Accident Analysis and Prevention 88 (2016) 9–19

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data

Ho-Chan Kwak[a,*], Seungyoung Kho[b]

[a] Green Transport and Logistics Institute, Korea Railroad Research Institute, Uiwang, Republic of Korea
[b] Department of Civil and Environmental Engineering, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 6 April 2015
Received in revised form 4 November 2015
Accepted 4 December 2015
Available online 19 December 2015

Keywords:
Crash risk prediction
Segment type
Traffic flow state
Genetic programming
Conditional logistic regression analysis
Loop detector

ABSTRACT

In order to improve traffic safety on expressways, it is important to develop proactive safety management strategies with consideration for segment types and traffic flow states because crash mechanisms have some differences by each condition. The primary objective of this study is to develop real-time crash risk prediction models for different segment types and traffic flow states on expressways. The mainline of expressways is divided into basic segment and ramp vicinity, and the traffic flow states are classified into uncongested and congested conditions. Also, Korean expressways have irregular intervals between loop detector stations. Therefore, we investigated on the effect and application of the detector stations at irregular intervals for the crash risk prediction on expressways. The most significant traffic variables were selected by conditional logistic regression analysis which could control confounding factors. Based on the selected traffic variables, separate models to predict crash risk were developed using genetic programming technique. The model estimation results showed that the traffic flow characteristics leading to crashes are differed by segment type and traffic flow state. Especially, the variables related to the intervals between detector stations had a significant influence on crash risk prediction under the uncongested condition. Finally, compared with the single model for all crashes and the logistic models used in previous studies, the proposed models showed higher prediction performance. The results of this study can be applied to develop more effective proactive safety management strategies for different segment types and traffic flow states on expressways with loop detector stations at irregular intervals.

1. Introduction

With the tremendous growth of Intelligent Transportation System (ITS) during the past decade, it is now possible to collect traffic parameters such as traffic volume, speed, and occupancy from a variety of detectors in real time. Several studies have attempted to demonstrate the potential application of loop detector data in order to reveal the relationship between crash occurrence and traffic parameters (Hughes and Council, 1999; Oh et al., 2001, 2005; Lee et al., 2002, 2003). Also, proactive safety management strategies utilizing Advanced Traffic Management Systems (ATMS) such as variable speed limits (VSL) and ramp metering have showed effects on improving traffic safety (Abdel-Aty et al., 2006, 2007; Lee et al., 2006a,b; Lee and Abdel-Aty, 2008). Within these studies, real-time crash risk prediction models were estimated to predict crash occurrence likelihood based on real-time traffic data.

The crash precursors mean traffic flow characteristics leading to crashes, which are identified by comparing the crash cases and non-crash cases. The crash precursors also will be different depending on traffic flow characteristics by surrounding environment on expressways. Especially, these traffic flow characteristics are divided by segment type and traffic flow state in Highway Capacity Manual (HCM). Therefore the real-time crash risk prediction models based on the crash precursors are required to develop for different segment types and traffic flow states on expressways. However, the majority of studies have developed a single model (e.g., aggregated model) for overall crashes (Abdel-Aty et al., 2004, 2008; Abdel-Aty and Pande, 2005; Ahmed and Abdel-Aty, 2012, 2013; Yu and Abdel-Aty, 2013). A major drawback of the aggregated modeling technique is that it cannot consider different effects of traffic flow characteristics on crash risk by roadway geometry and traffic flow states. Therefore, separated crash risk prediction models were developed by segment type and traffic flow state on expressways in this study. Also Korean expressways have irregular intervals between loop detector stations. Therefore, we investigated on the effect and application of the detector stations at irregular intervals for the crash risk prediction on expressways.

* Corresponding author.
E-mail addresses: kwak01@krri.re.kr, kwak.hochan@gmail.com (H.-C. Kwak).

http://dx.doi.org/10.1016/j.aap.2015.12.004
0001-4575/

Data used in this study were collected from the Gyeongbu expressway in Korea. The real-time traffic data (volume, speed and occupancy) was obtained from loop detectors and matched with historical crash data. The real-time crash risk prediction models were developed by segment type and traffic flow state. The mainline of expressways is divided into two segments based on ramp presence, basic segment and ramp vicinity, and the traffic flow states are classified into uncongested and congested conditions. Genetic programming technique was used in this study to develop the real-time crash risk prediction models. The genetic programming is a relatively new modeling technique that was proposed to solve classification and regression problems. Compared with traditional statistical regression methods and machine learning algorithms, two major advantages of genetic programming were proposed (Xu et al., 2013b). First, genetic programming can find a solution to a problem without any pre-specified functional forms. Second, in contrast with different machine learning algorithms, genetic programming can remove the "black box" effect and make the model understandable. Due to genetic programming models lack the ability to select significant variables that play a role to increase the reliability of prediction models, conditional logistic regression analysis was first estimated to select the most significant traffic variables contributing to crash occurrence. The conditional logistic regression analysis has an advantage which can control confounding factors by matching. Based on the chosen explanatory variables by conditional logistic regression analysis, the genetic programming models have been estimated and the relationship between traffic variables and crash risk has been investigated for each condition. Finally, the prediction performance of the proposed models have been compared to single model based on Receiver Operating Characteristics (ROC) curves and areas under the ROC curve (AUC).

2. Background

Real-time crash risk prediction models were estimated with the purpose of identifying the crash precursors and the results were applied in proactive safety management strategies. With the advanced traffic surveillance system (loop detectors, remote traffic microwave sensors, automatic vehicle identification systems), traffic flow characteristics prior to crash occurrence could be identified and matched with the crashes. Abdel-Aty et al. (2004) developed crash likelihood prediction model for freeway using matched case-control logistic regression based on loop detector data. The average occupancy at the upstream and the coefficient of variation in speed at downstream was found to affect crash occurrence most significantly. Abdel-Aty and Pande (2005) proposed probabilistic neural network to classify traffic speed pattern for crash and non-crash conditions from historical crash and loop detector data collected from the Interstate-4 corridor in the Orlando metropolitan area.

In order to increase prediction performance of the crash risk models, the separate models have been developed by the crash characteristics. Abdel-Aty et al. (2005) identified that multivehicle crashes under high- and low-speed traffic conditions on freeways were found to differ in severity and in their mechanism. The speed regime was divided into two speed conditions by the distribution of average speeds obtained immediately before the crash from the loop detector and the two models by speed condition were estimated using matched case-control logistic regression. Pande and Abdel-Aty (2006a,b) identified the traffic parameters leading to rear-end collisions and lane-change related collision from loop detector data on freeways. Classification tree was used in variable selection procedure, and then the classification models (crash versus non-crash) based on selected variables were developed by neural network. Xu et al. (2012) divided freeway traffic flow into different states, and evaluated the safety performance associated with each state. K-means clustering analysis was conducted to classify traffic flow into five different states, and conditional logistic regression models using case-controlled data were developed to identify the relationship between crash risks and traffic states. Hassan and Abdel-Aty (2013) investigated whether real-time traffic flow data, collected from loop detectors and radar sensors on freeways, can be used to predict crashes occurring at reduced visibility conditions. The result indicated that traffic variables leading to visibility related crashes are slightly different from those variables leading to clear visibility crashes. Xu et al. (2013a) developed a model that predicts the crash likelihood at different levels of severity with a particular focus on severe crashes. Crash severity was divided into three levels: fatal/incapacitating injury crashes (KA), non-incapacitating/possible injury crashes (BC), and property-damage-only crashes (PDO). The sequential logit model was used to link the likelihood of crash occurrences at different severity levels to various traffic flow characteristics identified from detector data.

