviation of speed that results from the oscillating traffic conditions on the likelihood of crash occurrence based on the case-controlled data [6].

Traditional statistical regression models usually require assumptions about the distribution of data and a well-defined functional form such as a linear functional form between dependent variable and independent variables. When the basic assumptions of the traditional statistical regression models were violated, inefficient estimations and incorrect inferences would be produced [22]–[24]. In response to the limitations associated with the statistical regression models, a few researchers have proposed nonparametric methods and artificial intelligence models for developing real-time freeway crash prediction models. These models include probabilistic neural networks [25], [26], Bayesian networks [27], [28], artificial neural networks [29]–[31], the classification and regression tree model [32], and the support vector machine (SVM) model [33]. The major limitation associated with the aforementioned artificial intelligence models is that these models work as black boxes, which cannot directly be used to identify the relationships between crash likelihood and various traffic flow variables. Thus, most times, these models are difficult for practical implementation.

The genetic programming (GP) model is a relatively new modeling technique that was proposed to solve classification and regression problems [34]. The GP model is rooted in the

Manuscript received March 15, 2012; revised July 28, 2012 and October 3, 2012; accepted October 19, 2012. Date of publication November 20, 2012; date of current version May 29, 2013. This work was supported in part by the National Key Basic Research Program of China under Grant 2012CB725400, the National High-Tech R&D Program of China under Grant 2012AA112304, the Ministry of Education of China through the Scholarship Award for Excellent Doctoral Students, and the Graduate Scientific Innovation Research Foundation of Jiangsu, China, under Grant CXZZ0164. The Associate Editor for this paper was S. C. Wong.

The authors are with the Key Laboratory of Traffic Planning and Management, School of Transportation, Southeast University, Nanjing 210096, China (e-mail: iamxcc1@gmail.com; wangwei@seu.edu.cn; liupan@seu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2012.2226240

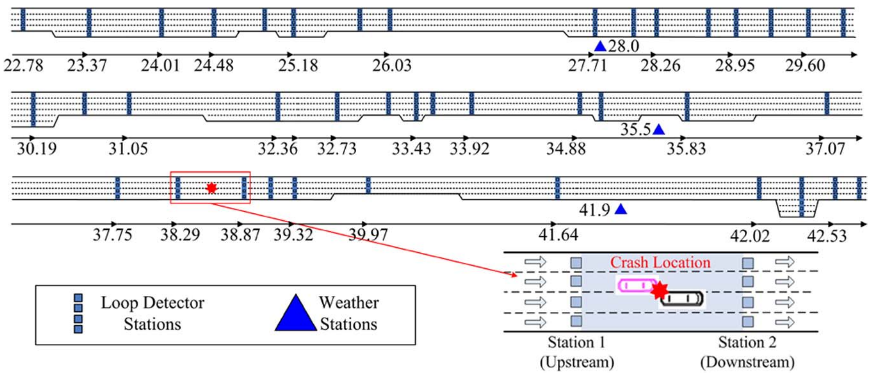


Fig. 1. Location of the loop detector and weather stations along the selected I-880N freeway section.

evolutionary theory and has recently been applied to classifications [35] and regression analyses [36] in transportation engineering. Compared with the traditional statistical regression models and artificial intelligence models, the GP model has two major advantages. First, the GP model can find a solution to a problem without any prespecified functional forms. The solutions of the GP model can be any functional forms describable by mathematics. In addition, the GP model could select the best functional form for the solution to the problem based on the training data. Second, in contrast to the artificial intelligence models, the GP models could remove the "black box" effect and make the model understandable. The output of a GP model is a readable mathematical model, which defines tangible relationships between dependent and independent variables. This allows the results to easily be applied in practical engineering applications. In addition, previous studies demonstrated that the GP model has better prediction performance over traditional modeling techniques [37]–[39]. So far, no applications of the GP model for real-time freeway crash prediction have been identified.

The primary objective of this paper is to investigate the applications of the GP model for real-time crash prediction on freeways. The random forest (RF) modeling technique was applied to select the contributing factors to crash risk under congested and uncongested traffic states. The RF model is a machine-learning method that consists of an ensemble of randomized classification and regression trees. It is one of the most efficient methods in evaluating variable importance. Based on the candidate variables selected by RF, the GP models were developed to identify the hazardous conditions that lead to crash occurrence under each traffic state. The prediction performance of the GP models would be compared with that of the binary logit models developed using the same training data set. In the rest of this paper, a brief description of the data used in this paper is presented, followed by the theoretical background of the RF and GP model. Then, the development and evaluation of the GP model for each traffic state are discussed. Finally, the prediction performance of the GP models is compared with to the binary logit model.

II. Data Sources

To accomplish the research objective, traffic, weather, geometry, and crash data were obtained from a 21-mi freeway section on the I-880N freeway in the United States. As shown in Fig. 1, a total of 40 loop detectors stations and 3 weather stations are located along the selected freeway section. The average spacing between detector stations was about 0.5 mi, and the average spacing between weather stations was about 7 mi. All the three weather stations were located within 1 mi from the I-880N freeway. Crash, traffic, and weather data were collected from January 1, 2008 to December 31, 2008 and from January 1, 2010 to December 31, 2010. A total of 807 crashes were identified and used for further data analysis.

The crash data reported at the selected freeway segment during the 24-month study period were obtained from the Statewide Integrated Traffic Records System (SWITRS) maintained by the California Department of Transportation (Caltrans). The actual location for each crash was identified by checking the variations in speed and occupancy near the reported location of crash. If there was no easily identifiable abrupt change in speed or occupancy, the reported crash location were used for further data analysis.

The traffic data were obtained from the Highway Performance Measurement System (PeMS) maintained by Caltrans. The PeMS database provided 30-s raw loop detector data, including vehicle count, vehicle speed, and traffic occupancy. The loop detectors sometimes suffer from hardware problems and random errors, which might result in invalid traffic data. Traffic data were excluded as invalid or not usable under one or more of the following conditions: 1) the average speed was greater than 100 mi/h; 2) the average occupancy was greater than 100%; 3) the flow rate was greater than 0 vph, whereas the occupancy was equal to 0%; 4) the average speed was greater than 0 mi/h, whereas the flow rate was equal to 0 vph; and 5) the occupancy was greater than 0%, whereas the flow rate was equal to 0 vph.

Traffic data were collected from the nearest upstream and downstream stations to each crash location, as shown in Fig. 1. The research team extracted 30-s raw traffic data in the time interval between 10 and 15 min prior to crash occurrence. The purpose of doing so was to identify hazardous traffic conditions ahead of the crash occurrence time to make preemptive measures possible [18], [31]. For example, if a crash occurred at 9:00 A.M., traffic data were extracted from 8:45–8:50 A.M. A similar time lag was also adopted in previous studies to develop real-time crash prediction models [26], [40]. The 30-s raw traffic data that were collected from upstream and downstream

TABLE I
CANDIDATE VARIABLES FOR THE GP MODEL

Symbol	Variables
AC_u	Average 30-second vehicle count at the upstream station (veh/30 s)
AO_u	Average 30-second detector occupancy at the upstream station (%)
AS_u	Average speed at the upstream station (mile/h)
SC_u	Std. dev. of 30-second vehicle counts at the upstream station (veh/30 s)
SO_u	Std. dev. of 30-second detector occupancies at the upstream station (%)
SS_u	Std. dev. of 30-second mean speeds at the upstream station (mile/h)
DC_u	Average absolute difference in 30-second vehicle counts between adjacent lanes at the upstream station (veh/30 s)
DO_u	Average absolute difference in detector occupancies between adjacent lanes at the upstream station (%)
DS_u	Average absolute difference in 30-second mean speeds between adjacent lanes at the upstream station (mile/h)
AC_d	Average 30-second vehicle counts at the downstream station (veh/30s)
AO_d	Average 30-second occupancy at the downstream station (%)
AS_d	Average speed at the downstream station (mile/h)
SC_d	Std. dev. of 30-second vehicle counts at the downstream station (veh/30s)
SO_d	Std. dev. of 30-second detector occupancies at the downstream station (%)
SS_d	Std. dev. of 30-second mean speeds at the downstream station (mile/h)
DC_d	Average absolute difference in flow between adjacent lanes at the downstream station (veh/30s)
DO_d	Average absolute difference in occupancy between adjacent lanes at the downstream station (%)
DS_d	Average absolute difference in speed between adjacent lanes at the downstream station (mile/h)
$DC_{u\text{-}d}$	Absolute difference in average vehicle counts between upstream and downstream stations (veh/30s)
$DO_{u\text{-}d}$	Absolute difference in average detector occupancies between upstream and downstream stations (%)
$DS_{u\text{-}d}$	Absolute difference in average speeds at upstream and downstream stations (mile/h)
WC	1 = adverse weather conditions(rain or fog); 0 = otherwise
$DT_{u\text{-}d}$	Distance between upstream and downstream stations (mile)
$Lane$	Number of lanes at the upstream stations
Wid_i	Inner Shoulder Width (ft)
Wid_o	Outer Shoulder Width (ft)
DAY	1 = Daylight; 0 = otherwise
$PEAK$	1= Peak period; 0 = otherwise

