Accident Analysis and Prevention 122 (2019) 378–384

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors

Ling Wang[a,b,*], Mohamed Abdel-Aty[a], Jaeyoung Lee[a], Qi Shi[c]

[a] Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816, USA
[b] College of Transportation Engineering, Tongji University, Shanghai 201804, China
[c] Research Institute of Highway, Ministry of Transportation, Beijing 10088, China

ARTICLE INFO

Keyword:
Real-time crash prediction
Expressway ramps
Socio-demographic predictors
Trip generation predictors
Support vector machine

ABSTRACT

There have been numerous studies on real-time crash prediction seeking to link real-time crash likelihood with traffic and environmental predictors. Nevertheless, none has explored the impact of socio-demographic and trip generation parameters on real-time crash risk. This study analyzed the real-time crash risk for expressway ramps using traffic, geometric, socio-demographic, and trip generation predictors. Two Bayesian logistic regression models were utilized to identify crash precursors and their impact on ramp crash risk. Meanwhile, four Support Vector Machines (SVM) were applied to predict crash occurrence. Bayesian logistic regression models and SVMs commonly showed that the models with the socio-demographic and trip generation variables outperform their counterparts without those parameters. It indicates that the socio-demographic and trip generation parameters have significant impact on the real-time crash risk. The Bayesian logistic regression model results showed that the logarithm of vehicle count, speed, and percentage of home-based-work production had positive impact on crash risk. Meanwhile, off-ramps or non-diamond-ramps experienced higher crash potential than on-ramps or diamond-ramps, respectively. Though the SVMs provided good model performance, the SVM model with all variables (i.e., all traffic, geometric, socio-demographic, and trip generation variables) had an overfitting problem. Therefore, it is recommended to build SVM models based on significant variables identified by other models, such as logistic regression.

1. Introduction

Traffic safety research includes wide-ranging areas, and the most prominent is assessing the safety of roadway facilities (e.g., intersections and segments) and roadway networks through crash frequency analyses, which use aggregated data from a long period, e.g., one year. The crash frequency analyses attempt to identify the factors that impact the number of crashes (Lord and Mannering, 2010), to find the potential for safety improvement (Lee et al., 2015a; Persaud et al., 1999), and to measure the safety impact of countermeasures (Gross et al., 2010; Park and Abdel-Aty, 2015).

In the recent decade, well-developed traffic management and information systems provide high-resolution traffic data, and technologies are capable to analyze and manage that data. Real-time safety studies have been focusing on the occurrence of each crash. The underlying assumptions of these studies are as follow: (1) some predictors, called crash precursors, are relatively more 'crash prone' than others are; (2) the crash potential will be significantly higher if these crash precursors are under certain conditions. To identify such crash precursors, real-time safety studies analyze traffic and environment predictors closely preceding each crash and non-crash event. Then, statistical methods are used to uncover the quantitative relationships between crash likelihoods and precursors.

The primary crash factors are traffic, environment, vehicle, and driver (Oh et al., 2001). Previous real-time safety analyses mainly focused on the impact of traffic and environment on crash occurrence. Among the traffic predictors, vehicle count in 5-min intervals, speed variance, average speed, and lane occupancy were found to be significant crash precursors (Abdel-Aty and Pande, 2005; Hossain and Muromachi, 2013a; Lee et al., 2002). Weather and geometric parameters are also crash precursors. Previous real-time studies found that hourly rainfall, visibility, and roadway surface conditions would impact crash risk (Abdel-Aty and Pemmanaboina, 2006; Wang et al., 2015b; Yu and Abdel-Aty, 2013a). Geometric parameters also play an important role in the occurrence of crashes (Shi et al., 2016; Wang et al., 2015a), e.g., shoulder width.

* Corresponding author at: Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816, USA and College of Transportation Engineering, Tongji University, Shanghai 201804, China.
E-mail address: lingwang@knights.ucf.edu (L. Wang).

http://dx.doi.org/10.1016/j.aap.2017.06.003
Received 3 December 2016; Received in revised form 4 June 2017; Accepted 5 June 2017
Available online 08 July 2017
0001-4575/

However, the driver factor has not been widely examined in real-time safety studies. The real-time safety studies usually need to analyze two types of events: crash and non-crash events. For crash events, crash reports could provide information for drivers who are involved in a traffic crash. In contrast, for non-crash events, driver information cannot be obtained from available data sources. Hence, it is difficult to directly consider driver characteristics as explanatory variables in the real-time safety studies except for some special studies, for example, naturalistic driving studies. The naturalistic driving studies would provide both driver characteristics and crash or near-crash conditions, and then these data would be used in the safety studies that directly consider driver characteristics as safety explanatory variables (Guo and Fang, 2013). On the other hand, trip generation and socio-demographic factors can reflect drivers' characteristics and behavior and can serve as the surrogate measures of drivers' characteristics and behavior. Their significant effects on traffic safety have been widely explored from a macroscopic perspective (Abdel-Aty et al., 2013; Lee et al., 2015a, 2015b; Wang et al., 2012). Nevertheless, no study has adopted trip generation and socio-demographic factors in real-time safety analyses.