Also the studies applying more advanced vehicle detection systems and modeling techniques have been performed to improve the prediction performance. Abdel-Aty et al. (2008) attempts to address the issues of transferability through analysis of crash data and loop detector data collected from Dutch freeways. In addition to these transferability issues, the application of a new data mining technique, random forests, was investigated for identifying significant variables associated with the crash. Ahmed and Abdel-Aty (2012) examined the identification of freeway locations with high crash potential using real-time speed data collected from automatic vehicle identification (AVI). Travel time and space mean speed data by AVI systems and crash data on the freeway network in Orlando were collected. Utilizing a random forest technique for significant variable selection and matched case-control method to account for the confounding effects, the log odds of crash occurrence were calculated. The results showed that the length of the AVI segment was found to be a crucial factor that affects the usefulness of the AVI traffic data and the likelihood of a crash is statistically related to speed data obtained from AVI segments within an average length of 1.5 mile. Hossain and Muromachi (2012, 2013a) employed random multinomial logit model to identify the most important predictors as well as the most suitable detector locations to build models for the basic freeway segment and the ramp areas of urban expressways. And Bayesian belief net (BBN) was applied to build the real-time crash prediction model. Hossain and Muromachi (2013b) identified crash influential factors and traffic patterns for different types of crashes on urban expressways. However, these studies were focused on crashes and traffic patterns on urban expressways rather than those on conventional regional expressways in this study. Compared to the conventional expressways, crashes and traffic patterns on urban expressways is different since urban expressways have a relatively lower speed limit and more frequent ramps. Xu et al. (2013b) evaluated the application of the genetic programming model for real-time crash prediction on freeways. Traffic, weather, and crash data were obtained from I-880N freeway in California, United States, and the random forest technique was conducted to select the variables that affect crash risk. The genetic programming models were found to increase the crash prediction accuracy compared with binary logit models. Ahmed and Abdel-Aty (2013) proposed a framework for real-time crash risk assessment on a freeway in Colorado fusing traffic data from two different detection systems (AVI and Remote Traffic Microwave Sensors (RTMS)), real-time weather data and roadway geometry data. Stochastic Gradient Boosting (SGB) is used to calibrate the model. Yu and Abdel-Aty (2013) introduced support vector machine (SVM) to evaluate real-time crash risk. Classification and regression tree (CART) model has been developed to select the most important traffic variables. The SVM models with

Table 1
Distribution of spacing between loop detector stations.

Spacing (m)	Distribution ratio	Descriptive statistics
0–250	1.9%	
250–500	2.2%	
500–750	5.3%	
750–1000	47.4%	Mean: 1064 m Std. dev.: 351 m Min.: 40 m Max.: 2540 m
1000–1250	22.0%	
1250–1500	12.7%	
1500–1750	4.6%	
1750–2000	0.6%	
>2000	3.3%	

different kernel functions have been developed and compared to the Bayesian logistic regression models.

3. Data source

3.1. Data collection

To estimate the relationships between traffic flow characteristics and crash risk, crash data and traffic data were collected from the Gyeongbu expressway in Korea. The Gyeongbu expressway, connecting Seoul to Busan, is the longest and most heavily traveled expressway in Korea. The entire length from Seoul to Busan is 416.0 km, and the yearly 371 million vehicles traveled through this expressway in 2010. The Gyeongbu expressway has irregular intervals between loop detector stations. Table 1 presents distribution of spacing between detector stations.

Crash data and traffic data were obtained for three years, 2008–2010. Crashes caused by vehicle defects, animal trespass, and vehicle fires were excluded from the dataset, because they have no relation with traffic conditions. In addition, only crashes that occurred from 7 a.m. to 10 p.m. were included in the dataset since crashes during late-night and early morning hours may be attributed mostly to human error rather than traffic conditions (Abdel-Aty and Pande, 2005).

The 5-min traffic volume, speed, and occupancy parameters were collected from loop detectors. Previous studies found that crash occurrence was mostly related to the six 5-min intervals during the 30 min prior to the crash (Pande et al., 2005; Xu et al., 2012) and the segments including three upstream stations and three downstream stations of the crash location (Abdel-Aty et al., 2008; Pande et al., 2011; Hassan and Abdel-Aty, 2013). Therefore, these time intervals and segments were considered in the data extraction process and modeling. Also the cases that ramps are within distance of six detector stations from the crash were excluded from dataset to remove the effect of ramps which are adjacent to the segment, and the distance of six detector stations is from 2.90 km to 12.30 km with an average length of 5.46 km and standard deviation of 1.00 km. The time-space diagram defining the time intervals and stations is shown in Fig. 1. All stations were labeled from "*a*" to "*f*", with "*a*" being the farthest upstream station and "*f*" being the farthest downstream station. A crash occurred between station *c* and station *d*. Similarly, the time intervals are labeled from *t1* to *t6*. The interval between time of the crash and 5 min prior to the crash was named as *t1*, interval between 5 and 10 min prior to the crash as *t2*, and so on. Time interval *t1* was excluded from the analysis, since it would not provide enough time for successful intervention to reduce crash risk in proactive safety management strategies.

To compare disruptive traffic conditions that led to a crash with normal traffic conditions that did not lead to a crash, traffic data for non-crash cases were also extracted in a similar fashion as described above. Non-crash cases were determined for the same season, day of the week, time, and location corresponding to each crash case. The same season is defined as duration from four weeks before to four weeks after the day of the crash. Therefore, a maximum of eight non-crash cases were matched to each crash case. Also, there is a crash happened within 5 h before crash time of corresponding crash case and within a 50-km radius of crash location of corresponding crash case in non-crash cases, the non-crash case is excluded from data set to remove effect by crashes. A total of 1256 crash cases and 8150 non-crash cases were included in the data set. The ratio of non-crash cases to crash case in each matched pair is not consistent from 4:1 to 8:1.

3.2. Data preparation

The traffic data used in this study were collected from loop detectors, and the average, standard deviation, coefficient of variation, and variation rate for three traffic parameters (volume, speed, occupancy) were calculated over detector stations or time intervals. The coefficient of variation is defined as the ratio of the standard deviation to the average and indicates the relative variation among traffic parameters. The variation rate is utilized as a variable to represent characteristic of detector system at irregular interval, which is calculated by dividing the difference in traffic parameters between two stations into the distance between them. It indicates how the traffic condition is drastically changed.

The traffic variables used in the analysis are expressed as XY_{α}-β. Here, *X* takes *A*, *S*, *C*, or *R* for average, standard deviation, coefficient of variation, or variation rate, respectively; *Y* takes *V*, *S*, or *O* for volume, speed, or occupancy. If a variable *XY* is calculated over detector stations, then α indicates detector stations (e.g., *a*, *b*, *c*, *d*, *e*, or *f*), while β represents time intervals (e.g., 2, 3, 4, 5, or 6). On the contrary, if a variable *XY* is computed over time intervals, then α represents the time intervals, while β refers to the detector stations.

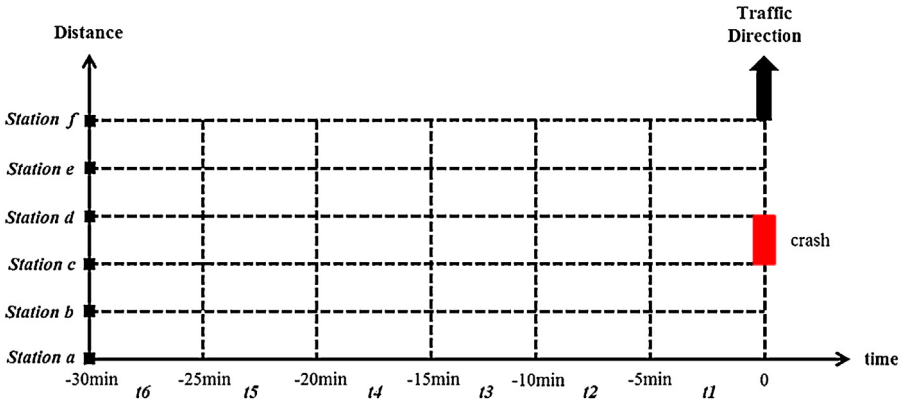

Fig. 1. Time-space diagram for dataset.