stations were further aggregated to a 5-min station level and converted into the 21 traffic flow variables, as shown in Table I. Weather conditions for each crash were extracted based on the time of the crash from the weather station that is nearest its location. Considering the sample size in each category, rain and fog were combined as adverse weather conditions. As a result, this paper considered the following two different weather conditions: 1) clear weather and 2) adverse weather.

Traffic and weather data for noncrash cases were randomly selected from crash-free days. For each selected crash case in the data set, the authors randomly selected m observations of noncrash cases. Different $m:1$ ratios between noncrash and crash cases were applied to develop the following GP models. The number of m was set from 1 to 10. The prediction accuracy was found to reach a maximum when m was set to be 10.

TABLE II
SAMPLE SIZE FOR EACH TRAFFIC STATE

Sample	Traffic State	Crash cases	Non-crash cases	Total
Training dataset	Uncongested	309	4510	4819
	Congested	175	335	510
Validation Sample	Uncongested	205	2993	3198
	Congested	118	232	350
Total		807	8070	8877

A total of 807 crash cases and 8070 noncrash cases were included in the data set. Because the traffic flow characteristics that contribute to the crash likelihood would be different between congested traffic and uncongested traffic [5], [8], the authors separately developed the GP models for congested and uncongested traffic states. The critical occupancy is often used to classify traffic flow conditions into congested and uncongested states. Based on the visual inspection of a flow-occupancy diagram that was developed using one-month traffic data, the critical occupancy was found to be 15%. Therefore, the traffic data for crash and noncrash cases were classified into congested and uncongested traffic states based on the mean value of traffic occupancy at upstream and downstream stations. If the average occupancy was larger than 15%, the traffic data were identified as a congested traffic state; otherwise, they were identified as an uncongested traffic state. Based on this rule, the original data set was separated into two subsamples: one subsample was for the congested traffic state, and the other subsample was for the uncongested traffic state. The distributions of crash and noncrash cases under different traffic states were summarized in Table II. The subsample for each traffic state was further randomly separated into a training data set and a validation data set with a ratio of 3 : 2. The training data sets were used to develop GP models under different traffic states, and the validation data sets were used to test the prediction performance of the developed GP models.

III. RESEARCH METHODOLOGY

A. RF for Variable Selection

RF is a machine-learning method that consists of an ensemble of randomized classification and regression trees [41]. In the RF model, a predetermined number of classification and regression trees are randomly generated and finally aggregated to give one single prediction. When solving classification problems, the RF model chooses the classification with the most votes from all the trees in the forest. In the training procedure of an RF model, each classification and regression tree is developed based on a bootstrap sample that is created by randomly selecting a number of samples with replacement from the original training data set. When building a classification and regression tree, the best split at each node is searched from a randomly selected subset of the whole predictors.

Currently, RF analysis is considered one of the most efficient methods in evaluating variable importance [42]. In the RF model, two measures, based on the Gini index and classification accuracy of out-of-bag (OOB) data, are usually used to evaluate the variable importance. In this paper, the measure based on

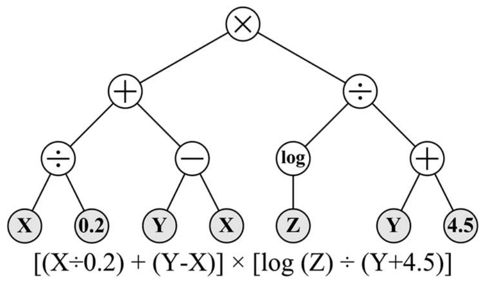


Fig. 2. Example of a function tree in GP.

the Gini index was used to evaluate the variable importance and to select the candidate variables for developing GP models. In this measure, the decrease in the Gini index at each node is calculated for the variable that is used to make the split. Then, the Gini-index-based variable importance measure for this variable is given as the average decrease in the Gini index over all trees in the forest.

B. GP

The GP model is an evolutionary computation method that was introduced by Koza [34]. The GP model can be used to generate mathematical models that represent approximate or exact solutions to a problem [34]. It can be considered an extension of the genetic algorithm (GA). The main difference between GP and GA is the representation of individuals. The individuals in a GA model are numbers that were coded as fixed-length binary strings, whereas the individuals in a GP model are mathematical models that were coded as function trees. Fig. 2 illustrates an example of a GP function tree. As shown in Fig. 2, the inner nodes represent mathematical functions such as "+" and "÷," and the leaf nodes represent predictors and constants. The mathematical model that is represented by the function tree in Fig. 2 is $f(X, Y, Z) = [(X/0.2) + (Y - X)] \times [\log(Z) + (Y/0.45)]$. In a particular problem, the set of functions and predictors should be prespecified. The mathematical models in GP are generated from a prespecified set of functions and predictors.

In general, GP works on a population of mathematical models (individuals) based on the evolution theory. In each generation, multiple models are stochastically selected based on their fitness and modified to form a new population of models by crossover, selection, and mutation operations. The new population of models is then used in the next iteration of the algorithm. A GP model will stop when the predetermined maximum number of generations has been produced or the predetermined fitness level has been reached for the population. Therefore, the evolution process is expected to continuously produce a better model for a problem that is intended to be solved.

1) Genetic Operation: The new models in a GP model are usually created by the following three genetic operators: 1) crossover; 2) mutation; and 3) reproduction. The reproduction operator simply selects a proportion of models and includes them into the next generation without any alterations. The crossover operator creates new or offspring models by combining information that was extracted from selected parents. Two parent models are randomly selected based on their fitness level, and subtrees are chosen from both parent models. Then, the crossover operator swaps the subtrees from the two parent models. Fig. 3 illustrates an example of the crossover operation.

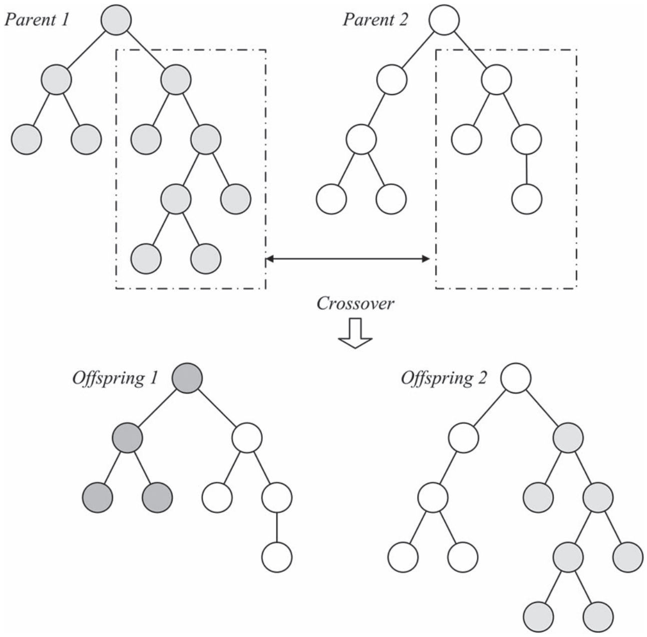


Fig. 3. Crossover operation in GP.

