Accident Analysis and Prevention 57 (2013) 30–39

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Predicting crash likelihood and severity on freeways with real-time loop detector data

Chengcheng Xu[a,*], Andrew P. Tarko[b], Wei Wang[a], Pan Liu[a]

[a] School of Transportation, Southeast University, Si Pai Lou #2, Nanjing, 210096, China
[b] Center for Road Safety, School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907, United States

ARTICLE INFO

Article history:
Received 27 August 2012
Received in revised form 4 March 2013
Accepted 31 March 2013

Keywords:
Crash severity
Real-time safety management
Crash risk prediction
Sequential logit model
Freeway

ABSTRACT

Real-time crash risk prediction using traffic data collected from loop detector stations is useful in dynamic safety management systems aimed at improving traffic safety through application of proactive safety countermeasures. The major drawback of most of the existing studies is that they focus on the crash risk without consideration of crash severity. This paper presents an effort to develop a model that predicts the crash likelihood at different levels of severity with a particular focus on severe crashes. The crash data and traffic data used in this study were collected on the I-880 freeway in California, United States. This study considers three levels of crash severity: 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 derived from detector data. The elasticity analysis was conducted to evaluate the effect of the traffic flow variables on the likelihood of crash and its severity.The results show that the traffic flow characteristics contributing to crash likelihood were quite different at different levels of severity. The PDO crashes were more likely to occur under congested traffic flow conditions with highly variable speed and frequent lane changes, while the KA and BC crashes were more likely to occur under less congested traffic flow conditions. High speed, coupled with a large speed difference between adjacent lanes under uncongested traffic conditions, was found to increase the likelihood of severe crashes (KA). This study applied the 20-fold cross-validation method to estimate the prediction performance of the developed models. The validation results show that the model's crash prediction performance at each severity level was satisfactory. The findings of this study can be used to predict the probabilities of crash at different severity levels, which is valuable knowledge in the pursuit of reducing the risk of severe crashes through the use of dynamic safety management systems on freeways.

1. Introduction

Real-time crash risk prediction models estimate the likelihood of crash occurrence for a given freeway segment over a short time period, such as 5 min. One of the important practical applications of real-time crash risk prediction models is identification of hazardous traffic conditions that may lead to a crash. Predicting the crash risk in real-time is an essential task in freeway dynamic safety management systems. Crash risk prediction helps identify hazardous traffic conditions where proactive crash prevention strategies are needed to mitigate the high crash risk. In recent years, numerous studies have developed freeway crash risk prediction models that link the crash risk with certain traffic flow characteristics measured with freeway traffic surveillance systems (Oh et al., 2001, 2005; Lee et al., 2003; Abdel-Aty et al., 2004, 2005; Zheng et al., 2010; Pande et al., 2011; Ahmed et al., 2012; Ahmed and Abdel-Aty, 2012; Xu et al., 2012a,b, in press; Li et al., 2012).

Oh et al. (2001, 2005) applied a Bayesian model to establish the statistical relationship between the crash risk and the real-time traffic flow states. The results showed that the standard deviation of speed estimated in five-minute intervals was a good indicator of hazardous traffic conditions where the crash potential was considerably higher than under other traffic conditions. Lee et al. (2003) used a log-linear model to estimate crash risks based on real-time traffic flow data collected from freeway loop detector stations. It was concluded that the coefficient of variation in speed, traffic density, and speed difference between upstream and downstream loop detector stations were significantly correlated with the crash risk.

Abdel-Aty et al. (2004) applied matched case-control logistic regression to link crash likelihood with real-time traffic flow characteristics. The traffic intervals preceding a crash were cases that were matched with the crash-free intervals used as controls. The results showed that the likelihood of crash occurrence was

* Corresponding author. Tel.: +86 13801580045.
E-mail addresses: iamxcc1@gmail.com (C. Xu), tarko@purdue.edu (A.P. Tarko), wangwei@seu.edu.cn (W. Wang), liupan@seu.edu.cn (P. Liu).

0001-4575/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2013.03.035

correlated with the average detector occupancy at the upstream loop detector station and with the coefficient of variation in speed observed at a downstream loop detector station. In a subsequent study, Abdel-Aty et al. (2005) developed real-time crash risk prediction models under high speed and low speed traffic conditions based on the matched case-control logistic regression. It was found that the mechanisms of multi-vehicle crashes were different in these two speed regimes. Abdel-Aty and Pande (2005) applied the probabilistic neural network (PNN) model to predict crash occurrences on freeways using multiple speed derivatives, which included the logarithms of the coefficient of the variation in speed. Pande and Abdel-Aty (2006) developed a crash risk prediction model based on the classification tree and neural network to identify hazardous traffic conditions potentially leading to lane-change-related collisions. The results indicated that the average speed, the difference in occupancy between adjacent lanes, and standard deviation of speed and volume contributed to lane-change-related crash risk.

Recently, Zheng et al. (2010) used a matched case-control logistic regression model to evaluate the impacts of the speed variance resulting from the oscillating traffic state on the likelihood of crash occurrence using case-controlled data. Hossain and Muromachi (2011) developed separate crash risk prediction models for basic freeway segments and ramp vicinities and found that the contributing factors to crash risk were quite different for the two areas. The mean and standard deviation of the difference in traffic flow parameters between adjacent lanes were the main contributing factors to high crash risks on the basic freeway segments while the high ramp flow and the variation in speed between downstream and upstream detector stations affected the crash risk within the ramp vicinities. Xu et al. (2012a) used a K-means clustering analysis to classify traffic flow states and to test the connection between these traffic states and the crash risks on freeways. Crash risk prediction models were then developed for different traffic states, and the results demonstrated that the impacts of traffic flow characteristics on crash risks were different across different traffic states.

Most of the real-time crash risk prediction models were developed using traffic data collected with loop detector stations. In later studies, traffic data collected with other surveillance technologies were used to develop crash risk prediction models. Hourdos et al. (2006) applied the binary logit model to identify crash prone conditions on freeways using traffic data captured with video cameras. Several traffic flow characteristics contributing to crash likelihood were identified, such as large speed differences between adjacent lanes and compression waves leading to abrupt changes in traffic flow. Ahmed and Abdel-Aty (2012) applied matched case-control logistic regression to develop a real-time crash risk prediction model using traffic data collected from the tag readers on toll roads known as Automatic Vehicle Identification (AVI) Systems. In a following study, Ahmed et al. (2012) applied the Bayesian semi-parametric Cox proportional hazards model to develop a real-time crash risk prediction model based on traffic data collected from an AVI system. The results demonstrated that the likelihood of crash occurrences on freeways was affected by the average speed and standard deviation of speed measured by the AVI system.

Crash severity is usually defined as the most serious injury among individuals involved in the crash. The severity of a crash can range from low-cost property damage to extremely costly severe injury and fatality. It is important to be able to predict the likelihood of crashes at various levels of severity to focus safety management systems on proactive prevention of severe injuries. The current most common approach is to consider the real-time risk of crash without distinguishing between different levels of severity. The first attempt to estimate the risk of a severe crash by using traffic flow characteristics measured with loop detectors was made by Golob et al. (2008). He developed a binary logit model of the probability of non-PDO crash conditional on crash occurrence, but the model cannot be used directly to predict the occurrences of severe crashes. Golob's study utilized data collected with a single detector station so the spatial differences in traffic flow states could not be included in the model. We were unable to find other studies aimed at developing real-time prediction models for crashes at various levels of severity.

