How to determine an optimal threshold to classify real-time crash-prone traffic conditions?

How to determine an optimal threshold to classify real-time crash-prone traffic conditions?

Kui Yang[1, 2]
Rongjie Yu[1, 2]*
Xuesong Wang[1, 2]
Mohammed Quddus[3]
Lifang Xue[4]

[1]School of Transportation Engineering
Tongji University
4800 Cao'an Road, 201804, Shanghai, China

[2]The Key Laboratory of Road and Traffic Engineering, Ministry of Education
4800 Cao'an Road, 201804, Shanghai, China

[3]School of Civil and Building Engineering
Loughborough University
Loughborough LE11 3TU
United Kingdom

[4] College of Arts and Sciences
Shanxi Agricultural University
No. 1 Mingxian Nan Road, 030801, Taigu, Shanxi, China

* Corresponding Author
yurongjie@tongji.edu.cn

March 2018

Abstract: One of the proactive approaches in reducing traffic crashes is to identify hazardous traffic conditions that may lead to a traffic crash, known as real-time crash prediction. Threshold selection is one of the essential steps of real-time crash prediction. And it provides the cut-off point for the posterior probability which is used to separate potential crash warnings against normal traffic conditions, after the outcome of the probability of a crash occurring given a specific traffic condition on the basis of crash risk evaluation models. There is however a dearth of research that focuses on how to effectively determine an optimal threshold. And only when discussing the predictive performance of the models, a few studies utilized subjective methods to choose the threshold. The subjective methods cannot automatically identify the optimal thresholds in different traffic and weather conditions in real application. Thus, a theoretical method to select the threshold value is necessary for the sake of avoiding subjective judgments. The purpose of this study is to provide a theoretical method for automatically identifying the optimal threshold. Considering the random effects of variable factors across all roadway segments, the mixed logit model was utilized to develop the crash risk evaluation model and further evaluate the crash risk. Cross-entropy, between-class variance and other theories were employed and investigated to empirically identify the optimal threshold. And K-fold cross-validation was used to validate the performance of proposed threshold selection methods with the help of several evaluation criteria. The results indicate that (i) the mixed logit model can obtain a good performance; (ii) the classification performance of the threshold selected by the minimum cross-entropy method outperforms the other methods according to the criteria. This method can be well-behaved to automatically identify thresholds in crash prediction, by minimizing the cross entropy between the original dataset with continuous probability of a crash occurring and the binarized dataset after using the thresholds to separate potential crash warnings against normal traffic conditions.

Keywords: Urban expressway safety management; Crash risk evaluation; Mixed logit model; Threshold selection method; Cross-entropy; Between-class variance

1 Introduction

Given the technological progress over the last decade in traffic data detection, storage and mining, real-time crash prediction has become a popular research topic within the safety community. Crash risk evaluation and threshold selection are the two essential steps of real-time crash prediction. Crash risk evaluation is related to the investigation of a relationship between the crash occurrence and geometric characteristics, real-time traffic flow parameters, weather conditions. The relationship is used to evaluate the probability of a crash occurring given a specific traffic condition and identify hazardous traffic conditions. The threshold selection is to investigate the algorithm to select the optimal cut-off point of the posterior probability ($0<$ posterior probability<1, also known as crash risk), which is used for separating potential crash

warnings from normal traffic conditions, and further triggering Active Traffic Management (ATM) control strategies (Abdel-Aty et al., 2006). Previous studies mostly focused on crash risk evaluation, with the purpose of identifying impact factors on crash occurrence in order to further understand the crash mechanisms (e.g. Abdel-Aty et al., 2005; Yu & Abdel-Aty, 2013a), modeling technique aimed at better classification accuracy (e.g. Abdel-Aty & Pande, 2005; Yu & Abdel-Aty, 2013b; Xu et al., 2013a). However, there is a dearth of study focused on how to determine a reliable threshold for real-time crash prediction.