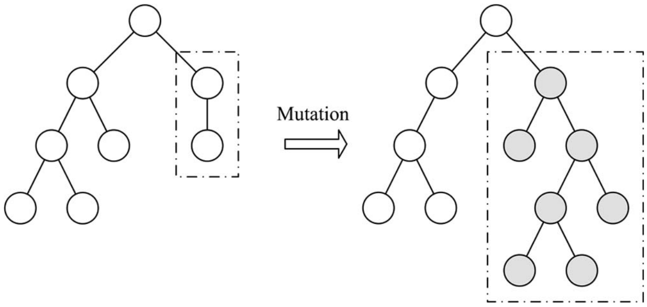


Fig. 4. Mutation operation in GP.

The purpose of the mutation operation is to introduce new information into the population and avoid the premature convergence of a GP model. In mutation, a single parent is randomly selected based on its fitness level. A random subtree on the parent model is selected and replaced with a new random tree created from the prespecified set of predictors and functions. This process is illustrated in Fig. 4. A new model is created by replacing the subtree on the left tree with a new randomly generated subtree.

2) Fitness Function: One of the most important components of a GP model is the fitness function, which determines how well a model in the population can solve the problem. The fitness function greatly varies across different types of problems. The fitness function is developed based on the error between the values predicted by the model and the actual data. For example, when a GP model is developed to set the time of a clock, the fitness function could be the summation of time that the clock is wrong.

The most commonly used fitness functions for classification problems include the number of hits, sensitivity or specificity, relative square error, and mean square error. In this paper, a fitness function for real-time freeway crash prediction was developed based on the number of hits and square errors. Assuming a data set $S = \{(y_1, x_1), (y_2, x_2), \ldots, (y_\mathrm{n}, x_\mathrm{n})\}$ of data (x_i) for crash $(y_\mathrm{i} = 1)$ and noncrash $(y_\mathrm{i} = 0)$, the functional form of the fitness function is expressed as follows:

$$F(B_j)=\sum_{i=1}^{n}\left(\beta^{y_i}\times|y_i-C(B_j(x_i))|+(y_i-B_j(x_i))^2\right) \quad (1)$$

where $F(B_j)$ denotes the fitness of the jth model B_j in the population, $B_j(x_\mathrm{i})$ is the value calculated by the jth model B_j in the population, and $C(B_j(x_\mathrm{i}))$ represents the function that converts the value calculated by the model B_j into 1 or 0. The functional form of this $C(B_j(x_\mathrm{i}))$ is expressed as follows:

$$C(B_j(x_i)) = \begin{cases} 1 & \text{if } B_j(x_i) > c \\ 0 & \text{if } B_j(x_i) < c \end{cases} \quad (2)$$

where c is the cut value. Note that the selection of cut value c in (2) does not change the prediction performance of the evolved models, because the GP model has the ability to compensate for any change in the cut value c in (2) by corresponding changes in the generated models. Previous studies also confirmed that the cut value in the fitness function does not change the distribution of types I and II errors of the evolved GP model [38], [39]. Therefore, the cut value in (2) was arbitrarily set to 0.5 in the following analysis.

Because the number of noncrash cases is much greater than that of crash cases in the training data set, the GP model might ignore the information from crash cases and classify all the observations as noncrash cases to improve the overall classification accuracy. To account for this problem, the weighting factor β was introduced in the fitness function. The weighting factor β was set to the ratio between the number of noncrash and crash cases in each training data set. As shown in (1), β^{y_i} is equal to 1 if $y_i = 0$ (noncrash), and β when if $y_i = 1$ (crash). Hence, in (1), correctly classifying a crash case will contribute more to the fitness than correctly classifying a noncrash case.

The fitness function based on the number of hits only is not sensitive to the marginal improvements in deviations from the target. Therefore, the square error was introduced in the fitness function to measure how closely the evolved outputs and the target outputs in the training data match up.

3) Procedure of GP: The GP model is based on a repetitive computational process. As shown in Fig. 5, the GP model uses the following steps to solve problems.

a) *Initialization.* Create at random an initial population of M models that represent potential solutions to the prediction of crash occurrence on freeways.
b) Execute each model in the current population on the training data set and evaluate the fitness of each model in the current population.
c) Select the parent models that will be used to produce offspring models.

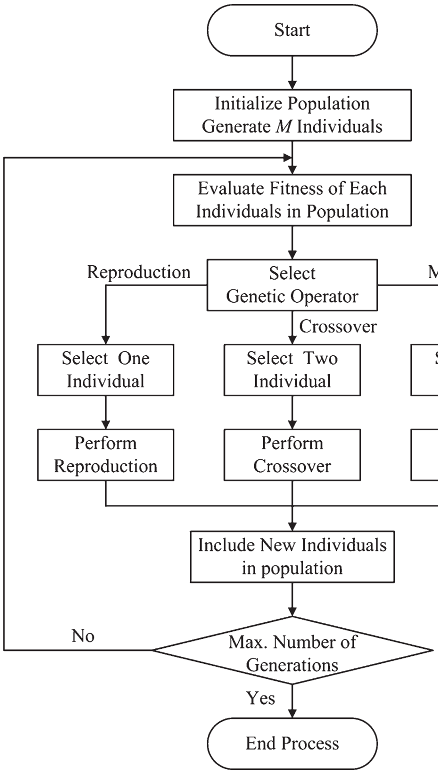


Fig. 5. Flowchart of GP.

d) Probabilistically select the reproduction, crossover, and mutation operators.
e) Generate a new model by performing one of the three genetic operators.
f) Repeat steps c–e until the predetermined population size M has been reached.
g) Replace the M old models by new generated M models.
h) Repeat steps b–g until the predetermined maximum generation N has been reached.
i) The model with the best fitness level in any generation is designated as the result of GPs.

C. Binary Logit Model

This paper aimed at developing a method of predicting crash occurrence on freeways based on traffic data that were collected from loop detector stations on freeways. This gave a binary outcome that can be coded as one if a crash occurs and zero if no crashes occur. Hence, the binary logit model was used to benchmark the GP model.

The binary logit model has been widely used for predicting a binary-dependent variable as a function of predictor variables in transportation engineering [43]–[45]. Using a binary logit model, the probability of crash occurrence can be estimated using the following equation [46]:

$$P(x_i) = \frac{1}{1+e^{-g(x_i)}} \quad (i = 1, 2, \ldots, n) \quad (3)$$

where $P(x_i)$ denotes the probability that the certain traffic flow conditions lead to crash occurrence, and $g(x)$ is the multiple

TABLE III
DISTRIBUTION OF CRASHES IN THIS PAPER

Factors	Categories	Percentages
Crash Type	Rear end	55.6%
	Sideswipe	23.9%
	Others	20.5%
Crash severity	Property damage only (PDO)	71.5%
	Injury Crash	28.5%
Peak hours	Pear hours	35.5%
	off-peak hours	64.5%
Spacing between upstream and downstream stations	Distance < 0.3 mile	5.6%
	0.3 mile < Distance < 0.6 mile	52.2%
	Distance > 0.6 mile	42.3%
Lighting conditions	Daylight	71.2%
	Dark (Street Light)	11.2%
	Dark (No Street Light)	13.2%
	Others	4.4%

linear combination of explanatory variables, which can be expressed as

$$g(x) = \ln \frac{P(x_i)}{1 - P(x_i)} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \quad (4)$$

where x_{ki} denotes the value of a variable k for sample i, and β_k is the coefficient of variable k. The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ can be estimated based on the log-likelihood function for (3), which is given by

$$\ln L(\beta, x_i) = \sum_{i=1}^{n} [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{ki} x_{ki} - \ln(1 + e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{ki} x_{ki}})] . \quad (5)$$

IV. DATA ANALYSIS AND RESULTS

A. *Preliminary Analysis*

A preliminary analysis of the crashes used in this paper was conducted in this section. The distributions of the crashes used in this paper were summarized in Table III. Most of the crashes (55.6%) in the data set were rear-end crashes, followed by sideswipe crashes of about 23.9%. With regard to the severity outcome, 71.5% of the crashes in the data set were property-damage-only crashes, and 28.5% were injury crashes. With regard to the crash occurrence time, 35.5% of the crashes occurred during peak hours, and 64.5% occurred during off-peak hours.