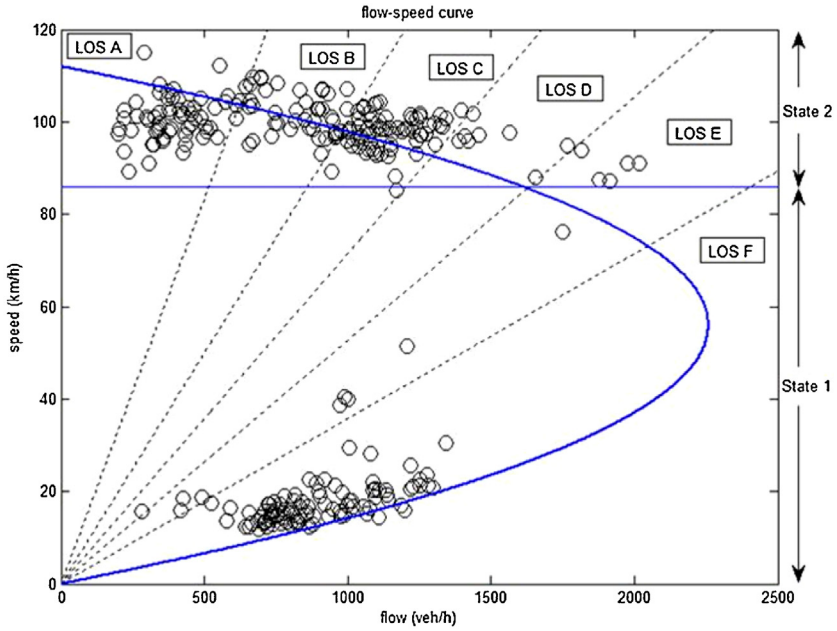

Fig. 2. Flow-speed diagram to identify the critical speed.

For example, AV_{abcd}_2 represents the average volume for station *a*, *b*, *c*, and *d* at *t2*, and SS_{234}_a represents the standard deviation of speed for *t2*, *t3*, and *t4* at station *a*.

In this study, separate crash risk prediction models are developed for different segment types and traffic flow states. For segment types, the mainline of expressways is divided into two segments based on ramp presence: basic segment and ramp vicinity. As mentioned above, the influence area of crashes refer to the terms upstream and downstream detectors in a relative spatial relationship based on order of occurrence rather than in the unit of distance. Hence, the basic segment has no ramp within spacing between upstream detectors and downstream detectors of crash location. The ramp vicinity has off-ramp at the upstream segment and on-ramp at the downstream segment of crash location, and this is the most conventional type on Gyeongbu expressway in Korea. According to Fig. 1, detector station *a* at upstream segment is placed in upstream of off-ramp and detector station *f* at downstream segment is placed in downstream of on-ramp. Thus the ramp vicinities in this study cover upstream of off-ramp, downstream of on-ramp and segment between off-ramp and on-ramp.

The traffic flow states are classified into uncongested and congested conditions based on a critical speed index as done in a previous study by Abdel-Aty et al. (2005). Fig. 2 presents the flow-speed diagram using real-time data obtained from a section without crashes on Gyeongbu expressway. The level of service (LOS) analysis in the flow-speed diagram was performed based on Korean Highway Capacity Manuel (KHCM). The uncongested condition means from LOS A to D, and the congested condition means from LOS E to F. According to the LOS analysis in this study, the critical speed was 86 km/h.

The original data set was separated into four subsamples: basic segment/uncongested condition, basic segment/congested condition, ramp vicinity/uncongested condition, and ramp vicinity/congested condition. Table 2 presents the dataset used in the study. Each subsample was further separated into a training data set and a validation data set with a ratio of 2:1. The former was used to develop the crash risk prediction models, and the latter was used to test the prediction performance of models developed in the study.

4. Methodology

4.1. Conditional logistic regression analysis

The matched case-control structure was used in this study. It was adopted to identify the significant traffic variables leading to crash occurrence while controlling for time of the day, day of the week, season, and location (i.e., geometric characteristics). Therefore, conditional logistic regression analysis is used in variable selection process, it is expected to provide accurate results as the effects of confounding factors are controlled by matching. The modeling is estimated under the conditional likelihood principle of statistical theory accounting for within-stratum differences between crash and non-crash traffic variables. Assume that there are N strata with one crash and m non-crash cases in stratum j, where $j = 1, 2, 3, \ldots, N$. $p_j(x_{ij})$ is the probability of the ith observation in the jth stratum being a crash; x_{ij} is the vector of traffic variable; $i = 0, 1, 2, 3, \ldots, m$; and $j = 1, 2, 3, \ldots, N$.

$$\text{logit}(p_j(x_{ij})) = \alpha_i + \beta x_{ij}$$

Note that the intercept term α summarizes the effect of factors used to form strata on the crash likelihood and differs across strata. Conditional likelihood is constructed to take account of the stratification in the analysis, and conditional likelihood function has the form of

$$L(\beta) = \prod_{j=1}^{n} \left[1 + \sum_{i=1}^{m} \exp\{\beta(x_{ij} - x_{0j})\} \right]^{-1}$$

Table 2
Sample size of crash and non-crash data for each subsample.

Segment type	Traffic flow state	Crash	Non-crash
Basic segment	Uncongested condition	819	5754
	Congested condition	159	709
Ramp vicinity	Uncongested condition	175	1239
	Congested condition	103	448
Total		1256	8150

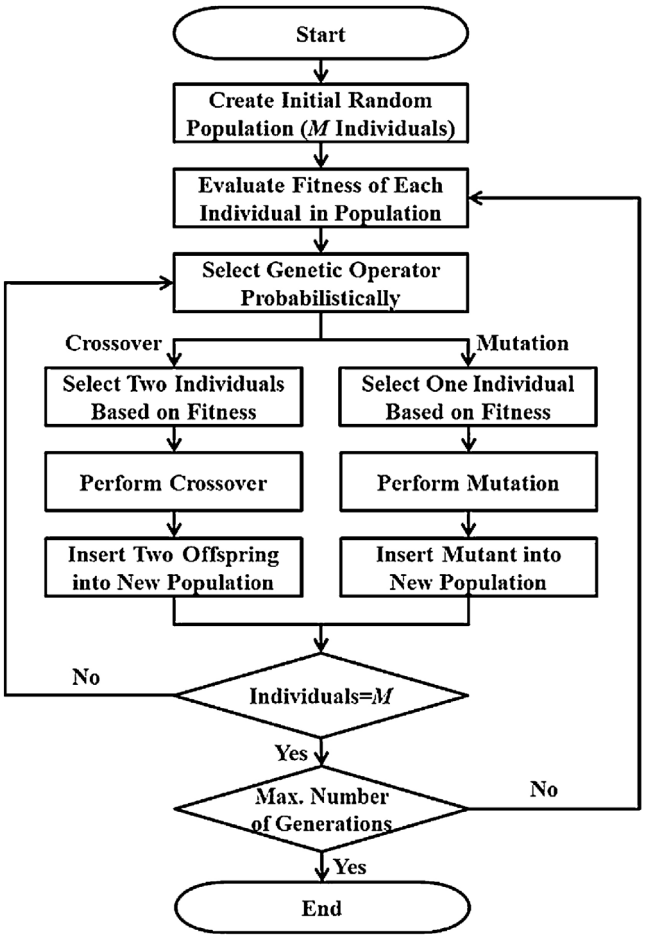

Fig. 3. Flowchart of genetic programming.