Real-time safety studies have been conducted for mainlines (Abdel-Aty and Pande, 2005; Kwak and Kho, 2016; Shi et al., 2016; Xu et al., 2016), ramp vicinities (Hossain and Muromachi, 2013a, 2013b), weaving segments (Wang et al., 2015a), and ramps (Lee and Abdel-Aty, 2008). However, using trip generation and socio-demographic factors as driver behavior surrogates is not suitable for all locations but is only appropriate for some locations, e.g., ramps. For crash and non-crash events on ramps, the origins or destinations of the vehicles are likely to be at nearby zones. Hence, if the trip generation and socio-demographic information of the zone in which a ramp lies can be captured, these points of data might act as surrogates for the characteristics of the drivers on the ramp. On the other hand, for the drivers who are on mainlines, they might only drive through the zones, and the characteristics of the zones might not be related to the drivers.

To explore crash contributing factors' impact on crash likelihood, the logistic regression model has been widely used since it is capable of handling categorical target variable (Abdel-Aty and Pemmanaboina, 2006; Kwak and Kho, 2016; Washington et al., 2010). It measures the relationship between the target variable and explanatory variables based on a logistic function. The model is easy for interpretation since the model results provide the coefficient value for each significant variable. However, the logistic regression assumes that the error term has a standard logistic distribution. In reality, this assumption might not be true. On the other hand, the data mining method may not be able to provide the impact of each independent variable on the target variable, but it does not have a restriction on the distribution of the parameters. Among numerous data mining methods, Support Vector Machine (SVM) has been applied in several transportation studies (Qu et al., 2012; Sun et al., 2014; Yu and Abdel-Aty, 2013b).

The two main objectives of this study are: (1) to find out whether and which socio-demographic and trip generation factors would contribute to real-time crash risk for expressway ramps; and (2) to explore the applicability of SVM in real-time crash analysis for ramps. The paper is organized into five sections. Following this section, the second section presents the methodologies of Bayesian logistic regression model and SVM. The third section describes the data and conducts descriptive analysis of the collected variables. The fourth section shows the model results, and the fifth section summarizes the findings, conclusions, and limitations of this study.

2. Methodology

2.1. Bayesian logistic regression

This study built two Bayesian logistic regression models to explore the relationship between ramp crash likelihood and the explanatory variables. Traditional logistic regression models treat the coefficients of independent variables as fixed values. However, the Bayesian logistic regression model assumes coefficients follow a distribution. Providing full distribution of the parameters is one of the most important advantages of Bayesian over traditional models, which only provide a point estimate of a parameter and adopt an asymptotic normality assumption to describe uncertainties (Reis and Stedinger, 2005).

For any given traffic event i, it has two exclusive states: crash or non-crash. In this study, the binary responses are crash ($y_i = 1$) and non-crash ($y_i = 0$). Their possibilities are p_i ($y_i = 1$) and 1-p_i ($y_i = 0$), respectively. The y_i follows a Bernoulli distribution whose success probability is p_i:

$$y_i \sim Bernoulli(p_i) \tag{1}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{r=1}^{R} \beta_r x_{ri} \tag{2}$$

where β_0 is the intercept, β_r the coefficient of r^{th} predictors, x_{ri} the value of r^{th} explanatory variable for i^{th} event. Both β_0 and β_r were assumed to follow normal distributions:

$$\beta_0 \sim N(\mu_0, \sigma_0^2)$$

$$\beta_r \sim N(\mu_r, \sigma_r^2)$$

where N stands for normal distribution, μ_0 and μ_r are the mean of β_0 and β_r, and σ_0 and $\sigma_r$are the standard deviation of β_0 and β_r. β_0 and β_r were specified to be vague normal distributed priors: Normal (0, 10^6) (Xu et al., 2013).

The Bayesian logistic regression models were estimated in WinBugs. Three chains of 10,000 iterations were set up, and the results from the second half of the iterations were used in order to exclude the impact of the initial values and obtain stable results (Gelman et al., 2014). The convergence of the chains has been confirmed by observing stable and overlapping trace plots of the three chains.

2.2. Support vector machine

Support Vector Machine (SVM) is used for classification analysis by constructing a hyperplane or a set of hyperplanes in a high- or infinite-dimensional space (Suykens and Vandewalle, 1999). The hyperplane with the largest distance to the nearest training data point is chosen, indicating that it provides the largest separation between two types of events. There are two types of SVM: linear and nonlinear. The choice of SVM type is based on the data type, e.g., a linear SVM is better if data are linearly separated. A nonlinear SVM is achieved by applying a kernel. By introducing a kernel, SVM is flexible in the choice of the separation form and can handle nonlinear data (Deng et al., 2012).