As aforementioned, a total of 40 loop detector stations are located along the selected 21-mi freeway section. The 40 loop detector stations divided the 21-mi freeway section into 39 freeway segments. Considering the crash location, 42.3% of the crashes occurred in a segment with a length larger than 0.6 mi. Among the 39 segments, only 7 segments are larger than 0.6 mi. This may imply that the geometric characteristic (segments length) was a significant variable that affects crash likelihood. Moreover, 71.2% of the crashes occurred during daylight, and 13.2% crashes occurred at night without street light.

B. *Variable Selection Using RF*

As shown in Table I, 28 candidate variables were obtained in this paper. The RF model was used to select the significant variables that affect crash risk under each traffic state. The "RF" package in the MATLAB software was used to develop the RF models [47]. When using the RF model, the number of trees in the forest and the number of variables tried at each node first need to be specified. To obtain stable estimations of variable importance, we conducted the RF model and calculated OOB error rates for different numbers of trees. It was found that 1000 trees were sufficient to obtain a constant minimum error rate for both congested and uncongested traffic states.

After the number of trees has been determined, the RF models were conducted with different numbers of variables tried at each node. The OOB error rate of the RF model for uncongested traffic state reached a minimum when the number of variables tried at each node was equal to 9. In addition, six variables tried at each node produced the minimum OOB error rate of the RF model for the congested traffic state.

The RF model for each traffic state was first conducted using all the 28 variables, as shown in Table I. The variable importance of the 28 candidate variables for congested and uncongested traffic states was illustrated in Fig. 6. As expected, the traffic flow variables that contribute to crash risk were quite different between uncongested and congested traffic states. The average speed, occupancy difference between adjacent lanes, speed difference between adjacent lanes, and speed variance were the main contributing factors to crash risk in uncongested traffic conditions, whereas the average occupancy, standard deviation of occupancy, average speed, and speed difference between upstream and downstream stations were the main contributing factors to crash risk in congested traffic conditions.

To select the number of important variables for developing GP models, the RF models were further conducted in a successive phase in which the number of input variables was set from 1 to 28. To be more specific, we first conducted the RF model using the top 1 variable in Fig. 6, then conducted the RF model using the top 2 variables in Fig. 6, and so on. The OOB error rate of the RF model for the uncongested traffic state reached a minimum when the top 12 important variables in Fig. 6(a) were used to develop the RF model. Therefore, the top 12 important variables in Fig. 6(a) were selected to develop the GP model for predicting the crash risk under the uncongested traffic state. In addition, the top 8 variables in Fig. 6(b) produced the minimum OOB error rate of the RF model for the congested traffic state. The GP model for the congested traffic state was developed based on these top 8 variables in Fig. 6(b).

C. *GP Model*

In this paper, two GP models were developed to separately predict the crash occurrence under congested and uncongested traffic states based on the traffic data collected from the loop detector stations on freeways. The research team developed the

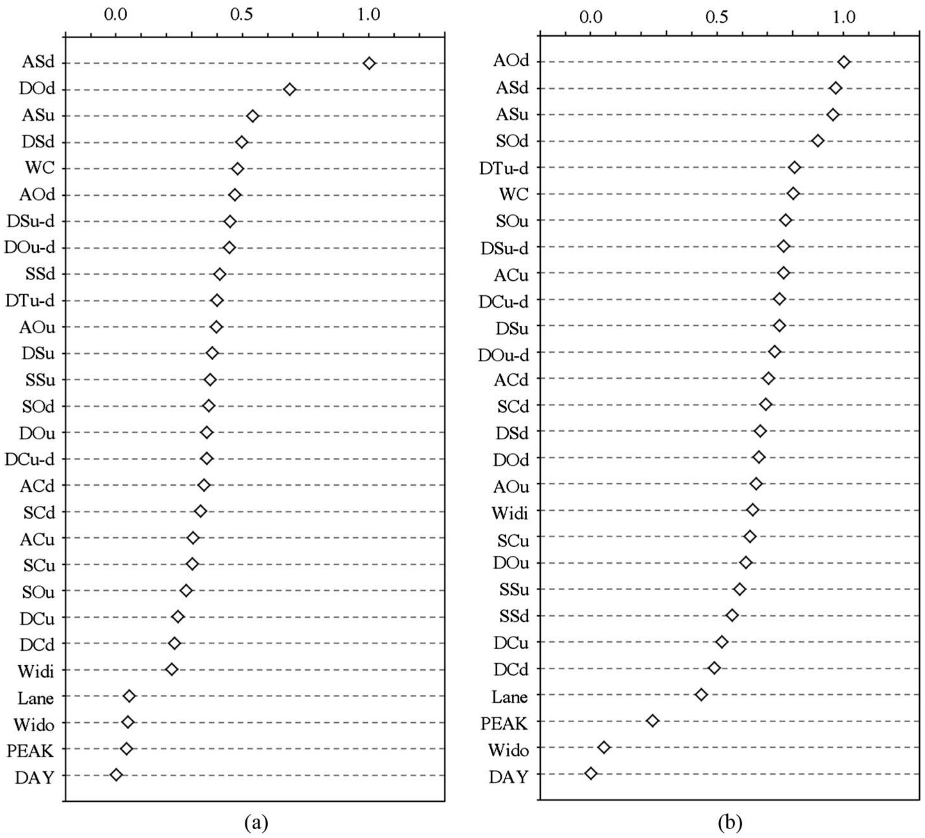


Fig. 6. Variable importance based on the normalized Gini index for each traffic state.

TABLE IV
SUMMARY OF THE CONFIGURATION PARAMETERS OF THE GP MODEL

Configuration Parameters	Selected Values
Number of individuals	1000
Number of generations	100
Depth limited to	30
Probability of crossover	automatic adaptation procedure [48]
Probability of mutation	automatic adaptation procedure [48]
Reproduction probability	0
Selection	Lexictour [49]
Initial population	Ramped half-and-half method
Initial maximum depth	6
Functions set	+, -, ×, ÷, protected square root, and protected natural logarithm
Terminal set	variables selected by RF, and random constant (between 0 and 10)

GP model using GPLab toolbox 3.0 [48], which is a popular GP software coded in MATLAB. The various parameters used in the GP models were given in Table IV. As shown in Table IV, the function set that was used to build the GP model contained six standard arithmetic operators, i.e., $+$, $-$, $\times$, $\div$, protected square root, and protected natural logarithm. The terminal set included the variables selected by RF and constant terminals (randomly generated floating-point numbers between 0 and 10). The population size was set to 1000, and the maximum number of generations was 100. With regard to the probability of operators, the reproduction probability was 0. The purpose of doing so was to let the crossover and mutation operation govern the evolutionary process. The probability of crossover and mutation was optimized during GP running by the automatic adaptation procedure [49]. In the automatic adaptation procedure, the probability values of operators will be increased if they have been producing models with better fitness. Otherwise, their probability values will be decreased.

The initial population of individuals was generated by a ramped half-and-half method [34], and the depth for trees in the initial population was limited to 6. The maximum depth for generating trees in other populations was 30. The selection process was used to choose the parent models for producing offspring models. In this paper, the lexictour method was chosen as the selection process. In the lexictour method, a random number of models are chosen from the population, and the model with the best fitness among the chosen models is selected. In addition, if two models have the same fitness, the model with fewer nodes will be chosen as the best [50].

1) GP Model Under the Uncongested Traffic State: Fig. 7 illustrates the crash prediction model under uncongested traffic conditions. Six traffic flow variables, weather conditions, and the spacing between upstream and downstream stations were found to be correlated with the crash risk in the GP model under uncongested traffic conditions. The GP model indicates a complex relationship of crash risk with traffic flow conditions, weather conditions, and geometry characteristics. Accordingly, Fig. 8 was developed to indicate the change in crash risk as the value of each continuous variable changes. Each continuous variable changed over a normal range of values when the other variables were kept at their sample mean.