This conditional likelihood function $L(\beta)$ is independent of the intercept terms $\alpha_1, \alpha_2, \ldots, \alpha_N$. Hence, the effects of matching factors cannot be estimated, and crash likelihood cannot be directly predicted using conditional logistic regression model. Thus, conditional logistic regression analysis was only used to select the most significant traffic variables in this study.

4.2. Genetic programming

Genetic programming works on a population of mathematical models that are coded as function trees (individuals) based on evolution theory (Koza, 1992). In each generation, multiple models are stochastically selected based on their fitness and modified to form a new population of models by crossover and mutation operators. The new population of models is then used in the next iteration of the algorithm. The algorithm stops when the predetermined number of generations has been reached. Therefore, the evolution process is expected to continuously produce a better model for the problem to be solved. Fig. 3 presents the flowchart of the genetic programming.

In genetic programming, the following three genetic operators are typically used to create new individuals: reproduction, crossover, and mutation (Silva, 2007). The reproduction operator selects a proportion of individuals and copies them into the next generation without suffering the action of the operators. The crossover operator creates new individuals (offspring) by combining information that was extracted from selected parents. Two individuals (parents) are randomly selected based on their fitness level, and random nodes are chosen from both parent trees. Then, the respective branches are swapped creating two offspring. The mutation is an important operator to introduce new information into the population and avoid the premature convergence of a genetic programming model. In mutation, a single individual (parent) is randomly selected based on their fitness level, and a random node is chosen from the parent tree and substituted by a new random tree created with the pre-specified set of mathematical operators and predictors. In this study, only crossover and mutation operators are used to improve the efficiency of model training procedure since a reproduction operator cannot contribute to improve a fitness of model.

One of the most important components in genetic programming is the fitness function, which determines how well a model in the population can solve the problem. The fitness function is developed based on the error between the predicted values and the actual values. For classification problems, the most commonly used fitness functions include the number of hits, sensitivity or specificity, and square error. In this study, a fitness function was developed based on the square errors. For the binary classification problem (crash and non-crash), given the training data

$$(x_1, y_1), \ldots, (x_i, y_i) \in \{0, 1\}$$

assuming that for the crash cases $y_i = 1$ and $y_i = 0$ for the non-crash cases. x_i represents the explanatory variables selected by the conditional logistic regression model. And the fitness function has the functional form

$$f(h) = \sum_i [\beta \cdot (y_i - h_j(x_i))^2]$$

where $f(h_j)$ denotes the fitness of the jth model h_j in the population, $h_j(x_i)$ is the value calculated by the jth model h_j in the population.

Because the number of non-crash cases is much greater than that of crash cases in the training data set, the genetic programming model might ignore the information from crash cases and classify all the observations as non-crash 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 crash and non-crash cases in each training data set.

5. Data analysis and results

5.1. Preliminary analysis

A preliminary analysis of characteristics for crashes used in this study was conducted to investigate the crash mechanism by segment type and traffic flow state on expressways. Table 3 presents the results of the preliminary analysis. On the basic segment, the percentage of crashes caused by unsafe speed, lack of visual attention, and following too closely under the congested condition is larger than that under the uncongested condition. Also, the crashes associated with multiple vehicles and occurred under adverse weather conditions had a higher percentage under the congested condition than under the uncongested condition. One possible explanation is that a crash occurs because drivers cannot reduce their speed gradually when they suddenly encounter a relatively higher traffic density. That is to say, the incorrect braking by careless driving under high traffic density with interference between vehicles might cause the crashes under the congested condition for the basic segment. On the other hand, crashes resulted from exaggerated steering control, single vehicle and vehicle-facility crashes, and crashes at the curve section had a higher percentage in the uncongested condition. According to these results, the crashes that occurred in the uncongested condition on the basic segment might be related to the high speed and incorrect steering at the curve

Table 3
Distribution of crashes.

Factors	Categories	Basic segments		Ramp vicinities	
		Uncongested conditions	Congested conditions	Uncongested conditions	Congested conditions
Crash causation	Unsafe speed	26.0%	36.4%	23.3%	23.8%
	Exaggerated steering control	28.6%	18.1%	33.1%	26.2%
	Lack of visual attention	15.4%	23.7%	18.6%	23.8%
	Following too closely	3.7%	6.4%	3.5%	11.9%
	Others	26.3%	15.4%	21.5%	14.3%
Crash type	Vehicle–facility	67.4%	58.3%	67.5%	54.8%
	Single vehicle	12.8%	12.2%	17.4%	9.5%
	Multiple vehicles	19.8%	29.5%	15.1%	35.7%
Road alignment	Straight	66.1%	75.5%	65.7%	85.7%
	Curve	33.9%	24.5%	34.3%	14.3%
Weather	Clear	83.3%	57.4%	74.4%	45.2%
	Adverse	16.7%	42.6%	25.6%	54.8%

section. On the ramp vicinity, the percentage of crashes caused by exaggerated steering control was the largest in the congested and uncongested conditions. This means that the crashes on the ramp vicinity were influenced by lane change. Also, the crashes on the ramp vicinity had a higher percentage on straight section and under adverse weather than that on the basic segment. These results are represented that the crashes on the ramp vicinity were more influenced by weather condition than that on the basic segment.

In conclusion, the preliminary analysis results demonstrate that there are some differences in the crash mechanism by segment type and traffic flow state on expressways. Thus, to better understand the relationship between crash risk and traffic flow characteristics for different segment types and traffic flow states, separate crash risk prediction models for each condition are developed in this study.

5.2. Variable selection

There are numerous combinations of traffic parameters collected from loop detectors; however, genetic programming lacks the ability to select significant variables that play a role to increase the reliability of prediction models. Thus, an exploration step was required to select the significant variables. The most significant traffic variables were selected to develop crash risk prediction models by simple conditional logistic regression analysis. Also, the variables with correlation coefficient above 0.6 were removed from final selection.

Table 4 shows the final selected variables that satisfied the 95% significance level by conditional logistic regression analysis. As expected, the traffic variables that contribute to crash risk were quite different by segment type and traffic flow state on expressways. For the uncongested condition on the basic segment, three traffic variables were selected: average speed for station c and d at $t2$ (AS_{cd}_2), standard deviation of speed from $t2$ to $t3$ at station d (SS_{23}_d), and variation rate of speed between upstream station c and downstream station d at $t2$ (RS_{cd}_2). Under the congested condition on the basic segment, the following traffic variables were selected: average speed for stations b and c at $t2$ (AS_{bc}_2), standard deviation of occupancy from $t2$ to $t4$ at station e (SO_{234}_e), and coefficient of variation of speed between upstream station c and downstream station d at $t2$ (CS_{cd}_2). For the uncongested condition on the ramp vicinity, also three traffic variables were selected: average speed between stations b and f at $t2$ (AS_{bcdef}_2), coefficient of variation of speed for stations a and b at $t2$ (CS_{ab}_2), and variation rate of speed between upstream station b and downstream station d at $t2$ (RS_{bd}_2). Under the congested condition on the ramp vicinity, the selected traffic variables were as follows: average occupancy between stations b and e at $t2$ (AO_{bcde}_2), coefficient of variation of occupancy from station a and station b at $t2$ (CO_{ab}_2), and coefficient of variation of occupancy between upstream station c and downstream station e at $t2$ (CO_{cde}_2). Especially, it was represented that the variation rate variables had significance under uncongested conditions, and it means that the interval between detector stations has mainly influence on crash risk prediction under uncongested conditions.