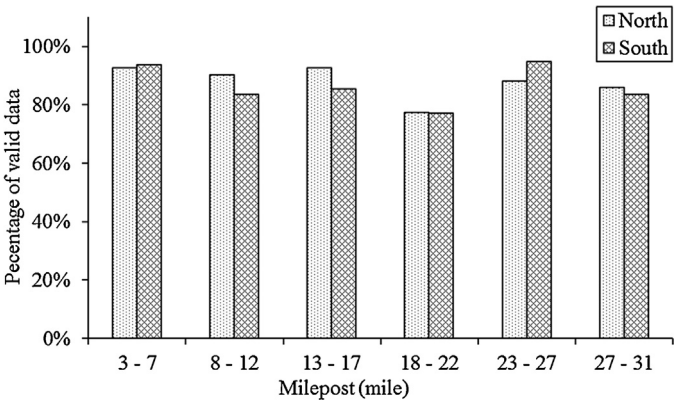

Fig. 1. Valid rate of traffic data along the I-880 Freeway in 2008.

The primary objective of this study is to explore the possibility of developing crash risk models for distinct levels of severity that are implementable in real-time freeway safety management systems. The practicality of such models is important, and this aspect will be addressed by rigorous testing of the predictive ability of the estimated models and by critical discussion of the models' performance from the standpoint of false alarms and their eroding effect on drivers' response to warning messages. The research results will promote a better understanding of the impact of traffic flow characteristics on the likelihood of severe crashes and will help transportation professionals develop effective crash prevention strategies that focus on consequential crash events.

2. Data sources

To accomplish the research objective, data were obtained from a 29-mile segment on the I-880 freeway in the San Francisco Bay area of California in the United States. There are 119 loop detectors stations in the northbound and southbound directions along the selected freeway section with an average spacing of 0.5 miles. The standard deviation of the spacing is around 0.3 mile. The minimum and maximum spacing are 0.15 and 1.68, respectively. A total of 5 weather stations are located along the selected freeway section. All the 5 weather stations are located within about 5 miles from the I-880N freeway. The collected crash, weather and traffic data cover the entire 2008 period. A total of 794 crashes were identified and used in the study.

The traffic data were obtained from the Highway Performance Measurement System (PeMS) maintained by the California Department of Transportation (Caltrans). Fig. 1 illustrates the percentages of valid traffic data along the I-880 freeway in 2008. As shown in Fig. 1, traffic data has reasonable valid rate on the selected freeway segment and the percentage of valid data is generally around 90%. The average speed, volume, and occupancy in 30-s aggregation intervals were collected in each lane. Traffic data were excluded as invalid or not usable under one or more of the following conditions: (1) the average occupancy was greater than 100%; (2) the average speed was greater than 0 mph while the flow rate was 0 vph; (3) the flow rate was greater than 0 vph while the occupancy was 0%; (4) the average speed was greater than 100 mile; or (5) the occupancy was greater than 0% while the flow rate was

Table 1
Variables considered for the models.

Symbol	Variables
$VehCnt_u$	Average 30-s vehicle count at the upstream station (veh/30 s)
$DetOcc_u$	Average 30-s detector occupancy at the upstream station (%)
$AvgSpd_u$	Average 30-s speed at the upstream station (mile/h)
$OccDev_u$	Std. dev. of 30-s detector occupancies at the upstream station (%)
$SpdDev_u$	Std. dev. of 30-s mean speeds at the upstream station (mile/h)
$CvSpd_u$	Coefficient of variation of 30-s mean speeds at the upstream station (mile/h)
$OccDif_u$	Average absolute difference in 30-s detector occupancies between adjacent lanes at the upstream station (%)
$SpdDif_u$	Average absolute difference in 30-s mean speeds between adjacent lanes at the upstream station (mile/h)
$VehCnt_d$	Average 30-s vehicle counts at the downstream station (veh/30 s)
$DetOcc_d$	Average 30-s detector occupancy at the downstream station (%)
$AvgSpd_d$	Average 30-s speed at the downstream station (mile/h)
$OccDev_d$	Std. dev. of 30-s detector occupancies at the downstream station (%)
$SpdDev_d$	Std. dev. of 30-s mean speeds at the downstream station (mile/h)
$CvSpd_d$	Coefficient of variation of 30-s mean speeds at the downstream station (mile/h)
$OccDif_d$	Average absolute difference in 30-s detector occupancies between adjacent lanes at the downstream station (%)
$SpdDif_d$	Average absolute difference in 30-s mean speeds between adjacent lanes at the downstream station (mile/h)
$AvgCnt_{u-d}$	Average absolute difference in vehicle counts between upstream and downstream stations (veh/30 s)
$AvgOcc_{u-d}$	Average absolute difference in detector occupancies between upstream and downstream stations (%)
$AvgSpd_{u-d}$	Average absolute difference in speeds between upstream and downstream stations (mile/h)
$DevCnt_{u-d}$	Std. dev. of absolute difference in vehicle counts between upstream and downstream stations (veh/30 s)
$DevOcc_{u-d}$	Std. dev. of absolute difference in detector occupancies between upstream and downstream stations (%)
$DevSpd_{u-d}$	Std. dev. of absolute difference in speeds between upstream and downstream stations (mile/h)
$DetDist_{u-d}$	Distance between upstream and downstream stations (mile)
$Width_s$	Road surface width (ft)
$Width_o$	1 = if outer shoulder width > 10 ft; 0 = otherwise
$Width_i$	1 = if inner shoulder width > 10 ft; 0 = otherwise
$Width_m$	Inner median width (ft)
Lanes	Number of lanes
On-ramp	1 = if there is an on-ramp between upstream and downstream stations; 0 = otherwise
Off-ramp	1 = if there is an off-ramp between upstream and downstream stations; 0 = otherwise
Curve	1 = curve section; 0 = otherwise
Peak	1 = peak period; 0 = otherwise
Weather	1 = adverse weather conditions (rain or fog); 0 = otherwise

equal to 0 vph. The 30-s raw detector readings from two consecutive upstream–downstream detector stations were aggregated into 5-min intervals and converted into the 22 traffic flow variables presented in Table 1.

The traffic flow variables in Table 1 consist of 5-min observations supplemented with a crash indicator (1 if a crash occurred between the upstream and downstream detectors, and 0 otherwise). The researchers extracted traffic data in the time interval between 5 and 10 min prior to crash occurrence. The purpose of doing so was to identify hazardous traffic condition ahead of the crash occurrence time to make preemptive measures possible (Pande et al., 2011; Lee et al., 2011; Xu et al., in press). For example, if a crash occurred at 10:00 pm, the traffic data were extracted from 9:50 to 9:55 pm. This time lag was also adopted in previous studies to develop real-time crash risk prediction models (Pande et al., 2011; Lee et al., 2011; Xu et al., in press). For each crash in the dataset, the researchers randomly selected 20 five-minute intervals without crashes from the crash-free days. These intervals were supplemented with the 22 traffic flow variables to form crash-free observations. To generate the dataset of non-crash cases, the time for each non-crash case was randomly chosen from the 527,040 1-min intervals in 2008 (60 min × 24 h × 366 days in 2008). Similarly, the upstream and downstream station for each non-crash case was randomly selected from the 119 loop detector stations. Then, each randomly selected combination of time and stations was used to extract 5-min detector data after the assigned time of the non-crash from the assigned upstream and downstream station of the non-crash. In addition, it was ensured that there were no crashes observed at the location of each non-crash case during the whole day. Each non-crash case was also assigned a random milepost location based on its upstream and downstream station. Figs. 2 and 3 illustrate the distributions of non-crash cases over time and space. The dataset of non-crash cases covers the normal traffic conditions at different time along the selected I-880 freeway section. Thus, the selected non-crash cases could generally represent the normal traffic flow conditions.