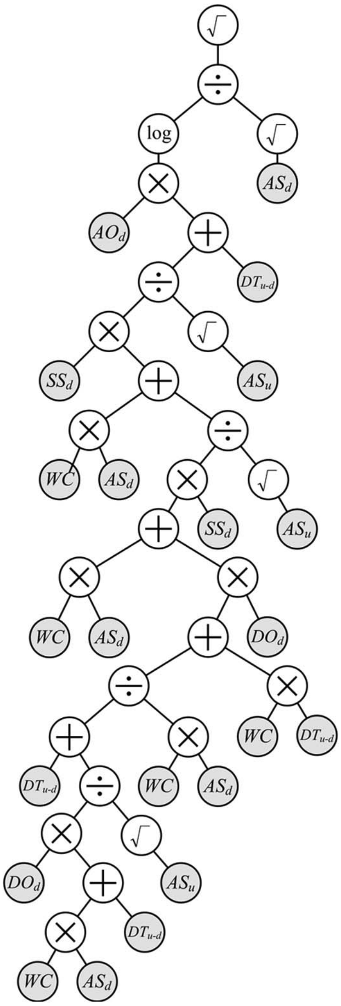


Fig. 7. GP model for the uncongested traffic state.

As shown in Figs. 8(a)–(d), the high occupancy and standard deviation of speed at the downstream station were found to be associated with the increase in crash risk. In addition, crash risk increases as the upstream and downstream speed decrease. The decreasing speed might represent the increase in traffic density and queue formations. These results are consistent with the findings of previous studies [4], [8], [10], [25]. As shown in Fig. 8(e), high occupancy difference between adjacent lanes at the downstream station results in an increase in crash risk. The large occupancy difference between adjacent lanes was found to be associated with high lane-change frequency [51]. Thus, high lane-change frequency increases the crash risk in uncongested traffic conditions.

The occupancy difference between upstream and downstream stations was found to be associated with the increase in crash risk [Fig. 8(f)]. The findings of a previous study also demonstrated that crashes were more likely to occur under traffic flow states with a large occupancy difference between upstream and downstream [8]. Moreover, as shown in Fig. 8(g), the spacing between upstream and downstream stations increases the crash risk between them, indicating the fact that crashes are more likely to occur in a freeway segment with a large length. This is consistent with the findings shown in the preliminary analysis.

2) GP Model Under the Congested Traffic State: The crash prediction model under congested traffic conditions is illustrated in Fig. 9. The occupancy at downstream stations, the standard deviation of occupancy at the upstream station, the speed difference between the upstream and downstream stations, weather conditions, and the spacing between the upstream and downstream stations were found to be correlated with the crash risk in the GP model under congested traffic conditions. Fig. 10 was developed to explore the relationship of crash risk with different predictors in the GP model.

As shown in Fig. 10(a), crash risk decreases with the increase in occupancy at the downstream station under congested traffic conditions. This is consistent with the findings in previous studies that drivers behave differently in spare and heavy traffic and crash risk may decrease with the increases in traffic density under congested traffic conditions [8], [52], [53]. As shown in Fig. 10(b) and (c), crash risk under congested traffic conditions tends to be high when the speed difference between the upstream and downstream stations and the standard deviation of occupancy at the upstream station are high. In addition, the spacing between the upstream and downstream stations was found to increase crash risk under congested traffic conditions.

D. Prediction Performance

The prediction performance of the GP model under each traffic state was tested. The authors developed receiver operating characteristic (ROC) curves to evaluate the prediction performance of the GP models with different thresholds. The ROC curve is a graphical plot of the sensitivity on the y-axis against (1—specificity) on the x-axis for different thresholds. In this paper, the sensitivity represents the proportion of crash cases predicted as a crash, and (1—specificity) represents one minus the proportion of noncrash cases predicted as a noncrash. To develop the ROC curve of the GP model for each traffic state, we calculated the sensitivity and (1—specificity) for multiple thresholds by using the validation data sets.

To test the relative prediction performance of the GP model for each traffic state, the research team benchmarked it against the binary logit model, which is a more prevalent methodology. For comparison, the authors developed the binary logit model for each traffic state using the same training data set. In addition, the validation data set for each traffic state was used to test the prediction performance of the binary logit models developed. The ROC curves for GP and the binary logit model under uncongested traffic conditions on the validation data set are illustrated in Fig. 11.

As shown in Fig. 11, the ROC curve for the GP model is always to the left of the curve for the binary logit model, indicating that the prediction accuracy of the GP model under uncongested traffic conditions is always greater than that of the binary logit model, no matter what threshold is selected. Fig. 12 presents the ROC curves for GP and the binary logit model under congested traffic conditions. The prediction accuracy of the GP model under congested traffic conditions was also found to be greater than that of the binary logit model.

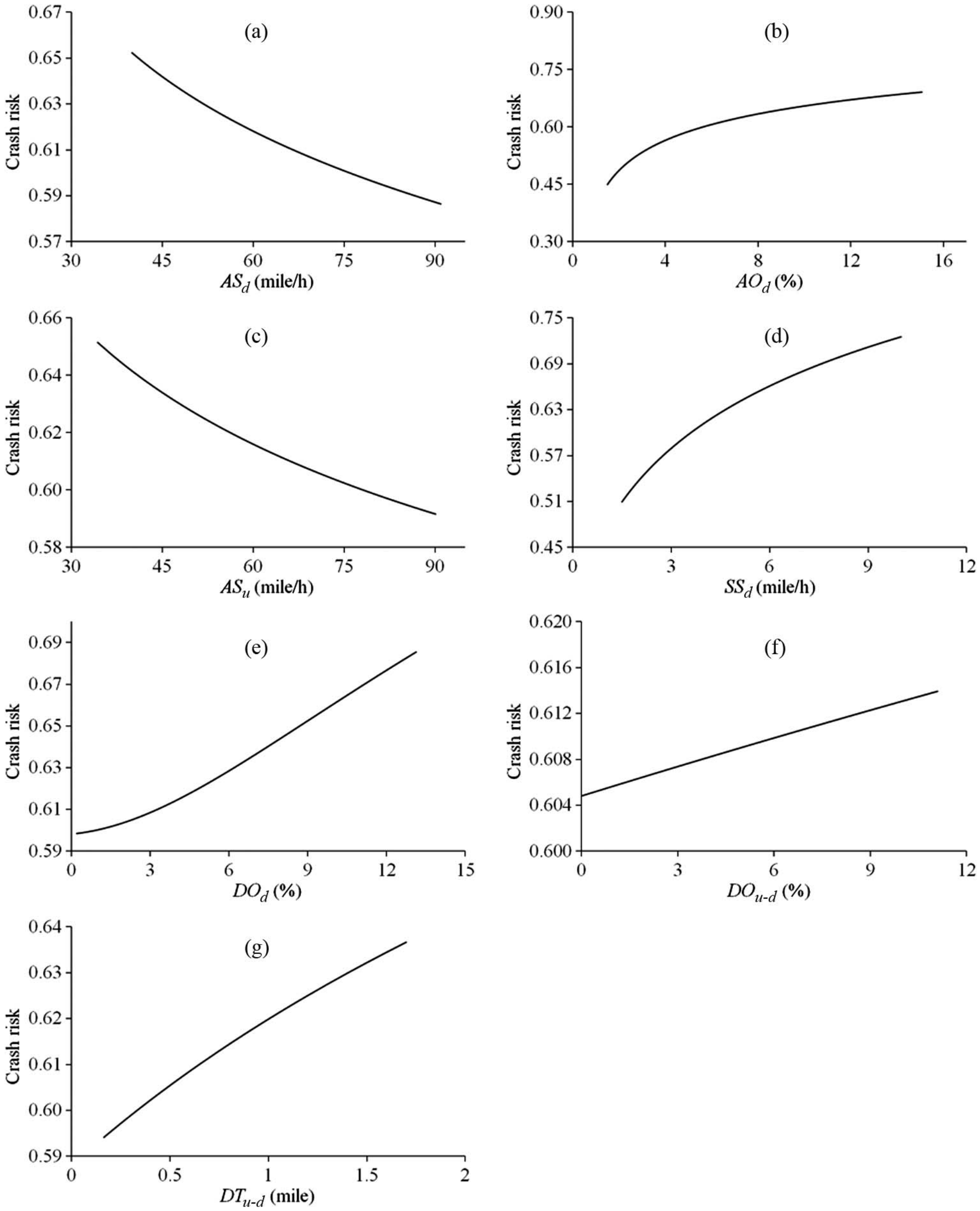


Fig. 8. Relationship between crash risk and different predictors in the GP model under uncongested traffic conditions.