Table 4
Final selected traffic variables.

Segment types	Traffic flow states	Traffic variables	$P>\|z\|$
Basic segment	Uncongested condition	AS_{cd}_2 ($X1$)	<.001
		SS_{23}_d ($X2$)	0.008
		RS_{cd}_2 ($X3$)	0.007
	Congested condition	AS_{bc}_2 ($X1$)	0.007
		SO_{234}_e ($X2$)	0.003
		CS_{cd}_2 ($X3$)	0.007
Ramp vicinity	Uncongested condition	AS_{bcdef}_2 ($X1$)	<.001
		CS_{ab}_2 ($X2$)	0.015
		RS_{bd}_2 ($X3$)	0.035
	Congested condition	AO_{bcde}_2 ($X1$)	0.005
		CO_{ab}_2 ($X2$)	0.014
		CO_{cde}_2 ($X3$)	0.035

5.3. Model results

The GPLAB toolbox 3.0, which is coded in MATLAB, was used to develop the genetic programming model. The parameters used in GPLAB are as follows. The population size was set to 1000, which is large enough to produce various individuals. The maximum number of generations was 100. In order to avoid the "bloat" phenomenon, excessive code growth without the corresponding improvement in fitness (Silva, 2007), the maximum tree depth was limited to 30. The function set contained six standard arithmetic operators, such as +, −, ×, ÷, protected square root, and protected natural logarithm, which can express most of mathematical models solving classification problem. These operators were generally used to build genetic programming model on this account (Das and Abdel-Aty, 2010; Das et al., 2010; Xu et al., 2013a,b). The terminal set included the traffic variables selected by conditional logistic regression analysis.

A total of four mathematical tree models to predict crash risk were developed for each condition. A complex relationship between crash risk and traffic variables was represented in genetic programming model. A change in crash risk is indicated as the value of each variable changes. The values of each variable were changed continuously over a normal range when the other variables were kept at their sample mean.

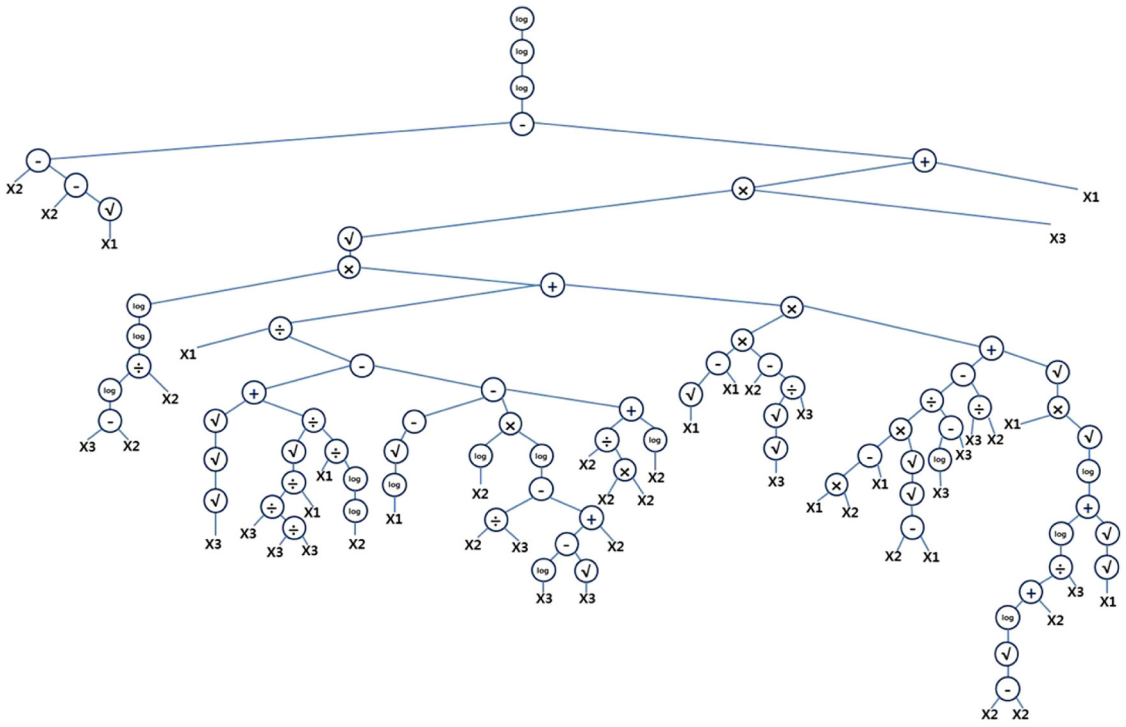

Fig. 4. The genetic programming model for uncongested condition on basic segment.

The genetic programming crash risk prediction models for uncongested and congested conditions on basic segment were presented in Figs. 4 and 5. Under the uncongested condition on the basic segment, as shown in Fig. 6, the crash risk increases as the average speed for the closest stations from crash location decreases. The decreasing speed might represent the increase in traffic density and in interference from different vehicles. The speed variation between the closest stations from crash location caused by the disruptions at the closest downstream stations was associated with the increase in crash risk. These results are consistent with the findings of previous studies (Abdel-Aty et al., 2005; Xu et al., 2013b). Especially, as the speed variation rate related to the

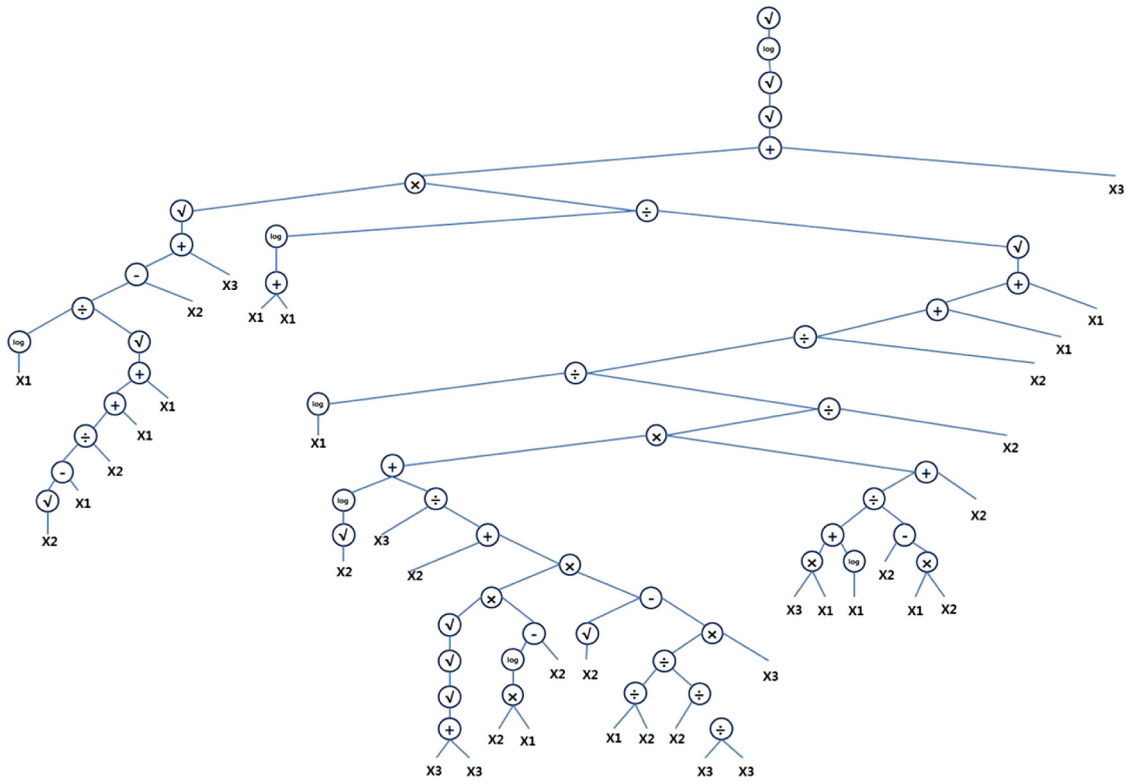

Fig. 5. The genetic programming model for congested condition on basic segment.