The crash occurrence outcome y is either 1 (crash) or -1 (non-crash). Training data D is a set of n events,

$$D = \{(x_i, y_i) | x_i \in R^P, y_i \in \{-1,1\}\}_{i=1}^{n} \tag{3}$$

where x is the matrix of independent variables, and P is the number of variables. The decision function is

$$f(x) = sign(w^T x + b) \tag{4}$$

$$w = [\omega_1 \omega_{2\ldots} \omega_p]^T \tag{5}$$

A hyperplane can be written as the set of points x satisfying

$$w^T x + b = 0 \tag{6}$$

$(w^T x_i + b)$ should be positive when $y_i = 1$, and negative when $y_i = -1$. Hence, the multiple of $(w^T x_i + b)$ and y_i should always be greater than zero ($y_i(w^T x_i + b) > 0$). The decision function is using a sign-function. This results in an uncertainty of distance or margin (Campbell and Ying, 2011). Hence, two parallel hyperplanes is constructed:

$$w^T x + b = 1 \tag{7}$$

and

$$w^T x + b = -1 \tag{8}$$

The distance between these two hyperplanes is $\frac{2}{\|w\|}$. The target of SVM is to maximize the distance between the two hyperplanes by minimizing $\frac{1}{2}\|w\|^2$. In order to prevent data points from falling into the margin between two hyperplanes, the following constraint is added: for each event i either

$$w^T x_i + b \geq 1, \; if \; y_i = 1 \tag{9}$$

or

$$w^T x_i + b \leq -1, \; if \; y_i = -1 \tag{10}$$

Combing Eqs. (9) and (10), produce the following new constrain:

$$y_i(w^T x_i + b) \geq 1, \; for \;\; all \;\; i \tag{11}$$

This is a constrained optimization problem, in which $\frac{1}{2}\|w\|^2$ is minimized subject to constrain (Eq. (11)). The optimization problem can be reduced to the minimization of the following Lagrange function,

$$L(w, b) = \frac{1}{2}(w.\, w) - \sum_{i=1}^{n} \alpha_i [y_i(w.\, x_i + b) - 1] \tag{12}$$

where α_i are Lagrange multipliers, and $\alpha_i > 0$. The Eq. (12) is taken the derivatives with respect to b and w, and set these derivatives to zero:

$$\frac{\partial L(w, b)}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{13}$$

$$\frac{\partial L(w, b)}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \tag{14}$$

Substituting Eqs. (13) and (14) back into Eq. (12), the formulation is obtained,

$$\begin{aligned} W(\alpha) &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (x_i.\, x_j) - \sum_{i=1}^{n} \alpha_i [y_i(\sum_{j=1}^{n} \alpha_j y_j x_j.\, x_i + b) - 1] \\ &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (x_i.\, x_j) - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (x_i.\, x_j) - \sum_{i=1}^{n} \alpha_i y_i b \\ &\quad + \sum_{i=1}^{n} \alpha_i \\ &= \sum_{i=1}^{n} \alpha_i - \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j K(x_i.\, x_j) \end{aligned} \tag{15}$$

Subject to

$$\alpha_i \geq 0 \;\; and \;\; \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{16}$$

When linear kernel is used, $K(x_i\,.\,x_j) = (x_i\,.\,x_j)$, but when the points are not linearly classified, there is a need to conduct another kernel. In this study, the Gaussian radial basis kernel was used,

$$K(x_i.\, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \; for \;\; \gamma > 0 \tag{17}$$

where γ was 0.5. Compared to a linear kernel, the Gaussian radial basis kernel has been proven to be better in a real-time safety study by Yu and Abdel-Aty (2013b). Meanwhile, the linear kernel was tested in this study, and the results showed that the SVMs with linear kernel performed worse than the SVMs with Gaussian radial basis kernel.

3. Data preparation

Three main expressways in Central Florida were chosen, i.e., State Road 408, State Road 417, and State Road 528. Along these three expressways, there were 141 ramps, including on- and off-ramps, which were selected as the study subjects. The study period was from July 2013 to March 2014. Five datasets were collected: crash, traffic, geometric, trip generation, and socio-demographic data.

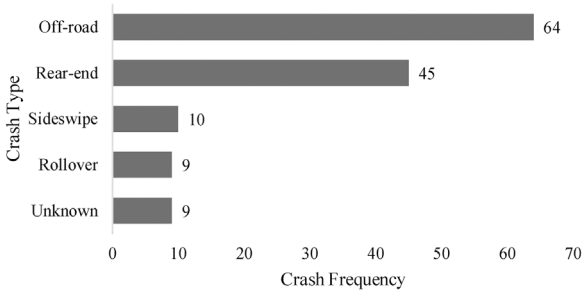

Fig. 1. Crash Types.

The crash data were from the Signal Four Analytics (S4A). S4A is an interactive, web-based system designed to support the crash mapping and analysis needs in the state of Florida. It is developed by the University of Florida and funded by the state of Florida. The crash data were collected by Florida Highway Patrol officers at crash sites throughout the state, and then are loaded into the Signal Four Analytics database (University of Florida, 2017). For each recorded crash event, S4A provides crash time, coordinate, type, severity, etc. There were 137 crashes that happened on the studied ramps in the study period, among which, 79 were single-vehicle (SV) crashes and 58 were multi-vehicles (MV) crashes. The crash type of six SV crashes and three MV crashes was unknown. The detailed crash type information is shown in Fig. 1.