The geometric data for the I-880 freeway were also obtained from the PeMS database. As shown in Table 1, 9 geometric variables were used in this study. The geometric data for each crash case and non-crash case were extracted based on their milepost location. The weather data were obtained from National Climate Data Center (NCDC) website which provides hourly weather information from weather stations across the United States. Weather conditions for each crash case and non-crash case were extracted based on their time and milepost location. Considering the sample size in each category, the rain and fog were combined as adverse weather conditions. As a result, the study considered two difference weather conditions, including clear weather and adverse weather.

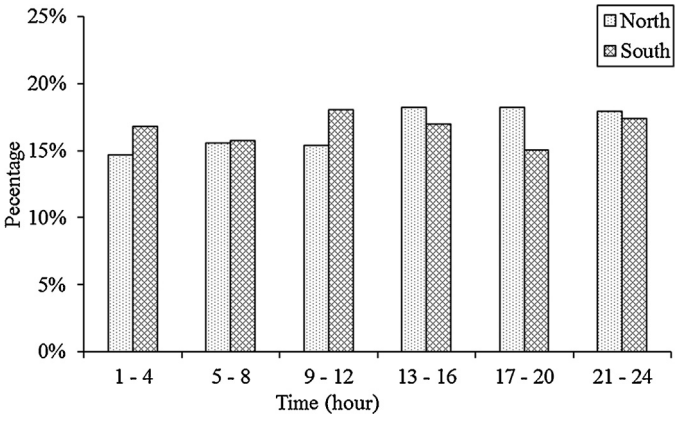

Fig. 2. The distribution of non-crash cases over time.

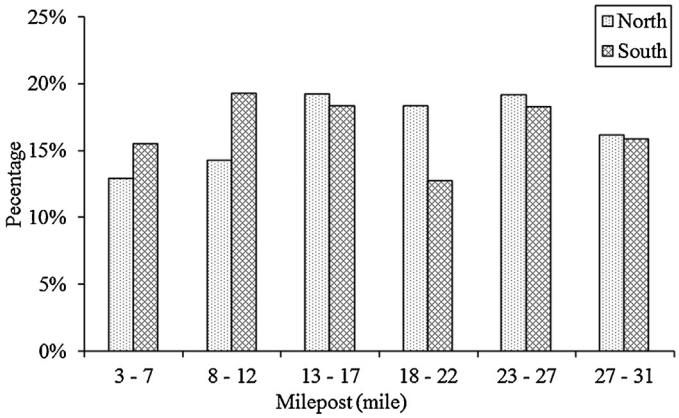

Fig. 3. The distribution of non-crash cases along the I-880 Freeway.

The crash data were obtained from the Statewide Integrated Traffic Records System (SWITRS) maintained by Caltrans. The crash dataset provided by SWITRS included crash location, crash time, and crash severity. The crash severity is divided into five levels, including fatal crash (K), incapacitating injury crash (A), non-incapacitating injury crash (B), possible injury crash (C), and property-damage-only crash (PDO). The definition of each crash severity level is:

A fatal injury is any injure that results in death within a 30 day period after the crash occurred.

An incapacitating injury is any injure other than a fatal injure, which prevents the injured person from normally continuing the activities the person was capable of performing before the injury occurred.

A non-incapacitating injury is any injure other than a fatal injury or an incapacitating injury, which is evident to observers at the scene of the crash.

A possible injury is any injure that includes complaint of pain without visible injury.

The original five levels of severity were combined into three levels: KA, BC, and PDO (Leckrone et al., 2011; Jung et al., 2010). This combining process acknowledged the similarity of the combined levels and increased the number of crashes at each new level, thereby improving the chance for more significant variables being included in the final models. Most of the crashes (56.0%) in the dataset were rear-end crashes followed by sideswipe crashes with about 22.3%. About 54.9% and 13.7% of injury crashes in dataset were rear-end crashes and sideswipe crashes respectively, and 20.2% were hit object crashes. The crash frequency for each severity level is shown in Table 2. A total of 794 crash cases and 15,880 non-crash cases were included in our database.

3. Research methodology

The ordered probit/logit model is the most common modeling approach that fits the data structure of an ordinal response. However, the main drawback associated with ordered probit/logit models is that the parameters estimates and the set of significant explanatory variables are the same over all the crash severity levels. Even though the parameters estimates of each explanatory variable could be different across different crash severity levels in the generalized ordered logit model, the set of significant explanatory variables is still assumed to be the same across different crash severity levels. Considering the fact that the multinomial and nested logit models cannot explicitly represent the ordinality in the discrete categories of the injury severity, this study applied a sequential logit model to capture the impacts of different traffic flow parameters on the crash likelihood at various severity levels.

Table 2
Frequency distribution of observations in the sequential logit model.

Crash severity level	Sequential structure		
	Stage 1	Stage 2	Stage 3
Fatal and incapacitating injury (KA)	59 (1[a])	59 (1)	59 (1)
Non-incapacitating and possible injury(BC)	203 (1)	203 (1)	203 (0)
Property damage only (PDO)	532 (1)	532 (0)	–
Non-crash	15,880 (0)	–	–
Total	16,674	794	262

[a] SAS coding of crash severity level are in parentheses.

3.1. Binary logit model

In this study, a binary logit model was applied at each stage to fit the developed sequential logit model. At each stage, a binary logit model was used to fit a sub-sample that excluded the observations of a certain level used in the previous stage. The binary logit regression model was used in previous studies for predicting a binary dependent variable as a function of the predictor variables in transportation engineering (Xu and Tian, 2008; Hubbard et al., 2009; Liu et al., 2007). Using the binary logit model, the probability of the occurrence of a crash can be estimated using the following equation:

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

where $P(x_i)$ denotes the probability of the occurrence of a crash and $g(x)$ is the multiple linear combination of explanatory variables, which can be expressed as:

$$g(x) = \ln \frac{P(x_i)}{1 - P(x_i)} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \tag{2}$$

where x_{ki} denotes the value of variable k for sample i and β_k is the coefficient of variable k. The parameters β_0, β_1, $\beta_2, \ldots, \beta_k$ can be estimated by solving the log-likelihood function for Eq. (2), which is given by:

$$\ln\ L(\beta, x_i) = \sum_{i=1}^{n} [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{ki} x_{ki} - \ln(1 + e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{ki} x_{ki}})] \tag{3}$$

As per findings of previous studies (Hauer and Hakkert, 1988; Elvik and Mysen, 1999; Hauer, 2006; Savolainen et al., 2011), less severe crashes are more likely to be under-reported and the under-reporting rate decreases with the increase in severity level. Thus, accident samples are usually over-represented by high severity level crashes. Further, it is prohibitive to include all the observations of non-crash cases in the dataset. 15,880 observations of non-crash cases were randomly selected to represent the normal traffic conditions. Finally, due to the missing or invalid real-time traffic data, the crashes that could not be matched with real-time traffic data were excluded from further data analysis. These factors will make the dataset an outcome (choice)-based sample.

When the conventional maximum likelihood estimator (MLE) is used to the outcome-based samples, the multinomial logit model could still produce unbiased estimates for model parameters except the constant terms (Cosslett, 1981a,b; Yamamoto et al., 2008; Patil et al., 2011). Similar to multinomial logit model, the parameter estimates in the binary sequential logit model are unbiased except constant terms (Yamamoto et al., 2008; Savolainen et al., 2011), because the binary sequential logit model is the combination of several binary logit models.