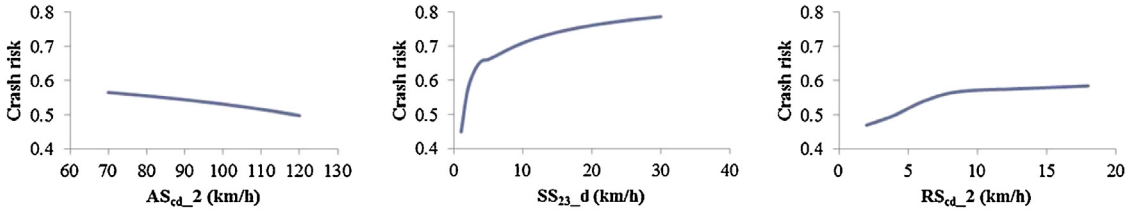

Fig. 6. Relationship between crash risk and traffic variables for uncongested condition on basic segment.

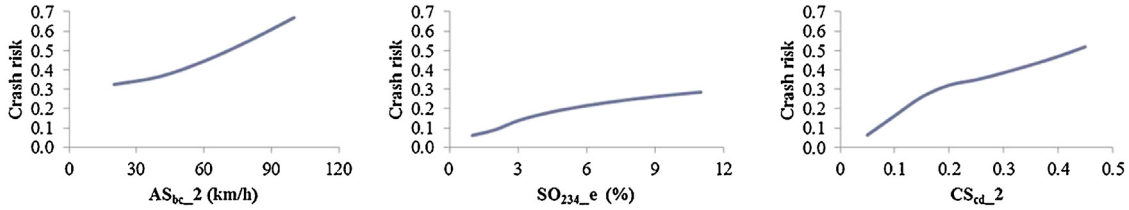

Fig. 7. Relationship between crash risk and traffic variables for congested condition on basic segment.

distance between detector stations was increased, crash risk was also increased. It represented that the drastic variation of speed under high speed condition was relatively increased crash risk.

As shown in Fig. 7, the high speed at the upstream stations and the speed difference between upstream and downstream stations were found to be associated with the increase in crash risk under the congested condition on the basic segment. Previous studies have demonstrated that crashes are more likely to occur when the traffic parameters have a large difference between upstream and downstream stations (Hossain and Muromachi, 2012, 2013b; Xu et al., 2012, 2013b). Also, queue formations by downstream congestion induce an increase in crash risk (Abdel-Aty et al., 2005). These results are consistent with the findings shown in the preliminary analysis. However, the pattern mentioned in the research by Hossain and Muromachi (2013b) – a rapidly progressing downstream trailed by a congested upstream – is not represented in this study. It may be because of the characteristic of urban expressways having frequent ramps than conventional regional expressways in this study.

The genetic programming models for uncongested and congested conditions on ramp vicinity were indicated in Figs. 8 and 9. As presented in Figs. 10 and 11, on ramp vicinities, the variation in speed or occupancy resulting from the ramp presence influenced the increase in crash risk. This might be associated with high lane change frequency on the ramp vicinity. This result was consistent

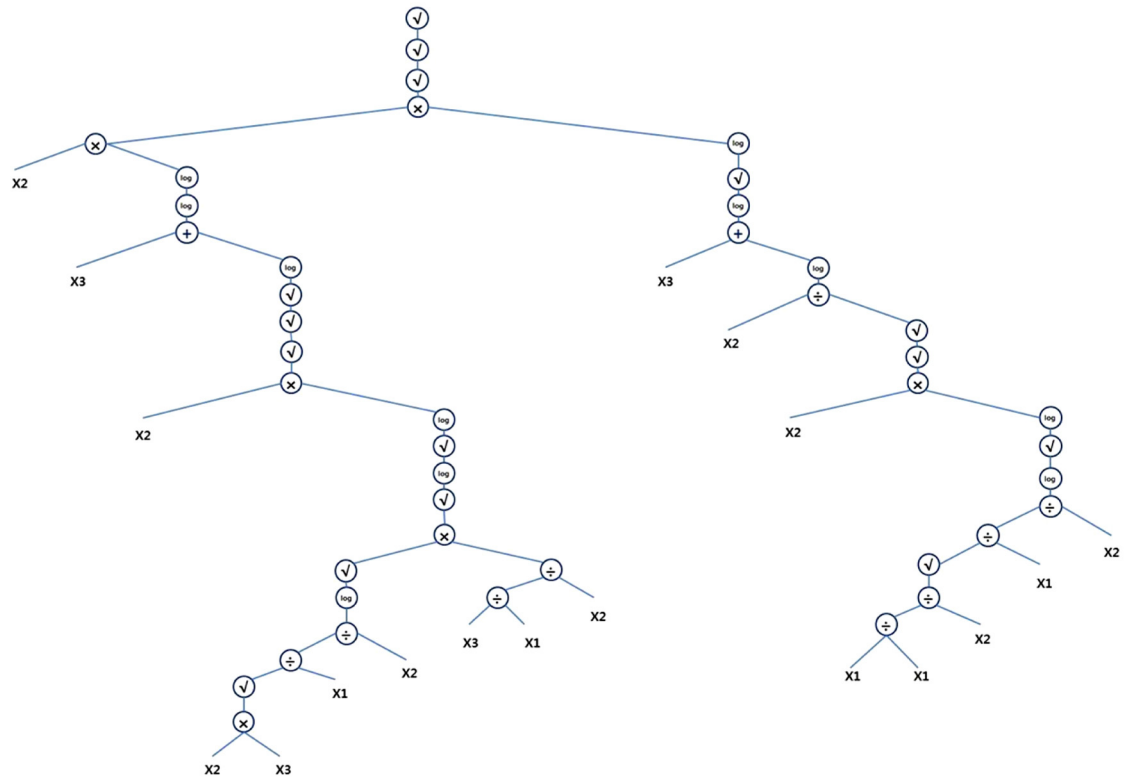

Fig. 8. The genetic programming model for uncongested condition on ramp vicinity.