In real-time crash prediction and its application in ATM, an appropriate threshold should be used to compare with the estimated probability which is the output of the crash risk evaluation model on the basis of real-time traffic data. If the probability exceeds the predetermined threshold, the case is predicted as a *potential crash scenario*, and then a crash warning is alerted, and further control strategies are triggered (Abdel-Aty et al., 2010). In addition, there is a dilemma in selecting the correct threshold (which is a metric ranging from 0 to 1) because a high threshold normally fails to identify many potential crash conditions whereas a low threshold falsely identifies normal traffic conditions as 'hazardous'. False alarms may affect the driver's compliance and raise the cost of ATM operations.

Only a few existing studies on real-time crash prediction models involved identifying a threshold when discussing the predictive performance of the models, and their methods are subjective. Moreover, the subjective approach cannot automatically identify the optimal thresholds in different traffic and weather conditions, aimed at capturing the temporal-spatial heterogeneity of crashes which has been proved to exist by researchers (e.g. Xu et al., 2013b; Yu et al., 2016). Additionally, in order to avoid subjective judgments, a theoretical method to select the threshold value is necessary (Li & Tzeng, 2009). In other fields of pattern recognition (e.g. image segmentation, medical field), different methods of the threshold selection have been investigated for more than half a century (Zhang, 2014). Their methods promoted the progress of technology in studies and applications.

This study aims to explore available threshold selection methods so as to automatically identify the optimal threshold for real-time crash prediction. Cross-entropy, between-class variance and other theories were utilized and their performances were compared on the basis of several evaluation criteria. Traffic data and historical crash data from Shanghai Urban Expressway System were used in this analysis. Considering the random effects of variable factors across all roadway segments, mixed logit model was employed to develop the crash risk evaluation model and further evaluate the crash risk. Different thresholds were selected by five threshold selection methods, and their predictive performances were further evaluated through different evaluation criteria. Besides, 5-fold cross-validation was used to test the thresholds for deriving a more accurate estimate of prediction performance.

The rest of this paper is divided into six sections. First, previous studies on threshold selection in the real-time crash prediction and other fields are summarized. The second section describes the

study area. The third section describes the data preparation procedures. Afterwards, the modeling techniques, threshold selection methods and several evaluation criteria are introduced. The fifth section presents the modelling and comparison results of different threshold selection methods. Finally, discussion and conclusions of this work are provided.

2 Literature review

2.1 Threshold selection involved in crash risk evaluation

In existing studies on real-time crash prediction, there is a dearth of research that focuses on threshold selection method. And only when discussing the prediction performance of the models, a few studies tried to choose the threshold subjectively.

When a matched case-control logistic regression model was used to develop a crash risk evaluation model, the average values of the explanatory variables associated with all non-crash cases within each matched stratum were calculated as the “normal traffic conditions”, and the odds ratio of each case relative to “normal traffic conditions” within each stratum was used as the crash risk index. Thus, the odds ratio with a value equal to one was selected as the threshold (e.g. Abdel-Aty et al., 2005; Ahmed & Abdel-Aty, 2012), which is known as fixed threshold based on the odds ratio. Similar to posterior probability, the odds ratio is an index indicating the crash risk level relative to non-crashes, and thus a value greater than one is regarded subjectively as “more hazardous” than “normal traffic conditions” by fixed threshold based on the odds ratio. Moreover, the method will not work if the modeling technique is not a matched case-control logistic regression model. Therefore, fixed threshold based on the odds ratio has the uniqueness of crash risk evaluation model development technique and the fixed threshold, which creates its limitations in range of application. And it has been not accepted by all researchers.

Due to the imbalance of the proportions of crash and non-crash cases in the sample, overall classification accuracy over validation dataset would not be a good measure for model performance evaluation. Therefore, the top 30 percentile of posterior probability (i.e. first three deciles) was decided as the threshold (Pande & Abdel-Aty, 2006a; Pande & Abdel-Aty, 2006b). Similarly, Pande et al. (2011) chose the top 20 percentile of posterior probability as the threshold. Since the threshold is mainly affected by the proportion of crashes in samples, different samples with the same proportion of crashes but different characteristic distributions cannot select obviously different thresholds.