Tables V and VI summarize the prediction performance of the GP model under uncongested and congested traffic conditions, respectively. The prediction performance is measured by the percentage of the correctly predicted crash cases for different false-alarm rates (1—specificity). As shown in Tables V and VI, there is a tradeoff between crash prediction accuracy and false-alarm rate. The prediction accuracy of crash cases increased as the false-alarm rate was increased. Thus, the threshold needs to carefully be determined to meet the requirement of the practical implementation or the preference of a specific traffic agency. After determining the threshold, the prediction performance of the GP model could easily be evaluated using the aforementioned ROC curves. For example, if a threshold value is selected to accept a 20% false-alarm rate for identifying crash occurrence under uncongested traffic conditions, the crash prediction accuracy of the GP model is found to be 61.0%.

For comparison, Tables V and VI also summarize the crash prediction accuracy of the binary logit model under each traffic state at the same false-alarm rate. As shown in Table V, the crash prediction accuracy of the GP model is much better than that of the binary logit model. The average difference in crash prediction accuracy between GP and the binary logit model is 8.2%, indicating that the GP model could increase the crash prediction accuracy under uncongested traffic conditions by an average of 8.2% compared with the binary logit model. For congested traffic flow conditions, the average difference in crash prediction accuracy between the two models was found to be 4.9%. Thus, the GP model could increase the crash prediction accuracy under congested traffic conditions by an average of 4.9%.

V. Conclusion

This paper investigated the application of the GP model for real-time crash prediction on freeways. Traffic, weather, geometry, and crash data were collected from the I-880N

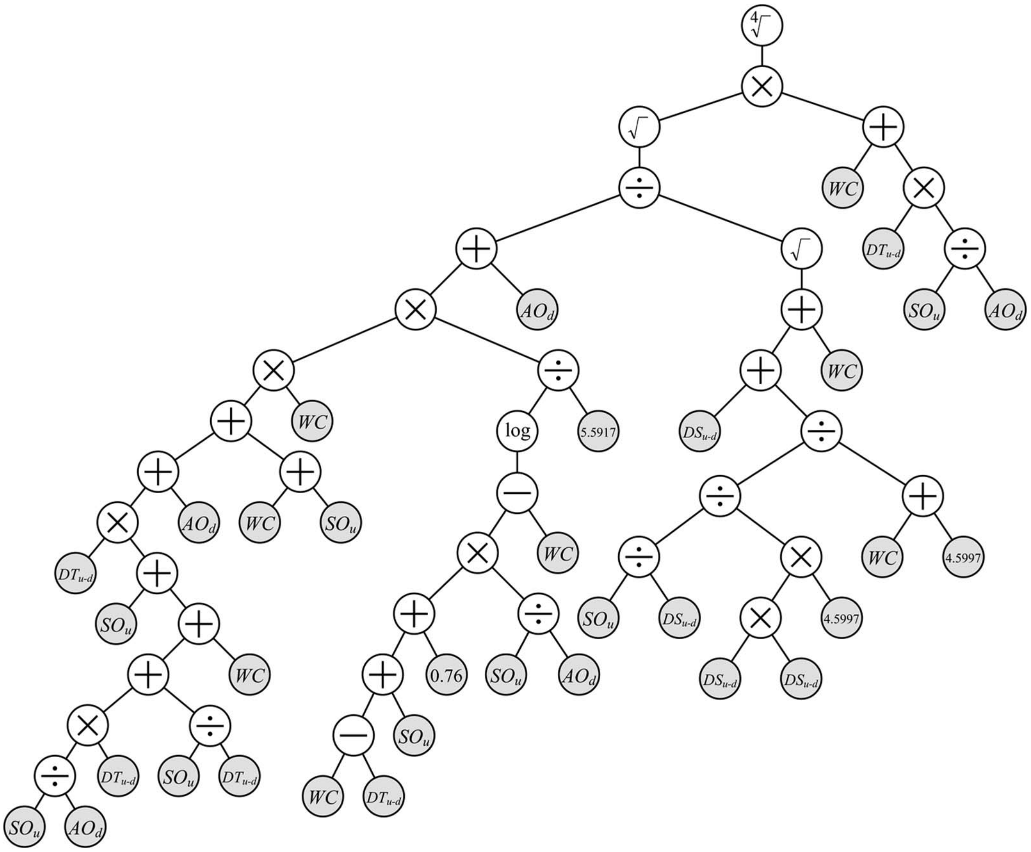


Fig. 9. GP model for the congested traffic state.

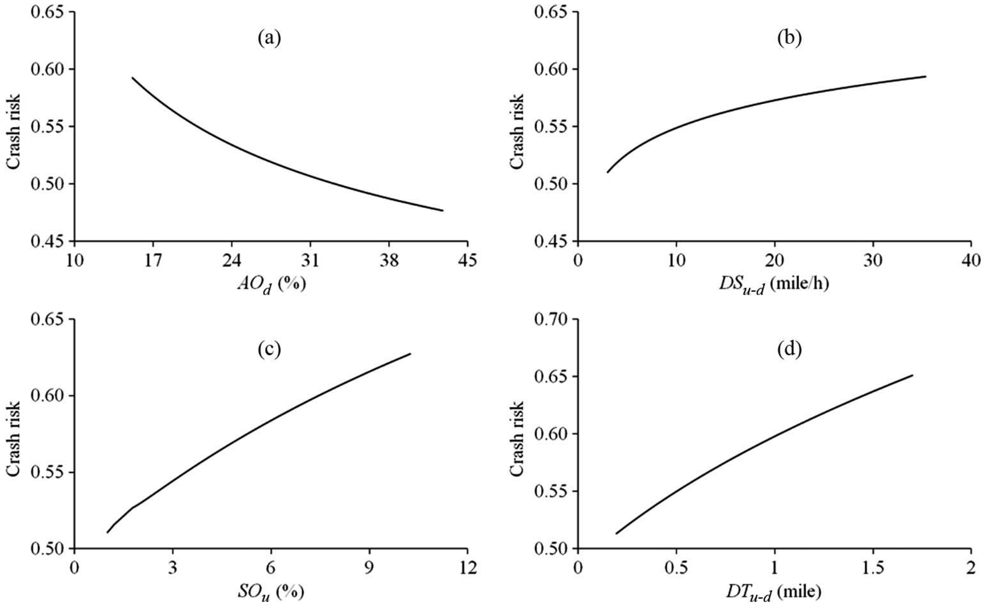


Fig. 10. Relationship between crash risk and different predictors in the GP model under congested traffic conditions.

freeway located in the United States, in 2008 and 2010. RF was applied to select the contributing factors to crash risk under uncongested and congested traffic conditions. Based on the candidate variables selected by RF, the GP model was conducted to develop the crash prediction model for each traffic state. The traffic flow characteristics that affect crash risk were found to be quite different between congested and uncongested traffic conditions. In general, the traffic density, speed variance, lane-change frequency, and occupancy difference between upstream and downstream were the main contributing factors to crash risk under the uncongested traffic conditions. Under the congested traffic conditions, crash risk decreases with the increase in traffic density. The speed difference between upstream and downstream stations and the standard deviation of occupancy also affected crash risk under the congested traffic conditions.

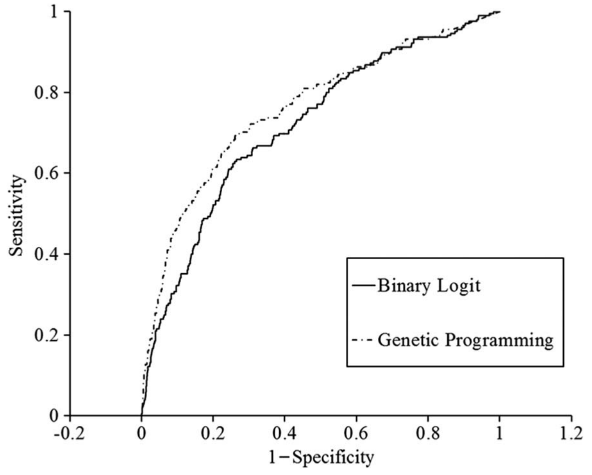


Fig. 11. ROC curves of GP and the binary logit model for the uncongested traffic state.