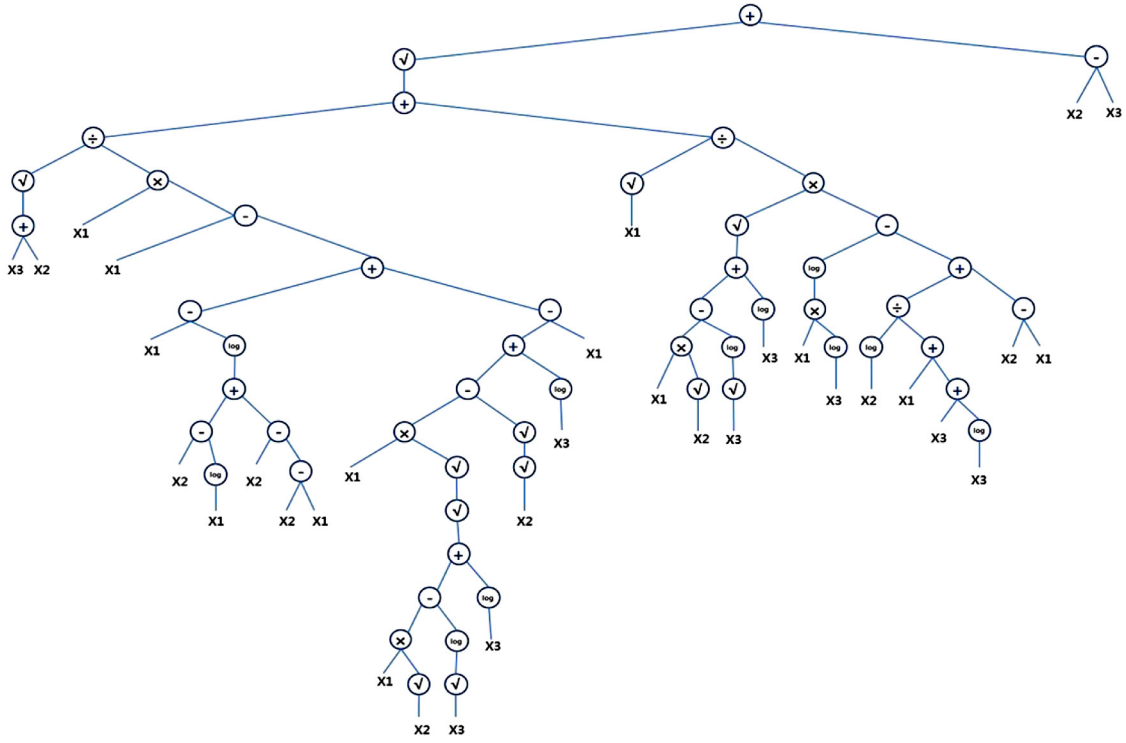

Fig. 9. The genetic programming model for congested condition on ramp vicinity.

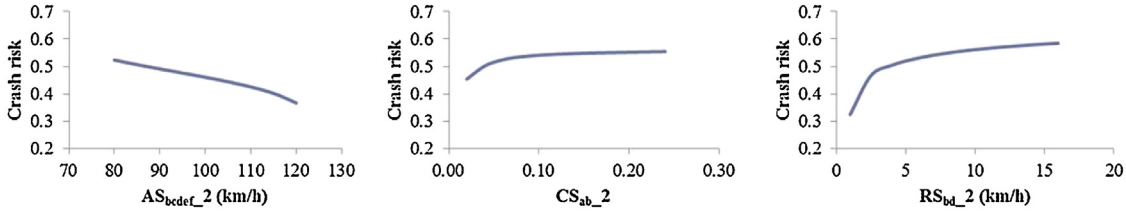

Fig. 10. Relationship between crash risk and traffic variables for uncongested condition on ramp vicinity.

with the findings shown in the preliminary analysis. Under the uncongested condition on the ramp vicinity, as shown in Fig. 10, crash risk increased as the average speed decreased. Also, high variation in speed between upstream stations *a* and *b* and high variation rate in speed between stations *b* and *d* had influence on the increase of the crash risk, and they were resulted from the presence of ramps. Also like on the basic segment, the crash risk was increased as the speed variation rate related to the distance between detector stations was increased.

Under the congested condition on the ramp vicinity, as shown in Fig. 11, low occupancy at the section between the upstream exit ramp and the downstream entrance ramp increased the crash risk. Also, high variation in occupancy at the upstream exit ramp vicinity and low variation in occupancy at the downstream entrance ramp vicinity were associated with the increase in crash risk.

Table 5
Prediction performance of the proposed models.

Segment types	Traffic flow states	AUC	
		GP model	Logistic model
Single model		0.5875	
Basic segment	Uncongested condition	0.6084	0.5863
	Congested condition	0.7422	0.6531
Ramp vicinity	Uncongested condition	0.7007	0.6756
	Congested condition	0.7892	0.7437

According to these results, the congestion level at the section between the upstream exit ramp and the downstream entrance ramp decreases temporarily due to the high outflow in upstream exit ramp and the low inflow in downstream entrance ramp.

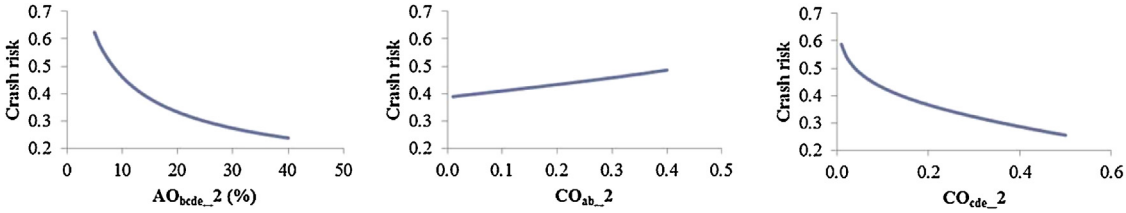

Fig. 11. Relationship between crash risk and traffic variables for congested condition on ramp vicinity.

Table 6
Comparison results of prediction accuracy.

False positive rate (1-specificity)	Sensitivity						
	Multiple models					Single model	Difference
	Basic/Uncon.	Basic/Con.	Ramp/Uncon.	Ramp/Con.	Total		
10%	17.9%	32.1%	27.6%	41.7%	23.0%	16.0%	7.0%
20%	35.4%	45.3%	44.8%	54.2%	39.5%	29.0%	10.5%
30%	46.3%	67.9%	58.6%	83.3%	53.8%	42.6%	11.2%
40%	57.5%	77.4%	65.5%	87.5%	63.6%	53.6%	10.0%

A crash occurs, because drivers cannot reduce their speed gradually after passing the upstream exit ramp and suddenly encounter a queue due to inflow from the downstream entrance ramp. These results are consistent with the findings shown in the previous study (Hossain and Muromachi, 2013b). The research identified that the difference in occupancy and level of congestion play important role in the ramp vicinities and level of congestion difference between the downstream and upstream leads to hazardous condition.

5.4. Prediction performance

Receiver Operating Characteristics (ROC) curves and areas under the ROC curve (AUC) were chosen to evaluate and compare the prediction performance of the proposed models. The ROC curve is a graphical plot of the sensitivity on the *y*-axis against (1-specificity) on the *x*-axis for different cutoff points. Sensitivity is the proportion of crashes that are correctly identified as crashes, while specificity is the proportion of non-crashes that are correctly identified as non-crashes by the estimated models (Agresti, 2002). Therefore, (1-specificity) represents a false positive rate, which is the proportion of non-crashes that are classified incorrectly as crashes. As mentioned before, the ROC curves are developed on the basis of the validation sample in this study.

To test the relative prediction performance of the proposed models, they were compared with the single model for all crashes and the logistic models used in previous studies. Table 5 illustrates the AUC value for proposed models and the compared models. Larger AUC values indicate better goodness-of-fit and classification performance. As represented in Table 5, the AUC values for the proposed models were always larger than for the single model and logistic models. In particular, the prediction performance of models for ramp vicinity and for congested conditions was clearly superior.

Table 6 shows the comparison results of prediction accuracy for multiple models constructed by segment type and traffic flow state and for the single model. The sensitivity for total of multiple models is calculated by the ratio of sum of samples corresponding to sensitivity for each condition to total samples by false positive rate. If a cutoff point is selected to accept 10% false positive rate, the sensitivity of multiple models is 7.0% higher than that of the single model. In the case that the false positive rate is 30%, the multiple models could increase the crash prediction accuracy by 11.2% compared to the single model. On the whole, the sensitivity of multiple models is improved as the false positive rate increases. However, it should be noted that higher false positive rate provides false alarm to drivers more frequently, and thus drivers are more likely to refuse to comply with crash risk warning. Therefore, more careful attention should be paid to the selection of an optimal threshold to set up a reasonable false positive rate.