Fig. 1 shows that the off-road crash (46.7%) was the major crash type for expressway ramps, and the next most frequent crash type was rear-end (32.8%). Similar result has also been found by other researchers (McCartt et al., 2004). On the other hand, the leading crash type on expressway mainlines is rear-end (Shi et al., 2016). Compared with mainlines, ramps are usually with sharp horizontal or vertical curve or both. This might be the main reason for the difference in crash types.

The traffic data were supplied by the Microwave Vehicle Detection System (MVDS) that is operated by the Central Florida Expressway Authority (CFX). MVDS detectors were set near the expressway gore area, between the expressway mainline and a ramp, which merges into or diverges from the mainline. The detectors record vehicle counts, average time mean speed, and lane occupancy every one minute for each lane. In order to alleviate the impact of temporal turbulences, the traffic data were aggregated at a 5-min interval for each ramp. First, if a ramp has more than one lane, several lanes' traffic data were aggregated at 1-min interval, including total vehicle counts, mean speed, and mean lane occupancy, among which, the mean speed and mean lane-occupancy were weighted by vehicle count, respectively. Then, the 1-min interval based data were aggregated to obtain traffic variables at 5-min intervals: total vehicle count, mean speed, mean lane occupancy, and the standard deviation of speed.

In addition to crash and traffic data, ramp geometric characteristics data were collected. There were two data sources: the Roadway Characteristic Inventory (RCI) and ArcGIS. The shoulder widths (left and right shoulder width) and ramp length information were obtained from RCI, which is maintained by the Florida Department of Transpiration (FDOT). Ramp type (on- and off-ramp), ramp configuration (diamond- and non-diamond-ramp), and the presence of a toll-booth were manually gathered by viewing ArcGIS map. For the studied ramps, there are mainly six ramp configurations (shown in Fig. 2). However, in this study, more than half of the ramps (58%) were diamond-ramps, and there were limited sample size for each of the other ramp configurations. Hence, the ramp configuration variable was set as a binary variable: diamond- and non-diamond-ramp.

The studied ramps exist in 69 Statewide Traffic Analysis Zones (SWTAZs). FDOT Central Office developed statewide planning data, network, and model based on SWTAZs. The data includes trip

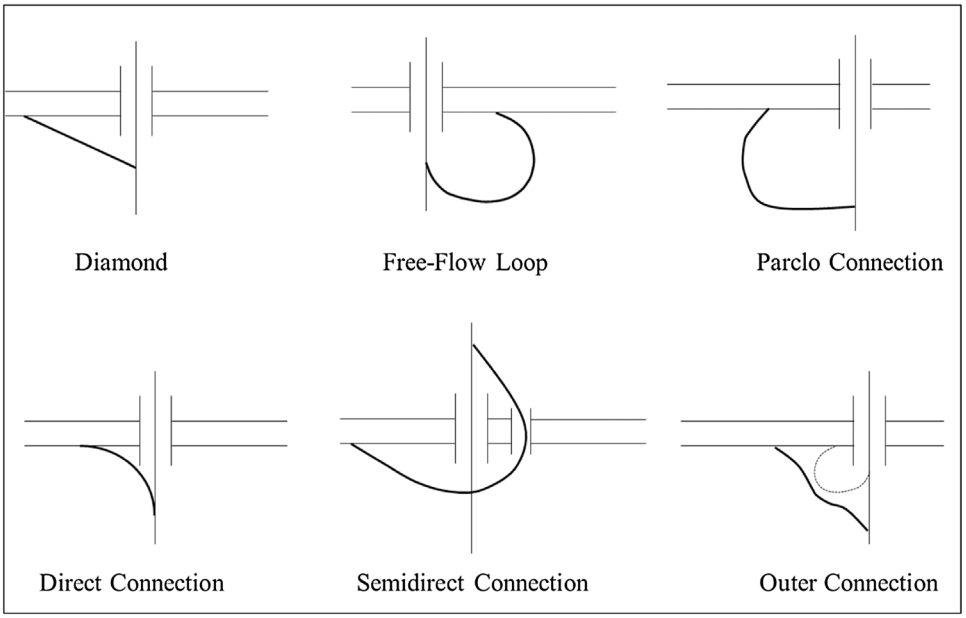

Fig. 2. Ramp configurations.

generation variables, which include trip productions and trip attractions. The production and attraction trips are provided by trip purposes (i.e., working, social or recreational, shopping, and total). The trip generation data were processed and converted to percentages by trip purposes. Furthermore, the SWTAZ model also provided socio-demographic variables including population density, employment density, and the number of employees by different industry type. The socio-demographic parameters were processed to provide percentage for each employee type.