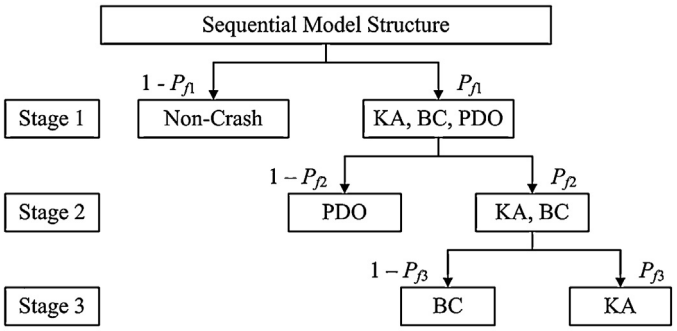

Fig. 4. The structure of the sequential logit model considered in this study.

The computation of the predicted probability is based on the full set of parameter estimates including the constant term, thus the predicted probability by the binary sequential logit model would be biased. To account for the biases caused by outcome-based sampling scheme, the intercept of the binary logit model at each stage was adjusted by using an offset (Scott and Wild, 1986). To adjust the estimated intercept, an offset value calculated as shown below was added to the original intercept.

$$\text{offset} = -\text{Ln}\left(\frac{SR_i}{PR_i}\right) \tag{4}$$

where SR_i represents the ratio of observations having outcome i to other observations in the sample and PR_i represents the ratio of observations having outcome i to other observations in the total population.

To evaluate the effect of the traffic flow variables on the likelihood of crash and its severity, the elasticity analysis was conducted. The elasticity represents the percentage change in the dependent variable resulting from a 1% change in an independent variable (Washington et al., 2003). The elasticity of the dependent variable Y with respect to a continuous independent variable x_i is given as:

$$E_i = \frac{\partial Y_i}{\partial x_i} \times \frac{x_i}{Y_i} = [1 - P(i)]\beta_i x_i \tag{5}$$

Although each observation in the dataset has an elasticity that depends on the value of x_i and the estimated probability of crash severity $P(i)$, it is customary to report the average elasticity in the sample. In the following analysis, both the average and standard deviation of the elasticity are given. Note that Eq. (5) cannot be used to calculate the elasticity for indictor variables. The sensitivity of an indicator variable x_i is made by computing a pseudo-elasticity using the following equation (Washington et al., 2003):

$$E_i = \left[\frac{\text{EXP}[\Delta(x'\beta)][1 + \text{EXP}(x_i\beta_i)]}{\text{EXP}[\Delta(x'\beta)][\text{EXP}(x_i\beta_i)] + 1} - 1\right] \times 100 \tag{6}$$

3.2. Model structure

Fig. 4 illustrates the structures of the binary sequential logit model used in this study. The severity level of crash in the sequential model varied from the lowest to the highest level. As shown in Table 2, the sequential model were conducted in this study as follows,

Stage 1. Crash types KA, BC, and PDO (binary response = 1) vs. non-crash cases (binary response = 0).
Stage 2. Crash type KA and BC (binary response = 1) vs. PDO (binary response = 0).
Stage 3. Crash types KA (binary response = 1) vs. crash types BC (binary response = 0).

Based on the estimated binary logit model at each stage of the sequential model, the likelihood of crash occurrence at different severity levels can be calculated as follows:

$$P(\text{Crash}) = P_{f1} \tag{7}$$

$$P(\text{PDO}) = P(\text{Crash})\,P(\text{PDO}|\text{Crash}) = P_{f1}(1 - P_{f2}) \tag{8}$$

$$P(\text{BC}) = P(\text{Crash})\,P(\text{KA } or \text{ BC}|\text{Crash})\,P(\text{BC}|\text{KA or BC}) = P_{f1}P_{f2}(1 - P_{f3}) \tag{9}$$

$$P(\text{KA}) = P(\text{Crash})P(\text{KA or BC}|\text{Crash})P(\text{KA}|\text{KA } or \text{ BC}) = P_{f1}P_{f2}P_{f3} \tag{10}$$

where $P(Y)$ is the probability of Y; $P(Y|X)$ is the probability of Y given that X happens; P_{fi} represents the estimated probability calculated by the binary logit model at each stage of the sequential structure as shown in Fig. 4.

3.3. Model validation technique

This study used the k-fold cross-validation method to estimate the predication accuracy of the developed models. The k-fold cross-validation method can minimize the bias associated with the random sampling of the training and validation dataset in estimating the prediction accuracy of a model (Olson and Delen, 2008). In the k-fold cross-validation method, the complete dataset is randomly partitioned into k mutually exclusive subsets of approximately equal size. Among the k datasets, each single subset is used as the validation dataset, and the other $k-1$ subsets are combined to form a training dataset. Hence, the prediction model is trained and tested for k times (the folds). In this study, a 20-fold cross-validation was conducted to estimate the prediction performance of the developed model.

4. Estimated models

The LOGISTIC procedure in SAS 9.2 was used to specify the binary logit model at each stage with a significance level of 0.1 for retaining the explanatory variables in the models (SAS, 2011). To account for the possible correlations between the candidate independent variables, the Pearson correlation parameters were calculated between different pairs of candidate independent variables and several combinations were generated that included the maximum number of uncorrelated variables. Stepwise variable selection was then applied to select the independent variables that should be included in the binary logit model at each stage. The log likelihood at the convergence of each model was compared, and the model with the highest log likelihood was considered the best model.

4.1. Results discussion

4.1.1. First-stage model – all crashes

The estimation results for the sequential logit model are shown in Table 3. In the first stage, six traffic flow variables: the upstream occupancy, the upstream speed variance, the downstream speed variance, the difference in occupancy between adjacent lanes at the downstream station, the difference in occupancy between upstream and downstream station, the difference in vehicle count between upstream and downstream station were found to be significantly correlated with the likelihood of crash. The results shown in Table 3 indicate that the crash likelihood tends to be high when the traffic density at the upstream detector station (approximated with DetOcc_u), the speed variance at the upstream detector

Table 3
Estimation results for the sequential logit model.

	Parameter	Estimate	Std. Error	Wald χ^2	Pr > Chisq	Elasticity
Stage 1	Crash (KA, BC, and PDO) vs. non-crash					
	$DetOcc_u$	0.074	0.007	99.586	<0.0001	0.467 (0.337[b])
	$SpdDev_u$	0.060	0.016	14.544	<0.0001	0.224 (0.142)
	$SpdDev_d$	0.050	0.016	9.079	0.003	0.189 (0.117)
	$OccDif_d$	0.119	0.011	109.175	<0.0001	0.299 (0.299)
	$AvgCnt_{u-d}$	0.092	0.039	5.593	0.018	0.09 (0.087)
	$AvgOcc_{u-d}$	0.026	0.013	3.663	0.056	0.038 (0.057)
	Weather	0.886	0.141	39.410	<0.0001	0.408 (0.041)
	$DetDist_{u-d}$	1.057	0.089	140.246	<0.0001	0.495 (0.365)
	$Width_s$	−0.049	0.008	38.836	<0.0001	−2.535 (0.281)
	$Width_o$	−0.856	0.150	32.365	<0.0001	−0.415 (0.028)
	Curve	0.508	0.121	17.646	<0.0001	0.243 (0.019)
	Intercept	−2.672 (−4.704[a])	0.420	40.545	<0.0001	–
	Summary statistics:					
	$-2L(c) = 6347.700$; $-2L(\beta) = 5408.806$					
	$-2[L(c) - L(\beta)] = 938.894$ (11df); $P < 0.0001$					
Stage 2	KA and BC vs. PDO					
	$DetOcc_u$	−0.033	0.013	6.087	0.014	−0.202 (0.184)
	$VehCnt_d$	−0.056	0.025	5.180	0.023	−0.303 (0.142)
	Peak	−0.335	0.174	3.698	0.054	−0.174 (0.011)
	Weather	−0.689	0.300	5.290	0.021	−0.338 (0.019)
	$Width_s$	−0.036	0.016	5.280	0.022	−0.956 (0.273)
	Intercept	2.129 (0.644)	0.822	6.707	0.010	–
	Summary statistics:					
	$-2L(c) = 1007.047$; $-2L(\beta) = 963.730$					
	$-2[L(c) - L(\beta)] = 43.317$ (5df); $P < 0.0001$					
Stage 3	KA vs. BC					
	$AvgSpd_u$	0.033	0.015	4.900	0.027	2.022 (0.448)
	$SpdDif_u$	0.067	0.020	10.723	0.001	0.859 (0.404)
	$VehCnt_d$	−0.117	0.042	7.577	0.006	−1.007 (0.464)
	Intercept	−3.510 (−1.971)	1.199	8.568	0.003	–
	Summary statistics:					
	$-2L(c) = 279.501$; $-2L(\beta) = 238.076$					
	$-2[L(c) - L(\beta)] = 41.426$ (3df); $P < 0.0001$					