6. Conclusion

In order to improve traffic safety on expressways, it is important to develop proactive safety management strategies, such as providing crash risk information and controlling traffic flow to reduce crash risk in real time, with consideration for segment types and traffic flow states because crash mechanisms had some differences by each condition. Hence, this study aimed to identify crash precursors which mean the traffic conditions leading to crash occurrence and to develop separate crash risk prediction models for different segment types and traffic flow states. Also, because Korean expressways have irregular intervals between loop detector stations, we investigated on the effect and application of the detector stations at irregular intervals for the crash risk prediction on expressways.

The relationships between traffic flow characteristics and crash risk were established through a detailed analysis of traffic data corresponding to crashes that occurred on the mainline of the Gyeongbu expressway in Korea. The mainline was divided into basic segment and ramp vicinity, and the traffic flow states were classified into uncongested and congested conditions.

Preliminary analysis results on the crash data indicated that the crash mechanisms by segment type and traffic flow state on expressways were quite different; hence, four models for each condition were considered to investigate the impacts of the traffic flow variables on crash risk. The most significant traffic variables were selected by conditional logistic regression analysis. There was a noticeable difference between selected traffic variables for each condition. Based on the selected variables, the crash risk prediction models were developed using genetic programming. The model estimation results showed that the traffic flow characteristics contributing to crash risk differed by segment type and traffic flow state. High traffic density and disruptions were the main contributing factors to crash risk under uncongested condition on the basic segment. Under the congested condition on the basic segment, a large difference in speed between upstream and downstream stations and queue formations at downstream stations increased crash risk. Also on ramp vicinities, variation in speed or occupancy due to lane changes or queue formations influenced the increase in crash risk. Especially, it was represented that the speed variation rate related to the interval between detector stations have mainly influenced on crash risk prediction under uncongested conditions.

The prediction performance of the proposed models was evaluated by ROC curve and AUC. Compared with the single model for all crashes and the logistic models used in previous studies, the proposed models showed higher prediction performance. The sensitivity of multiple models proposed in this study was 7.0–11.2% higher than that of the single model according to different false positive rates. These results indicate that estimating separate models by segment type and traffic flow state on expressways are more desirable to predict crash risk with traffic parameters.

The findings of this study can help us better understand the relationship between traffic flow characteristics and crash risk under different conditions for segment types and traffic flow states, which is important for developing crash prevention strategies under each condition. For example, providing a warning of crash risk to drivers may help to reduce crash occurrence since drivers will tend to drive more carefully while traveling a hazardous roadway segment. Moreover, a variable speed limit might be implemented to reduce

speed variability before reaching the end of queue under the congested condition. However, prior to field application, there is a need to conduct additional research for testing the transferability of the models using data collected from other expressways. Also, exploration of the optimal threshold is needed. A high false positive rate might affect driver compliance to the system, while a high false negative rate could miss actual crashes. Hence, a threshold which maximizes the effectiveness of the system should be chosen carefully. Finally, it is required to improve the prediction performance of models by combination of additional data such as weather related data and lane-based traffic data, and to present the methodology to enhance simultaneously the stability and accuracy of the models in case of the failure of partial detectors.

References

Abdel-Aty, M., Dhindsa, A., Gayah, V., 2007. Considering various ALINEA ramp metering strategies for crash risk mitigation on freeways under congested regime. Transp. Res. Part C 15, 113–134.

Abdel-Aty, M., Dilmore, J., Dhindsa, A., 2006. Evaluation of variable speed limits for real-time freeway safety improvement. Accid. Anal. Prev. 38, 335–345.

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Saf. Res. 36, 97–108.

Abdel-Aty, M., Pande, A., Das, A., Knibbe, W.J., 2008. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transp. Res. Rec.: J. Transp. Res. Board 2083, 153–161.

Abdel-Aty, M., Uddin, N., Pande, A., 2005. Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. Transp. Res. Rec.: J. Transp. Res. Board 1908, 51–58.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M.F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec.: J. Transp. Res. Board 1897, 88–95.

Agresti, A., 2002. Categorical Data Analysis, Second Ed. John Wiley & Sons, Inc, Hoboken, NJ.

Ahmed, M., Abdel-Aty, M., 2013. A data fusion framework for real-time risk assessment on freeways. Transp. Res. Part C 26, 203–213.

Ahmed, M.M., Abdel-Aty, M.A., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Trans. Intell. Transp. Syst. 13 (2), 459–468.

Das, A., Abdel-Aty, M., 2010. A genetic programming approach to explore the crash severity on multi-lane roads. Accid. Anal. Prev. 42, 548–557.

Das, A., Abdel-Aty, M., Pande, A., 2010. Genetic programming to investigate design parameters contributing to crash occurrence on urban arterials. Transp. Res. Rec.: J. Transp. Res. Board 2147, 25–32.

Hassan, H.M., Abdel-Aty, M.A., 2013. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. J. Saf. Res. 45, 29–36.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381.

Hossain, M., Muromachi, Y., 2013a. A real-time crash prediction model for the ramp vicinities of urban expressways. IATSS Res. 37, 68–79.

Hossain, M., Muromachi, Y., 2013b. Understanding crash mechanism on urban expressways using high-resolution traffic data. Accid. Anal. Prev. 57, 17–29.

Hughes, R.G., Council, F.M., 1999. On establishing the relationship(s) between freeway safety and peak period operations: performance measurement and methodological considerations. In: Presented at the 78th Transportation Research Board Annual Meeting, Washington, DC.

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

Lee, C., Abdel-Aty, M., 2008. Testing effects of warning messages and variable speed limits on driver behavior using driving simulator. Transp. Res. Rec.: J. Transp. Res. Board 2069, 55–64.

Lee, C., Hellinga, B., Ozbay, K., 2006a. Quantifying effects of ramp metering on freeway safety. Accid. Anal. Prev. 38, 279–288.

Lee, C., Hellinga, B., Saccomanno, F., 2003. Real-time crash prediction model for application to crash prevention in freeway traffic. Transp. Res. Rec.: J. Transp. Res. Board 1840, 67–77.

Lee, C., Hellinga, B., Saccomanno, F., 2006b. Evaluation of variable speed limits to improve traffic safety. Transp. Res. Part C 14, 213–228.

Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. Transp. Res. Rec.: J. Transp. Res. Board 1784, 1–8.

Oh, C., Oh, J.-S., Ritchie, S.G., 2005. Real-time hazardous traffic condition warning system: framework and evaluation. IEEE Trans. Intell. Transp. Syst. 6 (3), 265–272.

Oh, C., Oh, J.-S., Ritchie, S.G., Chang, M., 2001. Real time estimation of freeway accident likelihood. In: Presented at the 80th Transportation Research Board Annual Meeting, Washington, DC.

Pande, A., Abdel-Aty, M., 2006a. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38, 936–948.

Pande, A., Abdel-Aty, M., 2006b. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transp. Res. Rec.: J. Transp. Res. Board 1953, 31–40.

Pande, A., Abdel-Aty, M., Hsia, L., 2005. Spatiotemporal variation of risk preceding crashes on freeways. Transp. Res. Rec.: J. Transp. Res. Board 1908, 26–36.

Pande, A., Das, A., Abdel-Aty, M., Hassan, H., 2011. Estimation of real-time crash risk: are all freeways created equal? Transp. Res. Rec.: J. Transp. Res. Board 2237, 60–66.

Silva, S., 2007. GPLAB: A genetic programming toolbox for MATLAB, version 3. http://gplab.sourceforge.net/.

Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accid. Anal. Prev. 47, 162–171.

Xu, C., Tarko, A.P., Wang, W., Liu, P., 2013a. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39.

Xu, C., Wang, W., Liu, P., 2013b. A genetic programming model for real-time crash prediction on freeways. IEEE Trans. Intell. Transp. Syst. 14 (2), 574–586.

Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.