There were 137 crashes in total; however, 16 crashes did not have their corresponding traffic data. Hence, 121 crashes with complete traffic, geometric, socio-demographic and trip generation information were used in the model estimation. The traffic conditions that were 5–10 min before the crash time were chosen to represent the disturbance condition that contributes to the occurrence of a crash (Hossain and Muromachi, 2013a; Sun and Sun, 2016; Yu and Abdel-Aty, 2013a). For example, if a crash occurs at 8:00 A.M. on a ramp, the disturbance condition is from 7:50 to 7:55 A.M. of the same day on that ramp. The non-crash dataset was chosen from normal traffic conditions, which neither result in nor are impacted by any crashes. In this study, normal traffic conditions were more than two hours from a crash event, because crash clearance time is usually less than two hours (Moore et al., 2004). To be consistent with the crash events, non-crash events were also at 5-min intervals.

The non-crash dataset consisted of more than ten million events. It was not practical to use the entire non-crash dataset. Hence, this study adopted a case-control design: for each crash event in the crash population, a sample of ten non-crash events was randomly selected from the non-crash dataset. The case-control design has been widely used in real-time safety analyses (Caleffi et al., 2017; Hossain and Muromachi, 2012; Kwak and Kho, 2016; Wang et al., 2015a,b,c,d), because it could obtain valid coefficients of the independent variables for logistic regression models (Hosmer et al., 2013). The descriptive analysis of the variables of the modeling data is shown in Table 1.

4. Model estimation and results

This section first estimates two Bayesian logistic regression models to identify the significant variables and then applies SVMs in crash prediction. In order to find whether socio-demographic and trip generation variables have significant impacts on model performance, the first Bayesian logistic regression model considered these two types of variables, but the second Bayesian logistic regression model only considered traffic and geometric variables.

The whole dataset was randomly split into training and validation datasets with a ratio of 70:30. The model performance of the logistic regression models were measured by the deviance information criterion (DIC) and the area under the Receiver Operating Characteristic (ROC) curve. DIC is considered as a Bayesian measure of goodness-of-fit which is penalized by a complexity term (Spiegelhalter et al., 2002). A lower DIC indicates a better model. ROC is a standard for evaluating a model's ability of correctly predict a categorical event (Hosmer et al., 2013). The range of ROC is from 0.5 to 1.0, and a higher ROC is preferred.

Before the estimation of the Bayesian logistic regression models, the significance of each independent variable was tested by putting this variable and dependent variable (crash occurrence) in a logistic regression model. If this variable had a significant impact on crash occurrence at the 85% confidence interval (CI), the variable was kept; otherwise, the variable was excluded from the Bayesian logistic regression modeling. Then, in order to prevent high correlation between variables, the Pearson correlation test was done for the significant variables. If the absolute value of the correlation coefficient of two parameters was higher than 0.3, only the variable which resulted in a lower DIC was kept in the model. The logistic regression models' results are shown in Table 2. All the training and validation ROCs of the Bayesian logistic regression models were higher than 0.8. It indicates the models had a good ability to distinguish crash and non-crash events (Hosmer et al., 2013).

Only one of the socio-demographic and trip generation parameters was significant in the model. It was the percentage of home-based-work production (P_HBWP). This phenomenon might be because of the limited variation in zonal characteristics since the studied ramps were all from Central Florida. Table 2 shows that the standard deviations of some socio-demographic and trip generation parameters were very small compared to their mean value: for example, the coefficient of variation of the percentage of transportation employment was only 19%.

The model with the variable "percentage of Home-based-work production" was significantly better than the model without this variable. The DIC difference between Model 1 and 2 was more than five, which indicates the model with a lower DIC (Model 1) is substantially better than the other model (Model 2) (Bolker, 2008). The percentage of home-based-work production is positively related to crash risk. There might be two reasons for this positive relationship. First, drivers who

Table 1
Descriptive analysis for real-time ramp analysis.