[a] The intercept adjustment for each logit model.
[b] The standard deviation of elasticity for each variable.

station (represented by $SpdDev_u$), the speed variance at the downstream detector station (represented by $SpdDev_d$), the traffic inter-lane imbalance (measured with $OccDif_d$), the volume difference between upstream and downstream station (represented by $AvgCnt_{u-d}$), and the occupancy difference between upstream and downstream station (represented by $AvgOcc_{u-d}$) are high as well. These results are consistent with the high lane-change activities postulated by Gazis et al. (1962) for certain conditions and confirm the findings of previous statistical analyses (Lee et al., 2003; Abdel-Aty et al., 2004; Ahmed and Abdel-Aty, 2012; Xu et al., 2012a,b). Generally speaking, the small distances between vehicles in high-density traffic flow leave less time for taking crash avoidance maneuver. Speed fluctuations and traffic imbalances between lanes may encourage drivers to change lanes more frequently – a maneuver that can be quite dangerous in dense traffic.

Elasticity tells how many times the crash probability changes if the explanatory variable changes by 1% while the other variables remain fixed. Unlike the marginal effects, the elasticity is dimensionless, thus it is more convenient for comparing the effects of different variables. For example, the average elasticity values for the six traffic flow characteristics ($DetOcc_u$, $SpdDev_u$, $SpdDev_d$, $OccDif_d$, $AvgCnt_{u-d}$, and $AvgOcc_{u-d}$) are: 0.467, 0.224, 0.189, 0.299, 0.090, and 0.038, respectively. It means that the one-percent increase in these six traffic flow characteristics is associated with the 0.467%, 0.224%, 0.189%, 0.299%, 0.090%, and 0.038% increases in the crash probability, respectively.

Among the geometric variables, the spacing between upstream and downstream stations ($DetDist_{u-d}$), the road surface width ($Width_s$), the outer shoulder width ($Width_o$) and the curve section (Curve) were found to be significant in the crash probability model (first stage). The positive coefficient of variable $DetDist_{u-d}$ indicates that the probability of crash grows with the length of the road segment. The findings from the aggregate crash prediction models in previous studies also demonstrated that the increase in road segment resulted in an increase in the number of accidents (Anastasopoulos and Mannering, 2009). Both variables $Width_s$ and $Width_o$ had negative coefficients, indicating that the crash risk decreases with the increase in road surface width and out shoulder width. As indicated by the coefficient of the variable Curve, the curve segment could increase the crash risk on freeways. These results are consistent with the findings of previous studies (Anastasopoulos and Mannering, 2009; Das and Abdel-Aty, 2010; Shively et al., 2010). The positive coefficient of weather conditions indicates that adverse weather conditions could increase crash likelihood on freeways. Surprisingly, the model does not include the presence of on-ramps and off-ramps between the detector stations. It seems that the six traffic flow characteristics included in the model have already captured the impacts of the ramps.

As mentioned above, the parameter estimates in the binary sequential logit model are unbiased except constant terms when MLE is used to the outcome-based samples. To account for the biases caused by outcome-based sampling scheme, Eq. (4) was used to adjust the intercept of the model at the first stage as follows:

$$\text{offset} = -\text{Ln}\left(\frac{SR_{\text{crash}}}{PR_{\text{crash}}}\right) = -\text{Ln}\left(\frac{\text{Crash}_s}{\text{Crash}_p} \times \frac{\text{Non-crash}_p}{\text{Non-crash}_s}\right)$$
$$= -\text{Ln}\left(\frac{\text{Crash}_r}{\text{Crash}_p} \times \frac{\text{Crash}_s}{\text{Crash}_r} \times \frac{\text{Non-crash}_p}{\text{Non-crash}_s}\right) \quad (11)$$

where SR_{crash} represents the ratio of crash cases to non-crash cases in the data sample; PR_{crash} represents the ratio of crash cases to non-crash cases in the total population; $Non\text{-}crash_s$ and $Non\text{-}crash_p$ represent the number of non-crash cases in the data sample and total population, respectively; $Crash_s$ and $Crash_p$ represent the number of crashes in the data sample and total population, respectively; $Crash_r$ represents the reported number of crashes.

As shown in Eq. (11), three ratios were included in the offset to account for the biased caused by the outcome-based sampling scheme. The ratio between $Non\text{-}Crash_s$ and $Non\text{-}Crash_p$ is used to account for the reduction in the number of non-crash cases in the data sample. Due to the missing or invalid real-time traffic data, the number of crashes in the data sample is smaller than the number of reported crashes. The ratio between $Crash_s$ and $Crash_r$ is used to account for the reduction in the number of crashes caused by missing or invalid real-time traffic data. Finally, the ratio between $Crash_r$ and $Crash_p$ is used to account for the effect of underreporting of crash data. Although the actual value for reporting rate of all crashes is unknown, this value can be estimated using the results of previous studies about under reporting of crashes. The study conducted by Elvik and Mysen (1999) reported that the reporting rates for crash K, A, B, C, O are 95%, 69%, 27%, 11%, and 25%, respectively. By the combination of the above reporting rates and reported crashes in the crash data, the reporting rate of all crashes can be easily estimated. After determining the reporting rate of all crashes, the offset for adjusting the intercept of the model at the first stage can be estimated using Eq. (11). The offsets for adjusting the intercept terms at other stages shown in Table 3 can be estimated using the similar method shown above.

4.1.2. Second-stage model – injury crashes

At the second stage, the upstream traffic density (represented by $DetOcc_u$), the downstream traffic volume (represented by $VehCnt_d$), the weather conditions (represented by weather), the peak period (represented by Peak), and road surface width (represented by $Width_s$) were found to be significantly correlated with the risk of injury and fatality once a crash happens. The negative coefficients of the two traffic flow variables indicate that an injury is less likely to result from a crash that occurs in more intense and congested traffic. The interpretation here is that more congested traffic tends to be slower than less congested conditions; thus crashes tend to occur at relatively low speeds, thereby decreasing the risk of injury and fatality. The findings from the aggregate crash prediction models in previous studies also demonstrated that the crashes occurred in congested traffic conditions are likely to be less severe (Shefer, 1997; Chang and Xiang, 2003; Wang et al., 2009). The average elasticity for the upstream occupancy ($DetOcc_u$) was −0.202 which indicates that the 1% increase in the upstream occupancy will result in 0.202% reduction in the probability of injury crash. The average elasticity of −0.303 for the downstream flow indicates that the probability of injury crash decreases by 0.303% for each 1% increase in $VehCnt_d$. The variable $Width_s$ had negative coefficients, indicating that the risk of injury and fatality decreases with the increase in road surface width. Large road surface width gives drivers more space for taking crash avoidance maneuver and reducing crash severity. Both variables Peak and Weather had negative coefficient, indicating that the peak period and adverse weather conditions could decrease the risk of injury and fatality. These are consistent with the findings of previous studies (Das and Abdel-Aty, 2010; Khattak and Knapp, 2001; Kim et al., 2013).