Aimed at balancing the predictive accuracy of crash and non-crash, the cut-off point where the predictive accuracy of crashes was equal to that of non-crashes (i.e. the intersection of cumulative proportion curves of crash and non-crash cases), was chosen as the threshold by the intersection point method (e.g. Xu et al., 2013b). But the predictive accuracies of crashes and non-crashes among different models or strategies lack some comparability because of different weighting scores.

Totally, existing three types of threshold selection methods have no specific mathematical optimization functions. And it creates a certain amount of subjectivity and the limitations in application. Moreover, they are sensitive to a fraction of specific sample and cannot absorb all distribution information of the dataset. Thus, a theoretical method to select the threshold value is necessary so as to avoid subjective judgments.

2.2 Threshold selection in other fields

Threshold selection techniques are fundamental for image segmentation. Different techniques, ranging from a bilevel threshold selection with a single threshold to a multilevel threshold selection with multiple thresholds dividing pixels into categories, have been proposed (Yin, 2007).

The first category, known as histogram shape-based method, contains the approaches which determine the optimal threshold by analyzing the profile characteristics of the gray-level histogram of pixel in image, which is the basic information for image thresholding. With decades of experiments and applications, bimodal histogram threshold method (Weszka et al., 1974) and P-tile method (Doyle, 1962; Samopa & Asano, 2009) could achieve better binarized image, and became the two widely used shape-based methods. Bimodal histogram threshold selects the cut-off point of gray-level of pixel at valleys between two peaks as the threshold, while P-tile method requires the proportion of the object after being binarized will not be less than that in the original dataset.

The second category, known as optimization method, belongs to the techniques which determine the optimal threshold by optimizing a certain objective function or theory, which use some extra information such as spatial information or binarized images:

(i) Entropy-based methods: Maximum entropy method (Pun, 1980; Kapur et al., 1985) and minimum cross-entropy method (Yin, 2007; Li & Lee, 1993) can achieve better binarized image, and became the two widely used methods. Maximum entropy method maximizes the entropy of foreground and background regions, while the minimum cross-entropy method minimizes the cross-entropy between the original and binarized image.

(ii) Variance-based methods: the maximum between-class variance method (Otsu, 1979) selects an optimal threshold by maximizing the separability of the resultant classes in gray levels for image segmentation. And it is one of the best threshold selection methods for general real-world images.

Since the gray-levels of all pixels in the image are used to identify the optimal threshold, the two categories can absorb all distribution information. And they have objective optimization theory, such as the entropy and the variance. The key of image segmentation is to select the threshold to turn a gray-level image into a binary image, and even select multilevel threshold when multiple-levels are needed. Similarly, the purpose of threshold selection in real-time crash prediction is also to select the optimal cut-off point (i.e. threshold), to divide the probability into binary values

presenting crash and non-crash scenarios. Therefore, threshold selection methods achieving good performances in other fields (i.e. bimodal histogram threshold method, P-tile method, maximum entropy method, minimum cross-entropy method, and maximum between-class variance method) could be utilized to identify a threshold in real-time crash prediction, and further overcome the disadvantages of existing three types of methods.

3 Study area

The study was conducted on the Shanghai Urban Expressway System in China. The expressway stretch used for the analysis is 236 km that includes Yan'an elevated road, North-South elevated road, Middle ring elevated road, Inner ring elevated road, Yixian elevated road, and Humin elevated road. As shown in **Fig. 1**, Yan'an elevated road is east–west, and North–South elevated road has a north–south trajectory; Yixian elevated road and Humin elevated road are radial; the Inner ring and the Middle ring elevated roads are two loops that cover mostly the urban areas of Shanghai. Moreover, the expressway system in Puxi area has about 238 roadway segments separated by adjacent ramps, and the average length is about 949 meters.