Variables	Description	Mean	Std.	Min	Max
Traffic Parameters					
Vehcnt	Vehicle count in 5-min intervals (veh/5 min)	18.4	20.7	1	171
Speed	Average speed in 5-min intervals (mph)	52.7	9.1	3.8	103.6
Std_spd	Standard deviation of speed in 5-min intervals (mph)	4.1	3.2	0	34.0
Occupancy	Average lane occupancy in 5-min intervals (%)	2.5	3.7	0	47.0
Geometric Parameters					
Length	Ramp length (mile)	0.32	0.24	0.07	1.66
Sldwth_R	Right shoulder width (feet)	1.9	1.9	1.0	6.0
Sldwth_L	Left shoulder width (feet)	4.3	2.9	1.0	12.0
Type	1 = if the ramp is an off-ramp; 0 = otherwise	0.46	0.50	0	1
Configuration	1 = if the ramp is a diamond-ramp; 0 = otherwise	0.58	0.49	0	1
Toll	1 = if there is a toll booth on the ramp; 0 = otherwise	0.29	0.46	0	1
Trip Generation Parameters					
Production	Total productions (trips/day)	5601	5,910	84	25,010
Attraction	Total attractions (trips/day)	5666	7,663	20	33,742
P_HBWA	Percentage of Home-based-work attractions (%)	16.4	9.3	0	74.8
P_HBWP	Percentage of Home-based-work productions (%)	14.8	7.1	0	27.6
P_HBSRA	Percentage of Home-based-social-recreational attractions (%)	8.4	3.2	3.2	19.1
P_HBSRP	Percentage of Home-based-social-recreational productions (%)	7.1	4.2	1.4	31.0
P_HBSHA	Percentage of Home-based-shopping attractions (%)	9.3	7.4	0	27.2
P_HBSHP	Percentage of Home-based-shopping productions (%)	15.8	5.8	3.9	25.0
Socio-demographic Parameters					
Pop_density	Population density (people/square mile)	2215	2,038	0	10,312
Emp_density	Employment density (people/square mile)	1577	2,633	0	13,295
Enr_density	Enrollment density (people/square mile)	902	2607	0	14,945
P_agriculture	Percentage of agriculture employment (%)	1.3	0.3	0	2.2
P_service	Percentage of service employment (%)	50.0	9.5	25.0	66.7
P_construction	Percentage of construction employment (%)	3.0	2.2	0	10.0
P_manufacturing	Percentage of manufacturing employment (%)	2.7	2.1	0	8.3
P_wholesale	Percentage of wholesale employment (%)	3.1	2.3	0	10.0
P_retail	Percentage of retail employment (%)	19.3	10.4	0	48.8
P_financial	Percentage of financial employment (%)	6.7	1.3	3.3	9.5
P_public	Percentage of public administration employment (%)	8.5	1.5	5.0	11.1
P_transportation	Percentage of transportation employment (%)	5.3	1.0	2.5	7.0

travel from home to work have to arrive at destinations on time. They may want to avoid being late and may rush to get to work. Thus, they might be more susceptible to errors. Second, drivers may be fatigued after a whole day of work, so the crash potential of work-to-home trip may be higher than other trips.

The logarithm of vehicle count in 5-min intervals was positively related to ramp crash likelihood, indicating that high traffic count resulted in high crash risk on a ramp. Traffic volume is one of the most common exposure variables in previous traffic safety analyses, and there is a significant positive relationship between traffic volume and crash count or real-time crash potential has been widely found by researchers (Caleffi et al., 2017; Yu and Abdel-Aty, 2013a). Additionally, many previous studies indicated that the crash frequency and the crash risk were not linearly related to volume, hence, a logarithmic transformation was applied (Shi and Abdel-Aty, 2015; Wang et al., 2013; Wang et al., 2015a,b,c,d). Speed has a positive impact on crash occurrence. A higher speed means longer braking and reaction distance; hence, a vehicle travelling at a higher speed would be more likely involved in a collision with other objects (fixed object in SV and other vehicles in MV).

Two geometric factors, ramp type and configuration, are significant in the model. Off-ramps experience more crashes than on-ramps, and

Table 2
Bayesian logistic regression models.

Variables	With socio-demographic and trip generation parameters (Model 1)			Without socio-demographic and trip generation parameters (Model 2)		
	Mean	Std.	95% BCI[a]	Mean	Std.	95% BCI
Intercept	−7.469	1.064	(−9.618, −5.447)	−6.759	1.097	(−8.885, −4.673)
Log(Vehcnt)	0.824	0.140	(0.553,1.089)	0.888	0.148	(0.610,1.180)
Speed	0.044	0.016	(0.012,0.074)	0.044	0.016	(0.013,0.075)
Configuration	−1.233	0.275	(−1.786, −0.704)	−1.153	0.266	(−1.683, −0.644)
Type	0.528	0.277	(0.005,1.087)	0.524	0.278	(−0.033,1.025)[b]
P_HBWP	5.863	2.073	(1.694,9.991)			
			Model Performance			
DIC			438.492			445.702
Training ROC			0.813			0.801
Validation ROC			0.815			0.812

[a] Bayesian credible interval.
[b] Significant at the 90% BCI.

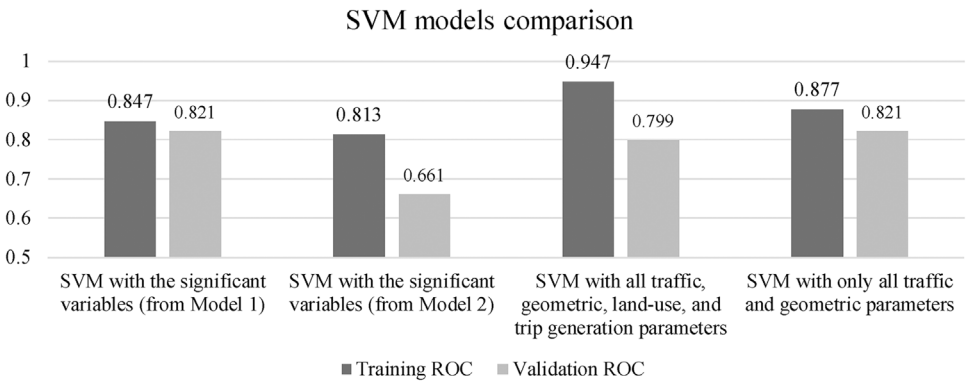

Fig. 3. SVM models comparison.

the models' results show that the crash odds for off-ramps is about 1.70 times of that for on-ramps. A previous study by McCartt et al. (2004) also found that the crash rate of off-ramps was much higher than that of on-ramps. Vehicles on off-ramps need to decelerate to adjust to the lower speed limits on ramps; meanwhile, they have to decrease speed in order to prepare to brake or even stop at a cross-street intersection. If a following vehicle does not react or decelerate in time, it will collide with the vehicle ahead. This inference is supported by the field data: the main crash type for off-ramps was rear-end crashes (50.0%).