4.1.3. Third-stage model – fatal and incapacitating injury crashes

The model estimated in the third stage implies that the average speed measured at the upstream detector station ($AvgSpd_u$), the difference in speeds between adjacent lanes at the upstream station ($SpdDif_u$) and the downstream traffic flow (represented by $VehCnt_d$) were significantly correlated with the risk of fatal or incapacitating injury upon crash occurrence. Both variables $AvgSpd_u$ and $SpdDif_u$ had positive coefficients, indicating a high risk of a fatal or incapacitating injury outcome from a crash if the crash occurs at high speed and considerable speed difference between lanes. On the other hand, the negative coefficient of the downstream flow variable ($VehCnt_d$) indicates that a fatality and incapacitating injury are less likely in a crash that occurs in a high-volume traffic flow. Therefore, crashes that occur in less congested traffic flow conditions with high speeds and high speed differences between adjacent lanes are prone to produce fatal and incapacitating injuries.

The average elasticity for average speed variable ($AvgSpd_u$) was 2.022; and for the speed difference across lanes ($SpdDif_d$), the average elasticity was 0.859, indicating that a 1% increase in $AvgSpd_u$ and $SpdDif_u$ increases the probability of a fatal and incapacitating injury crash by 2.022% and 0.859%, respectively. The average elasticity of −1.007 for the flow intensity (represented by $VehCnt_d$) implies that the probability of a fatal and incapacitating injury crash decreases 1.007% for 1% increase in $VehCnt_d$.

4.2. Prediction performance

4.2.1. Test design

This study applied the 20-fold cross-validation method to estimate the prediction performance of the developed models. The whole sample was randomly partitioned into 20 mutually exclusive sub-samples of approximately equal size. Then, each sub-sample was used as a validation sample and the other 19 sub-samples were combined as a training sample. The sequential logit models were developed for different training samples. Note that the explanatory variables included in the binary logit model at each stage were the same variables shown in Table 3. Eqs. (7)–(10) were used to calculate the crash likelihood at different severity levels for the observations in each validation sample.

4.2.2. Measure of performance and results

The prediction accuracy of a model of binary outcome (event = 1 and non-event = 0) can be measured with two complementary indicators: (1) the proportion of events predicted as an event (true positive rate) called *sensitivity* in SAS, and (2) the proportion of non-events predicted as a non-event (true negative rate) called *specificity* in SAS. The model predicts an event if the predicted probability of event exceeds a pre-specified threshold. The sensitivity and specificity depend on the same threshold varying between 0 and 1. A convenient and meaningful tool for evaluating the prediction performance of a logit model is a curve called the Receiver Operating Characteristic (ROC), which compares the sensitivity and the 1 − specificity for a threshold running from 0 to 1. Thus, the ROC curve is a graphical plot of the sensitivity (y-axis) vs. 1 − specificity (x-axis). To generate the ROC curve for each severity level, we calculated the sensitivity and 1 − specificity for multiple thresholds by using all of the validation samples. Fig. 5 presents the ROC curves for four possible outcomes for the sequential logit model: (1) crash at any severity level, (2) property-damage-only crash (PDO), (3) non-incapacitating or possible injury crash (BC), and (4) fatal or incapacitating injury crash (KA).

The areas under the ROC curves for the four possible outcomes were found to be 0.784, 0.803, 0.758, and 0.797 respectively, indicating that the sequential logit model can provide good predictive performance. Table 4 summarizes the crash prediction performance at different severity levels for the developed sequential logit model. The prediction performance is measured by the percent of predicted crashes at each severity level and for several false alarm (false positive) rates. As shown in Table 4, the prediction accuracy of crashes increased as the false alarm rate was increased. This

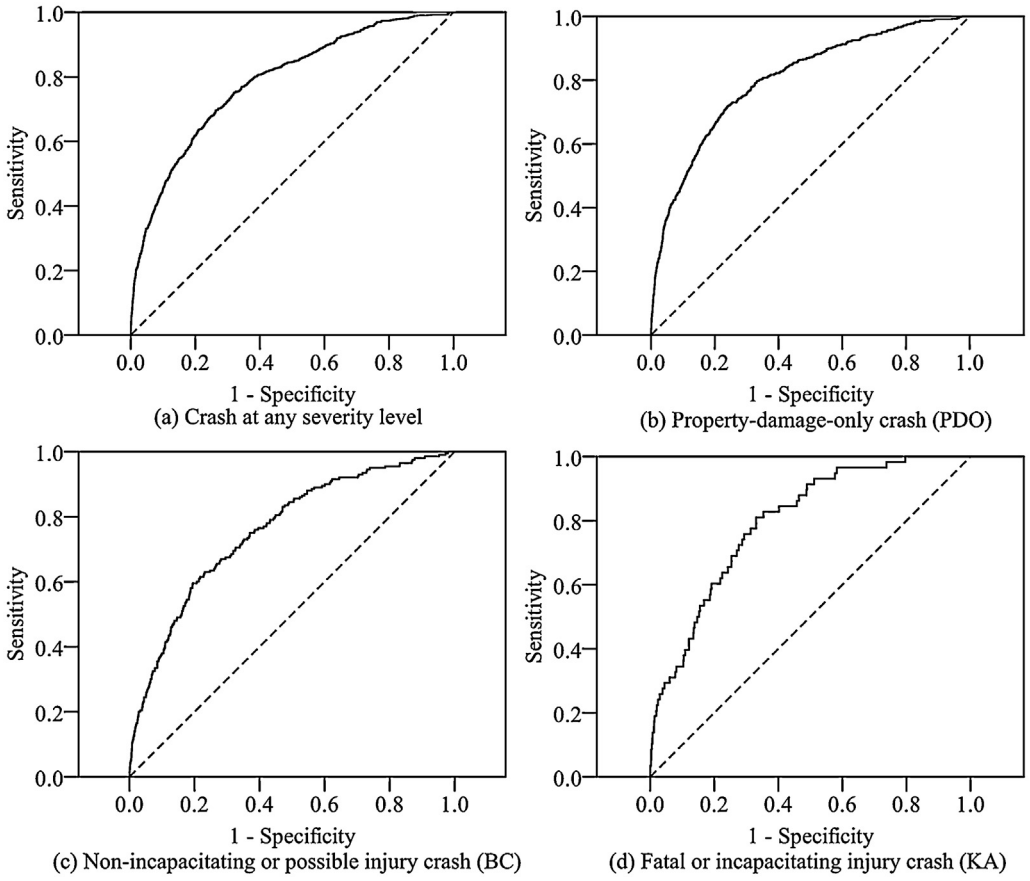

Fig. 5. The ROC curve of each severity level for the sequential logit model.

Table 4
Prediction performance at different false alarm rates.