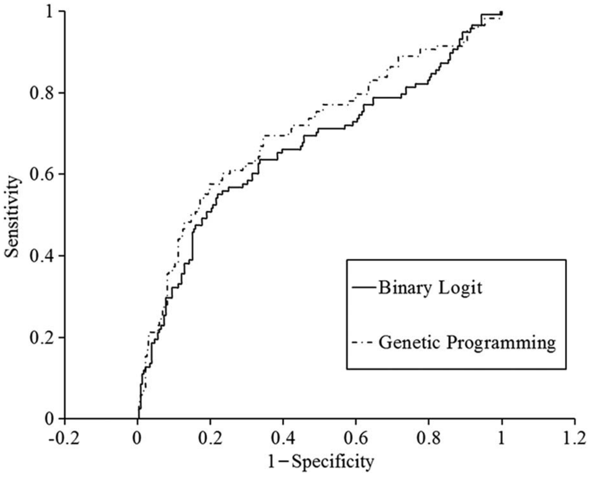


Fig. 12. ROC curves of GP and the binary logit model for the congested traffic state.

TABLE V
PREDICTION PERFORMANCE OF THE GP MODEL FOR THE UNCONGESTED TRAFFIC STATE

1−Specificity	Sensitivity of GP	Sensitivity of Logit	Difference
0.1	46.3%	32.7%	13.6%
0.2	61.0%	52.2%	8.8%
0.3	71.7%	64.4%	7.3%
0.4	76.1%	69.8%	6.3%
0.5	82.0%	77.1%	4.9%

The ROC curves were developed to evaluate the prediction performance of the GP model under each traffic state. The validation results demonstrated that the prediction performance of the GP models was deemed satisfactory. For comparison, the binary logit model was also developed for each traffic state using the same training data set. In general, the prediction performance of the GP models was better than that of the binary logit models, no matter what threshold is selected. Compared with the binary logit model, the GP model could increase the crash prediction accuracy under uncongested traffic conditions by an average of 8.2% and increase the crash prediction accuracy under congested traffic conditions by an average of 4.9%.

TABLE VI
PREDICTION PERFORMANCE OF THE GP MODEL FOR THE CONGESTED TRAFFIC STATE

1−Specificity	Sensitivity of GP	Sensitivity of Logit	Difference
0.1	37.3%	32.2%	5.1%
0.2	57.6%	50.8%	6.8%
0.3	62.7%	57.6%	5.1%
0.4	69.5%	66.1%	3.4%
0.5	75.4%	71.2%	4.2%

The real-time crash prediction model developed by the GP algorithm can be used in ATMSs such as variable-speed-limit systems [54] and ramp metering [55], [56] to improve traffic safety on freeways. For example, when a freeway segment is found to be susceptible to crash occurrence, the variable speed limit system could be activated to reduce crash risk by predetermined management strategies. However, before the findings in this paper are used in practical engineering applications, additional research is still needed to test the transferability of the GP model using data that were collected from other freeways. In addition, this paper partly demonstrated the benefits of using the GP model in real-time crash prediction over the traditional logit model. However, the authors did not compare the GP model to other artificial intelligence models such as SVM and artificial neural network models. The authors recommend that future studies may focus on these issues.

ACKNOWLEDGMENT

The authors would like to thank the Highway Performance Measurement System (PeMS) and California Department of Transportation (Caltrans) for providing the crash and traffic data that were used in this paper and X. Jiang for her help in the data collection.

REFERENCES

[1] C. Lee, B. Hellinga, and K. Ozbay, “Quantifying effects of ramp metering on freeway safety,” *Accid. Anal. Prev.*, vol. 38, no. 2, pp. 279–288, Mar. 2006.

[2] M. Abdel-Aty, J. Dilmore, and A. Dhindsa, “Evaluation of variable speed limits for real-time freeway safety improvement,” *Accid. Anal. Prev.*, vol. 38, no. 2, pp. 335–345, Mar. 2006.

[3] M. Abdel-Aty, J. Dilmore, and V. Gayah, “Considering various ALINEA ramp metering strategies for crash risk mitigation on freeways under congested regime,” *Transp. Res. C: Emerging Technol.*, vol. 15, no. 2, pp. 113–134, Apr. 2007.

[4] M. Abdel-Aty, N. Uddin, F. Abdalla, A. Pande, and L. Hsia, “Predicting freeway crashes from loop detector data using matched-case-control logistic regression,” *Transp. Res. Rec.*, vol. 1897, no. 1, pp. 88–95, 2004.

[5] M. Abdel-Aty, N. Uddin, and A. Pande, “Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways,” *Transp. Res. Rec.*, vol. 1908, no. 1, pp. 51–58, 2005.

[6] Z. Zheng, S. Ahna, and C. Monsere, “Impact of traffic oscillations on freeway crash occurrences,” *Accid. Anal. Prev.*, vol. 42, no. 2, pp. 626–636, Mar. 2010.

[7] M. Abdel-Aty and P. Rajashekar, “Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data,” *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 167–174, Jun. 2006.

[8] C. Xu, P. Liu, W. Wang, and Z. Li, “Evaluation of the impacts of traffic states on crash risks on freeways,” *Accid. Anal. Prev.*, vol. 47, pp. 162–171, Jul. 2012.

[9] M. Abdel-Aty, H. Hassan, M. Ahmed, and A. Al-Ghamdi, “Real-time prediction of visibility-related crashes,” *Transp. Res. C*, vol. 24, pp. 288–298, Oct. 2012.

[10] M. Ahmed and M. Abdel-Aty, “The viability of using automatic vehicle identification data for real-time crash prediction,” *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 459–468, Jun. 2012.

[11] M. Ahmed, M. Abdel-Aty, and R. Yu, “A Bayesian updating approach for real-time safety evaluation using AVI data,” presented at the Annu. Meeting Transport. Res. Board, Washington, DC, 2012.

[12] C. Lee, F. Saccomanno, and B. Hellinga, “Analysis of crash precursors on instrumented freeways,” *Transp. Res. Rec.*, vol. 1784, pp. 1–8, 2002.

[13] C. Lee, B. Hellinga, and F. Saccomanno, “Real-time crash prediction model for application to crash prevention in freeway traffic,” *Transp. Res. Rec.*, vol. 1840, pp. 67–77, 2003.

[14] J. Oh, C. Oh, S. Ritchie, and M. Chang, “Real-time estimation of accident likelihood for safety enhancement,” *J. Transp. Eng.*, vol. 131, no. 5, pp. 358–363, May 2005.

[15] M. Ahmed, M. Abdel-Aty, and R. Yu, “Assessment of the interaction between crash occurrence, mountainous freeway geometry, real-time weather and AVI traffic data,” presented at the Annu. Meeting Transport. Res. Board, Washington, DC, 2012.

[16] C. Lee, M. Abdel-Aty, and L. Hsia, “Potential real-time indicators of sideswipe crashes on freeways,” *Transp. Res. Rec.*, vol. 1953, no. 1, pp. 41–49, 2006.

[17] T. Golob, W. Recker, and Y. Pavlis, “Probabilistic models of freeway safety performance using traffic flow data as predictors,” *Safety Sci.*, vol. 46, no. 9, pp. 1306–1333, Nov. 2008.

[18] C. Lee, P. Park, and M. Abdel-Aty, “Lane-by-lane analysis of crash occurrence based on driver’s lane-changing and car-following behavior,” *J. Transp. Safety Security*, vol. 3, no. 2, pp. 108–122, 2011.

[19] C. Xu, P. Liu, W. Wang, and C. Yu, “Exploration and identification of hazardous traffic flow states before crash occurrences on freeways,” presented at the Annu. Meeting Transport. Res. Board, Washington, DC, 2011.

[20] C. Xu, P. Liu, W. Wang, and Z. Li, “Development of a crash risk index to identify real-time crash risks on freeways,” presented at the Annu. Meeting Transport. Res. Board, Washington, DC, 2012.

[21] Z. Christoforou, S. Cohen, and M. Karlaftis, “Identifying crash type propensity using real-time traffic data on freeways,” *J. Safety Res.*, vol. 42, no. 1, pp. 43–50, Feb. 2011.