The models show that diamond-ramps had significantly lower crash risk than non-diamond-ramps. A previous off-ramp safety study by Chen et al. (2013) also found diamond-ramps were safer than non-diamond-ramps. Non-diamond-ramps have smaller turning radii, which might lead to a loss of vehicle control and result in crashes. In this study, the off-road and the rollover crash percentage for non-diamond-ramp was 67.6%, but that percentage for diamond-ramps was only 20.0%.

In addition to the Bayesian logistic regression models, SVMs with Gaussian radial basis kernel were tested using the same training and validation datasets as the logistic regression models. The model results are presented in Fig. 3.

The SVM with the significant variables from Model 1 was better than the SVM with the significant variables from Model 2. The only difference was because of the trip generation variable, the percentage of home-based-work production. This result indicates the substantial importance of the trip generation variable, and it is consistent with the finding of the two Bayesian logistic regression models.

The ROCs of the SVM with only all traffic and geometric parameters were lower than that of the SVM with all variables (traffic, geometric, socio-demographic, and trip generation parameters), except for the validation ROC of the SVM with all variables. The main reason for this exception was the great number of the socio-demographic and trip generation variables resulting in an overfitting issue for the SVM. Too many independent variables may cause an SVM model to "memorize" training data instead of finding the underlying relationship between dependent and independent variables. The similar phenomenon was also found by other researchers (Yu and Abdel-Aty, 2013b).

5. Conclusion

Several real-time safety studies have been carried out for mainlines, ramp vicinities, weaving segments, and ramps. These studies found that traffic and environmental factors have significant impact on crash occurrence. However, no study has been conducted to analyze the effects of socio-demographic and trip generation parameters on crash risk. This study explored real-time crash risk for expressway ramps using traffic, geometric, socio-demographic, and trip generation predictors.

Two Bayesian logistic regression models were utilized to find the variables that effected ramp crash risk: one considered socio-demographic and trip generation parameters and the other did not. The results showed that the first model outperformed the second model, and it indicated that the socio-demographic and trip generation parameters would help in improving the estimation accuracy for ramp crash risk. The models identified that the logarithm of traffic count and speed had positive impacts on crash risk, and off-ramps and non-diamond-ramps had higher crash risk. As for the trip generation and socio-demographic parameters, the percentage of home-based-work production compared to other parameters was found to have a positive impact on crash risk.

Subsequently, four SVM models were applied to predict ramp crash occurrence. It was found that the SVMs with the socio-demographic and trip generation parameters were generally better than the SVMs that only considered the traffic and geometric parameters. However, the SVM with all variables had an overfitting issue as it provided a very high training ROC but a significantly lower validation ROC. Therefore, instead of using all collected variables, it would be better to build SVMs based on significant variables identified by other models, such as logistic regression.

There are limitations to this study. Since the studied ramps were from three expressways in Central Florida, the variations in zonal characteristics were limited. Hence, the models ended with a limited number of significant socio-demographic and trip generation predictors. The follow-up study should extend the study area in order to increase SWTAZs and improve the variation of the zonal characteristics. Thus, the effect of trip generation and socio-demographic elements can be better interpreted. Additionally, more years' data could be added in future, then, the final model might be more precise since increasing the sample size would further decrease the standard deviation of the coefficient of parameters.

Acknowledgements

The authors thank the Southeastern Transportation Center UTC consortium for funding of this research. The authors also thank the Central Florida Expressway Authority and Florida Department of Transportation for providing data.

References

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Saf. Res. 36, 97–108.
Abdel-Aty, M., Pemmanaboina, R., 2006. Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data. IEEE Trans. Intell. Transp. Syst. 7, 167–174.
Abdel-Aty, M., Lee, J., Siddiqui, C., Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. Transp. Res. Part A Policy Pract. 49, 62–75.
Bolker, B.M., 2008. Ecological Models and Data in R. Princeton University Press.
Caleffi, F., Anzanello, M., Cybis, H., 2017. A multivariate-based conflict prediction model for a Brazilian freeway. Accid. Anal. Prev. 98, 295–302.
Campbell, C., Ying, Y., 2011. Synthesis lectures on artificial intelligence and machine learning. Learning with Support Vector Machines 5. pp. 1–95 (1).
Chen, H., Lee, C., Lin, P.-S., 2013. Motorcycle safety investigation at exit ramp section from crash data and rider's perception. In: Transportation Research Board 92nd Annual Meeting. Washington, D.C.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2014. Bayesian Data Analysis. Taylor & Francis.