1 – specificity	Sensitivity of the sequential model		
	PDO	BC	KA
0.02	22.8%	16.5%	22.4%
0.1	48.5%	38.0%	37.9%
0.2	66.2%	59.5%	60.3%
0.3	75.7%	67.5%	75.9%
0.4	82.3%	76.5%	84.5%
0.5	87.2%	84.5%	91.4%

trade-off between the prediction accuracy and the false alarm rate must be considered when setting a threshold value and needs to be determined carefully to meet the requirement of the practical implementation or the preference of a specific traffic agency. After determining the threshold, the ROC curve can be easily used to estimate the predictive performance. For example, if a threshold value is selected to accept a 30% false alarm rate, the prediction accuracy of PDO, BC, KA crashes is found to be 75.7%, 67.5% and 75.9%, respectively on the ROC curve. For comparison purpose, we also briefly summarized the previous studies regarding the real-time crash risk prediction modes in Table 5. Compared with the predictive performance of the models in previous studies in Table 5, the predictive performance of the models in this study is good.

Table 5
Summary of predictive performance of the real-time crash risk models in previous studies.

Authors	Prediction accuracy of crash	Prediction accuracy of non-crash	False alarm rate	Sample sized of crash	Sample size of non-crash
Oh et al. (2001)	55.8%	72.1%	27.9%	52	4787
Oh et al. (2005)	35.2%	73.5%	26.5%	52	4787
Abdel-Aty et al. (2004)	69.4%	52.8%	47.2%	375	1875
Abdel-Aty and Pande (2005)	73.9%	71.3%	28.7%	377	2857
Abdel-Aty et al. (2005)	56.0%	80.0%	20.0%	1528	6112
Pande and Abdel-Aty (2006)	57.0%	70.0%	30.0%	162	3650
Hossain and Muromachi (2010)	63.3%	80.0%	20.0%	250	23,068
Hassan and Abdel-Aty (2011)	67.2%	64.7%	35.3%	67	201
Pande et al. (2011)	45.0%	80.0%	20.0%	533	2660
Ahmed and Abdel-Aty (2012)	69.9%	54.9%	45.2%	670	2680
Ahmed et al. (2012)	72.9%	57.9%	42.0%	447	1788
Abdel-Aty et al. (2012)	73.1%	60.3%	39.7%	106	670
Yu and Abdel-Aty (2013)		AUC = 0.75		265	1017
Xu et al. (in press)	61.0%	80.0%	20.0%	807	8070

4.2.3. Implementation discussion

One possible real-time implementation of a model such the one presented in this paper is calculating the probabilities of crashes at certain levels of severity between specific detector stations and warning drivers entering the segment about the high risk. It is imperative, however, in such an implementation, that the rate of false positives is not too high because of the anticipated erosion of drivers' attention to the message. Although drivers cannot tell in advance which alerts will not be followed by actual crash events, they will soon realize the high frequency of such alerts. For example, let us assume the false alarm rate is 50% and a driver who travels daily to work on a five-mile freeway section with 10 pairs of detector stations and who will be subject to five alarms on average during a one-way trip. Even if drivers are not familiar with the frequency of crashes, they may become immune to the alerts quickly if no crashes are apparent. If the false alarm rate is set at much lower rate, say 1/50, then the traveler in the considered example would witness one false alarm a week. However, setting the false alarm rate at 0.02% makes the prediction performance rather low. Fortunately, it seems that predictive accuracy of the most severe crashes is not too bad. Being able to predict approximately 22% of incapacitating injury or fatal crashes and preventing, hopefully, most of them without eroding the alertness of motorists might be sufficient justification for implementing real-time warning systems utilizing models such the one presented in this paper.

4.3. Temporal and spatial transferability

Previous studies have demonstrated that the real-time crash prediction models cannot be directly transferred from one road to another due to the difference in driver population and traffic patterns (Pande et al., 2011; Ahmed and Abdel-Aty, 2012). The Bayesian updating approach could be used to improve the temporal and spatial transferability of the developed sequential model above. The Bayesian updating approach could update the old model as new data becomes available. Assuming that a real-time crash prediction model has been developed based on the historical data Y_1, and that the new data or data from other road Y_2 are then obtained, the posterior distribution can be updated using Bayes' theorem as follows:

$$\pi(\theta|Y_1, Y_2) \propto f(Y_1, Y_2|\theta)\pi(\theta) = f(Y_2|Y_1, \theta)f(Y_1|\theta)\pi(\theta) \propto f(Y_2|\theta)\pi(\theta|Y_1) \quad (12)$$

Thus, when updating a new model, we can use the estimation results of the sequential logit model above to develop informative prior distribution $\pi(\theta|Y_1)$ and incorporate the new data or data from other road Y_2 in a new updated posterior distribution. It is obvious that the developed sequential logit model in this study can be easily updated by the Bayesian updating approach, but this application remains as a future study task.

5. Conclusions

Most of the existing studies consider the likelihood of crash without considering the crash outcome severity based on the differing contribution of traffic flow characteristics to the crash probability at different severity levels. However, due to their much higher social and economic impacts than less severe crashes, the ability to predict the occurrence likelihood of severe crashes is important. This study not only further developed real-time crash risk prediction models to identify hazardous traffic conditions that potentially lead to crashes, but addressed as well the possibility of predicting crashes at various levels of severity in real-time using traffic data collected with freeway loop detectors.

The sequential logit model was applied to link the likelihood of crash occurrences at different severity levels with various traffic flow characteristics. The real-time traffic and crash data utilized in this study were obtained on the I-880 freeway in California, United States. The model estimation results showed that the traffic flow characteristics contributing to crash probability were found to vary substantially across different crash severity level. In general, the low severity crashes (PDO) tended to occur in congested traffic flow conditions with highly variable speed and frequent lane changes. The injury crashes (KA and BC) were found to occur more often in less congested traffic flow conditions. The KA crashes, in particular, occurred under uncongested traffic flow conditions as well as with large differences in speed between adjacent lanes. The elasticity analysis was conducted at each stage of the sequential logit model estimation to evaluate the effects of the traffic flow variables on the likelihood of crash severity.

The 20-fold cross-validation method was applied to evaluate the predictive performance of the developed sequential logit model. The validation results demonstrate that the predictive performance of the developed models was deemed satisfactory. The predictive accuracy of PDO, BC, KA crashes is found to be 75.7%, 67.5% and 75.9% when the false alarm rate was equal to 30%. As expected, there was a strong trade-off between the false alarm rate and the percentage of crashes predicted by the model, which is an important aspect. A model should have a reasonably low false alarm rate in order to reduce the danger of losing drivers' responsiveness to alerts about high-risk traffic conditions. The set of developed models were able to correctly identify high-risk severe crash conditions in 22% of the cases with the false alarm rate set at a safe level of 2%. This result provides hope that the real-time prediction of severe crashes is possible, and implementation of this study's results, based on warning motorists, is practical.

Other advanced dynamic safety management systems, such as variable speed limits and ramp metering, may benefit from the real-time detection of high-risk conditions as well if the connection between these traffic control methods and safety is better understood. In these cases, the false alarm rate is not as critical as it is for warning-based solutions because the road users are not aware of the reason for a reduced speed limit or a ramp metering rate change.

Acknowledgement

This research was jointly sponsored by China's National Key Basic Research Program (No. 2012CB725400), China's National High-tech R&D Program (No. 2011AA110303-03), the Scholarship Award for Excellent Doctoral Student granted by the Ministry of Education of China, and Scientific Research Foundation of Graduate School of Southeast University.

References

Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., Hsia, L., 2004. Predicting freeway crashes based on loop detector data using matched case-control logistic regression. Transportation Research Record 1897, 88–95.