[22] L. Mussone, A. Ferrari, and M. Oneta, “An analysis of urban collisions using an artificial intelligence model,” *Accid. Anal. Prev.*, vol. 31, no. 6, pp. 705–718, Nov. 1999.

[23] D. Delen, R. Sharda, and M. Bessonov, “Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks,” *Accid. Anal. Prev.*, vol. 38, no. 3, pp. 434–444, May 2006.

[24] P. Miaou and D. Lord, “Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes,” *Transp. Res. Rec.*, vol. 1840, pp. 31–40, 2003.

[25] C. Oh, J. Oh, and S. Ritchie, “Real-time hazardous traffic condition warning system: Framework and evaluation,” *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 3, pp. 265–272, Sep. 2005.

[26] M. Abdel-Aty and A. Pande, “Identifying crash propensity using specific traffic speed conditions,” *J. Safety Res.*, vol. 36, no. 1, pp. 97–108, Jan. 2005.

[27] M. Hossain and Y. Muromachi, “Development of a real-time crash prediction model for urban expressway,” *J. Eastern Asia Soc. Transp. Stud.*, vol. 8, pp. 2092–2107, 2010.

[28] M. Hossain and Y. Muromachi, “Evaluating location of placement and spacing of detectors for real-time crash prediction on urban expressways,” presented at the Annu. Meeting Transport. Res. Board, Washington, DC, 2010.

[29] A. Pande and M. Abdel-Aty, “Assessment of freeway traffic parameters leading to lane-change-related collisions,” *Accid. Anal. Prev.*, vol. 38, no. 5, pp. 936–948, 2006.

[30] A. Pande and M. Abdel-Aty, “Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways,” *Transp. Res. Rec.*, vol. 1953, pp. 31–40, 2007.

[31] A. Pande, A. Das, M. Abdel-Aty, and H. Hassan, “Real-time crash risk estimation: Are all freeways created equal?” *Transp. Res. Rec.*, vol. 2237, pp. 60–66, 2011.

[32] M. Hossain and Y. Muromachi, “Understanding crash mechanism and selecting appropriate interventions for real-time hazard mitigation on urban expressways,” *Transp. Res. Rec.*, vol. 2213, pp. 53–62, 2011.

[33] X. Qu, W. Wang, P. Liu, and D. Noyce, “Real-time prediction of freeway rear-end crash potential by support vector machine,” presented at the Annu. Meeting Transport. Res. Board, Washington, DC, 2012.

[34] R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press, 1992.

[35] A. Das and M. Abdel-Aty, “A genetic programming approach to explore the crash severity on multilane roads,” *Accid. Anal. Prev.*, vol. 42, no. 2, pp. 548–557, Mar. 2010.

[36] A. Das, M. Abdel-Aty, and A. Pande, “Genetic programming to investigate design parameters contributing to crash occurrence on urban arterials,” *Transp. Res. Rec.*, vol. 2147, pp. 25–32, 2010.

[37] H. Etemadi, A. Rostamy, and H. Dehkordi, “A genetic programming model for bankruptcy prediction: Empirical evidence from Iran,” *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3199–3207, Mar. 2009.

[38] E. McKee and T. Lensberg, “Genetic programming and rough sets: A hybrid approach to bankruptcy classification,” *Eur. J. Oper. Res.*, vol. 138, no. 2, pp. 436–451, Apr. 2002.

[39] T. Lensberg, A. Eilifsen, and E. McKee, “Bankruptcy theory development and classification via genetic programming,” *Eur. J. Oper. Res.*, vol. 169, no. 2, pp. 677–697, Mar. 2006.

[40] M. Abdel-Aty, A. Pande, C. Lee, V. Gayah, R. Cunningham, A. Dhindsa, and J. Dilmore, “Linking crash patterns to ITS-related archived data—Phase II, Volume I: Real-time crash risk assessment models BD-550-5,” Florida Dept. Transp., Tallahassee, FL, 2007.

[41] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[42] R. Harb, X. Yan, E. Radwan, and X. Su, “Exploring precrash maneuvers using classification trees and random forests,” *Accid. Anal. Prev.*, vol. 41, no. 1, pp. 98–107, Jan. 2009.

[43] F. Xu and Z. Tian, “Driver behavior and gap-acceptance characteristics at roundabouts in California,” *Transp. Res. Rec.*, vol. 2071, pp. 117–124, 2008.

[44] P. Liu, J. Lu, H. Zhou, and G. Sokolow, “Operational effects of U-turns as alternatives to direct left-turns,” *J. Transp. Eng.*, vol. 133, no. 5, pp. 327–334, May 2007.

[45] S. Hubbard, D. Bullock, and F. Mannering, “Right turns on green and pedestrian level of service: Statistical assessment,” *J. Transp. Eng.*, vol. 135, no. 4, pp. 153–159, Apr. 2009.

[46] S. Washington, M. Karlaftis, and F. Mannering, *Statistical and Econometric Methods for Transportation Data Analysis*. Boca Raton, FL: CRC Press, 2003, pp. 250–253.

[47] *RandomForest: Breiman and Cutler’s Random Forests for Classification and Regression*. [Online]. Available: http://code.google.com/p/randomforest-matlab/

[48] S. Silva, *GPLAB—A Genetic Programming Toolbox for MATLAB*, MathWorks, Natick, MA. [Online]. Available: http://gplab.sourceforge.net/

[49] L. Davis, “Adapting operator probabilities in genetic algorithms,” in *Proc. 3rd Int. Conf. Genetic Algorithms*, 1989, pp. 61–69.

[50] S. Luke and L. Panait, “Lexicographic parsimony pressure,” in *Proc. GECCO*, 2002, pp. 829–836.

[51] D. Gazis, R. Herman, and G. H. Weiss, “Density oscillations between lanes of a multilane highway,” *Oper. Res.*, vol. 10, no. 5, pp. 658–667, Sep./Oct. 1962.

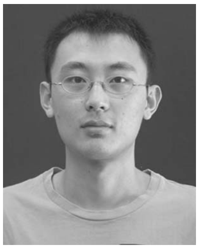
[52] M. Lord, A. Manar, and A. Vizioli, “Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments,” *Accid. Anal. Prev.*, vol. 37, no. 1, pp. 185–199, 2005.

[53] E. Hauer, *Observational Before–After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. New York: Pergamon, 2002.

[54] R. Carlson, I. Papamichail, and M. Papageorgiou, “Local feedback-based mainstream traffic flow control on motorways using variable speed limits,” *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1261–1276, Dec. 2011.

[55] J. Frejo and E. Camacho, “Global versus local MPC algorithms in freeway traffic control with ramp metering and variable speed limits,” *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1556–1565, Dec. 2012.

[56] D. Zhao, X. Bai, F.-Y. Wang, J. Xu, and W. Yu, “DHP method for ramp metering of freeway traffic,” *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 990–999, Dec. 2011.



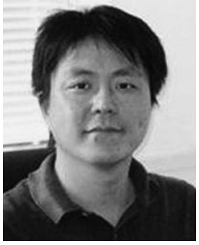
Chengcheng Xu received the B.Eng. and M.Eng. degrees in transportation planning and management from Southeast University, Nanjing, China, in 2008 and 2010, respectively. He is currently working toward the Ph.D. degree with the Key Laboratory of Traffic Planning and Management, School of Transportation, Southeast University.

His research interests include traffic safety and intelligent transportation systems.

Wei Wang received the M.Eng. and Ph.D. degrees in civil engineering from Southeast University, Nanjing, China, in 1985 and 1989, respectively.

He is currently a Professor with the Key Laboratory of Traffic Planning and Management, School of Transportation, Southeast University. His research interests include urban transportation and intelligent transportation systems.

Dr. Wang is a Member of the Model Traffic Technology Panel of the National High-Tech R&D Program of China (863 Program) and the panel of the National Natural Science Foundation of China. He received the National Distinguished Teacher Award of China in 2007.



Pan Liu received the Ph.D. degree in civil engineering from the University of South Florida, Tampa, in 2006.

He is currently a Professor with the Key Laboratory of Traffic Planning and Management, School of Transportation, Southeast University, Nanjing, China. His research interests include traffic design, traffic safety, and intelligent transportation systems.