Gross, F., Persaud, B., Lyon, C., 2010. A Guide to Developing Quality Crash Modification Factors. FHWA-SA-10-032.
Guo, F., Fang, Y., 2013. Individual driver risk assessment using naturalistic driving data. Accid. Anal. Prev. 61, 3–9.
Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression. John Wiley & Sons.
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.
Kwak, H.C.H., Kho, S., 2016. Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data. Accid. Anal. Prev. 88, 9–19.
Lee, C., Abdel-Aty, M., 2008. Two-level nested logit model to identify traffic flow parameters affecting crash occurrence on freeway ramps. Transp. Res. Rec. J. Transp. Res. Board 145–152.
Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. Transp. Res. Rec. J. Transp. Res. Board 1–8.
Lee, J., Abdel-Aty, M., Choi, K., Huang, H., 2015a. Multi-level hot zone identification for pedestrian safety. Accid. Anal. Prev. 76, 64–73.
Lee, J., Abdel-Aty, M., Jiang, X., 2015b. Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accid. Anal. Prev. 78, 146–154.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44, 291–305.
McCartt, A.T., Northrup, V.S., Retting, R.A., 2004. Types and characteristics of ramp-related motor vehicle crashes on urban interstate roadways in Northern Virginia. J. Saf. Res. 35, 107–114.
Moore, J.E., Giuliano, G., Cho, S., 2004. Secondary accident rates on Los Angeles freeways. J. Transp. Eng. 130, 280–285.
Oh, C., Oh, J.-S., Ritchie, S.G., Chang, M., 2001. Real-time estimation of freeway accident likelihood. In: Transportation Research Board 80th Annual Meeting. Washington, D.C.
Park, J., Abdel-Aty, M., 2015. Development of adjustment functions to assess combined safety effects of multiple treatments on rural two-lane roadways. Accid. Anal. Prev. 75, 310–319.
Persaud, B., Lyon, C., Nguyen, T., 1999. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transp. Res. Rec. J. Transp. Res. Board 7–12.
Qu, X., Wang, W., Wang, W., Liu, P., Noyce, D.A., 2012. Real-time prediction of freeway rear-end crash potential by support vector machine. In: Transportation Research Board 91st Annual Meeting. Washington, D.C.
Reis, D.S., Stedinger, J.R., 2005. Bayesian MCMC flood frequency analysis with historical information. J. Hydrol. 313, 97–116.
Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C Emerg. Technol. 58, 380–394.
Shi, Q., Abdel-Aty, M., Lee, J., 2016. A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. Accid. Anal. Prev. 88, 124–137.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 64, 583–639.
Sun, J., Sun, J., 2016. Real-time crash prediction on urban expressways: identification of key variables and a hybrid support vector machine model. IET Intell. Transp. Syst. 10, 331–337.
Sun, J., Sun, J., Chen, P., 2014. Use of support vector machine models for real-time prediction of crash risk on urban expressways. Transp. Res. Rec. J. Transp. Res. Board 91–98.
Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9, 293–300.
University of Florida, 2017. About Signal Four Analytics. https://s4.geoplan.ufl.edu/. (Accessed 3 June 2017).
Wang, X.S., Jin, Y., Abdel-Aty, M., Tremont, P.J., Chen, X.H., 2012. Macrolevel model development for safety assessment of road network structures. Transp. Res. Rec. J. Transp. Res. Board 100–109.
Wang, C., Quddus, M., Ison, S., 2013. A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. Transp. A Transp. Sci. 9, 124–148.
Wang, J., Ph, D., Wang, M., 2015a. Analysis on sideswipe collision precursors considering the spatial-temporal characters of freeway traffic. J. Transp. Eng. 142.
Wang, L., Abdel-Aty, M., Shi, Q., Park, J., 2015b. Real-time crash prediction for expressway weaving segments. Transp. Res. Part C Emerg. Technol. 61, 1–10.
Wang, L., Shi, Q., Abdel-Aty, M., 2015c. Predicting crashes on expressway ramps with real-time traffic and weather data. Transp. Res. Rec. J. Transp. Res. Board 2514, 32–38.
Wang, X., Fan, T., Chen, M., Deng, B., Wu, B., Tremont, P., 2015d. Safety modeling of urban arterials in Shanghai, China. Accid. Anal. Prev. 83, 57–66.
Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2010. Statistical and Econometric Methods for Transportation Data Analysis. CRC press.
Xu, C., Wang, W., Liu, P., 2013. Identifying crash-prone traffic conditions under different weather on freeways. J. Saf. Res. 46, 135–144.
Xu, C., Liu, P., Wang, W., 2016. Evaluation of the predictability of real-time crash risk models. Accid. Anal. Prev. 94, 207–215.
Yu, R., Abdel-Aty, M., 2013a. Multi-level Bayesian analyses for single-and multi-vehicle freeway crashes. Accid. Anal. Prev. 58, 97–105.
Yu, R., Abdel-Aty, M., 2013b. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.