Abdel-Aty, M., Uddin, N., Pande, A., 2005. Split models for predicting multi-vehicle crashes during high-speed and low-speed operating conditions on freeways. Transportation Research Record 1908, 51–58.

Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. Journal of Safety Research 36 (1), 97–108.

Abdel-Aty, M., Hassan, H., Ahmed, M., Al-Ghamdi, A., 2012. Real-time prediction of visibility related crashes. Transportation Research Part C 24, 288–298.

Ahmed, M., Abdel-Aty, M., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Transactions on Intelligent Transportation Systems 13 (2), 459–468.

Ahmed, M., Abdel-Aty, M., Yu, R., 2012. A Bayesian updating approach for real-time safety evaluation using AVI data. In: Presented at the 91th Annual Meeting of the Transportation Research Board, CD-ROM, Washington, DC.

Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41, 153–159.
Chang, G., Xiang, H., 2003. The Relationship Between Congestion Levels and Accidents. Maryland State Highway Administration, Baltimore, MD.
Cosslett, S.R., 1981a. Efficient estimation of discrete-choice methods. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Choice Data with Econometric Applications. MIT Press, Cambridge, MA, pp. 51–111.
Cosslett, S.R., 1981b. MLE for choice-based samples. Econometrica 49, 1289–1316.
Das, A., Abdel-Aty, M., 2010. A genetic programming approach to explore the crash severity on multi-lane roads. Accident Analysis and Prevention 42, 548–557.
Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting; meta-analysis of studies made in 13 countries. Transportation Research Record 1665, 133–140.
Golob, T., Recker, W., Pavlis, Y., 2008. Probabilistic models of freeway safety performance using traffic flow data as predictors. Safety Science 46 (9), 1306–1333.
Gazis, D., Herman, R., Weiss, G.H., 1962. Density oscillations between lanes of a multilane highway. Operations Research 10, 658–667.
Hauer, E., Hakkert, A., 1988. Extent and some implications of incomplete accident reporting. Transportation Research Record 1185, 1–10.
Hauer, E., 2006. The frequency-severity indeterminacy. Accident Analysis and Prevention 38, 78–83.
Hassan, H., Abdel-Aty, M., 2011. Exploring visibility-related crashes on freeways based on real-time traffic flow data. In: Presented at 90th Annual Meeting of the Transportation Research Board, CD-ROM, Washington, D.C.
Hossain, M., Muromachi, Y., 2010. Evaluating location of placement and spacing of detectors for real-time crash prediction on urban expressways. In: Presented at 89th Annual Meeting of the Transportation Research Board, CD-ROM, Washington, D.C.
Hossain, M., Muromachi, Y., 2011. Understanding crash mechanism and selecting appropriate interventions for real-time hazard mitigation on urban expressways. Transportation Research Record 2213, 53–62.
Hourdos, N., Garg, V., Michalopoulos, G., Davis, G., 2006. Real-time detection of crash-prone conditions at freeway high-crash locations. Transportation Research Record 1968, 83–91.
Hubbard, S., Bullock, D., Mannering, F., 2009. Right turns on green and pedestrian level of service: statistical assessment. Journal of Transportation Engineering 135 (4), 153–159.
Jung, S., Qin, X., Noyce, D., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accident Analysis and Prevention 42 (1), 213–224.
Khattak, A., Knapp, K., 2001. Interstate highway crash injuries during winter snow and nonsnow events. Transportation Research Record 1746, 30–36.
Kim, J., Ulfarsson, F., Kim, S., Shankar, V., 2013. Driver-injury severity in single-vehicle crashes in California: a mixed logit analysis of heterogeneity due to age and gender. Accident Analysis and Prevention 50, 1073–1801.
Lee, C., Saccomanno, F., Hellinga, B., 2003. Real-time crash prediction model for the application to crash prevention in freeway traffic. Transportation Research Record 1840, 67–77.
Lee, C., Park, P., Abdel-Aty, M., 2011. Lane-by-lane analysis of crash occurrence based on driver's lane-changing and car-following behavior. Journal of Transportation Safety and Security 3 (2), 108–122.
Leckrone, S.J., Tarko, A.P., Anastasopoulos, P.C., 2011. Improving safety at high-speed rural intersections. In: Presented at 3rd International Conference on Road Safety and Simulation, CD-ROM, Indianapolis, IN.
Liu, P., Wang, X., Lu, J., Sokolow, G., 2007. Headway acceptance characteristics of U-turning vehicles at unsignalized intersections. Transportation Research Record 2027, 52–57.
Li, Z., Chung, K., Liu, P., Wang, W., Ragland, D., 2012. Surrogate safety measure for evaluating rear-end collision risk near recurrent bottlenecks. In: Presented at the 91th Annual Meeting of the Transportation Research Board, CD-ROM, Washington, DC.
Oh, C., Oh, J., Ritchie, S., 2001. Real-time estimation of freeway accident likelihood. In: Presented at 80th Annual Meeting of the Transportation Research Board, CD-ROM, Washington, D.C.
Oh, C., Oh, J., Ritchie, S., 2005. Real-time hazardous traffic condition warning system: framework and evaluation. IEEE Transactions on Intelligent Transportation Systems 6 (3), 265–272.
Olson, D., Delen, D., 2008. Advanced Data Mining Techniques. Springer, Berlin, Germany.
Pande, A., Abdel-Aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accident Analysis and Prevention 38 (5), 936–948.
Pande, A., Dasand, A., Abdel-Aty, M., Hassan, H., 2011. Real-time crash risk estimation are all freeways created equal? Transportation Research Record 2237, 60–66.
Patil, S., Geedipally, S., Lord, D., 2011. Analysis of crash severities using nested logit model—accounting for the underreporting of crashes. Accident Analysis and Prevention 45, 646–653.
Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43, 1666–1676.
Scott, A.J., Wild, C.J., 1986. Fitting logistic models under case-control or choice based sampling. Journal of the Royal Statistical Society Series B48 (2), 170–182.
SAS Institute Inc, 2011. SAS/STAT(R) 9.2 User's Guide, second edition.
Shively, T., Kockelman, K., Damien, P., 2010. A Bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics. Transportation Research Part B 44, 699–715.
Shefer, D., 1997. Congestion and safety on highways: towards an analytical model. Urban Studies 34 (4), 679–692.
Wang, C., Quddus, M., Ison, S., 2009. The effects of area-wide road speed and curvature on traffic casualties in England. Journal of Transport Geography 17 (5), 385–395.
Washington, S., Karlaftis, M., Mannering, F., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & HALL/CRC.
Xu, C., Liu, P., Wang, W., Li, Z., 2012a. Evaluation of the impacts of traffic states on crash risks on freeways. Accident Analysis and Prevention 47, 162–171.
Xu, F., Tian, Z., 2008. Driver behavior and gap-acceptance characteristics at roundabouts in California. Transportation Research Record 2071, 117–124.
Xu, C., Liu, P., Wang, W., Li, Z., 2012b. Development of a crash risk index to identify real-time crash risks on freeways. In: Presented at the 91th Annual Meeting of the Transportation Research Board, CD-ROM, Washington, DC.
Xu, C., Wang, W., Liu, P. A genetic programming model for real-time crash prediction on freeways. IEEE Transactions on Intelligent Transportation Systems, in press.
Yamamoto, T., Hashiji, J., Shankar, V., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis and Prevention 43, 1320–1329.
Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis and Prevention 51, 252–259.
Zheng, Z., Ahna, S., Monsere, C., 2010. Impact of traffic oscillations on freeway crash occurrences. Accident Analysis and Prevention 42, 626–636.