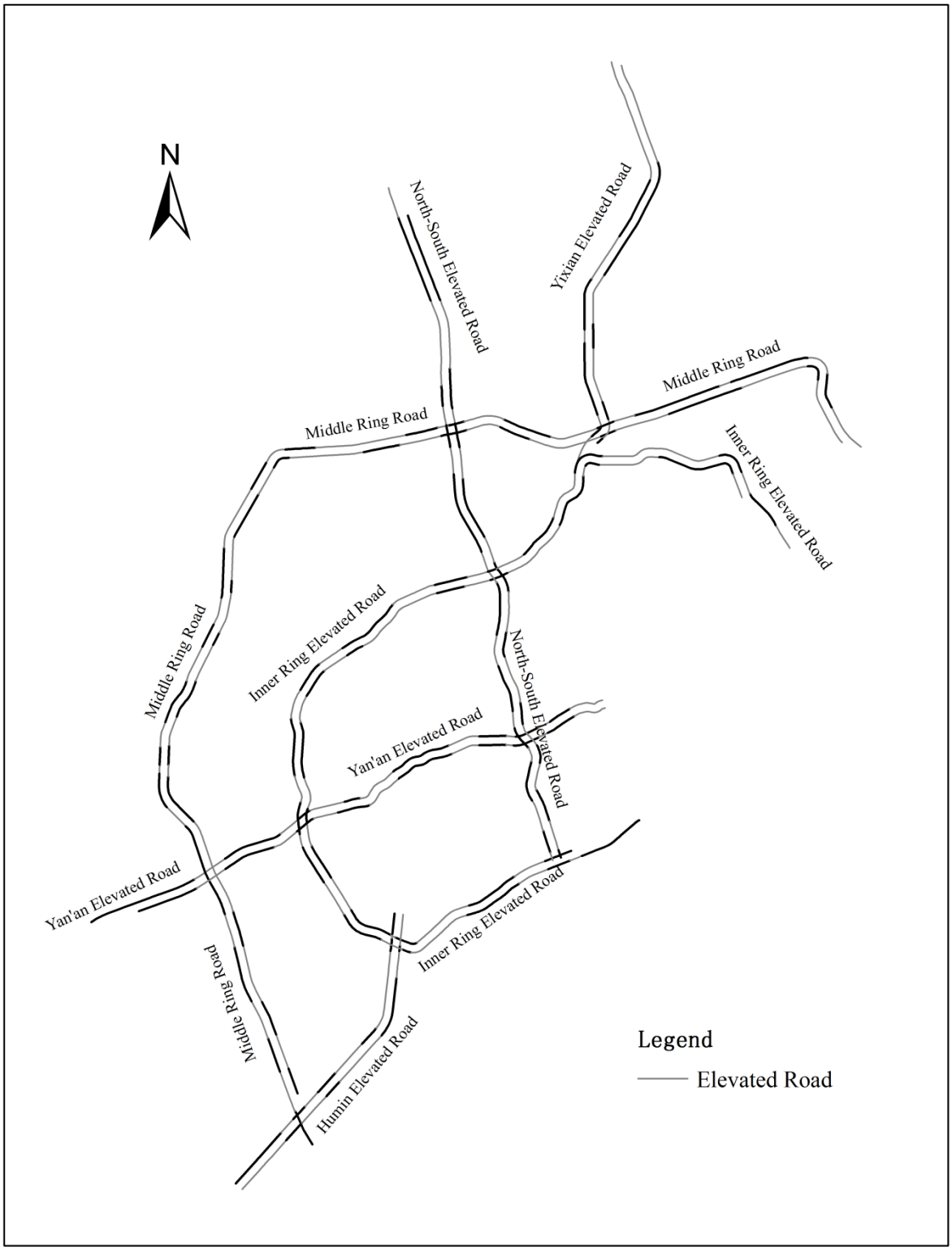

Fig. 1. Shanghai urban expressway network (Puxi area).

A total of 1,947 crashes occurred on the study corridor in September 2013, including 10 single-vehicle crashes, 1,735 two-vehicle crashes and 202 multi-vehicle crashes. Crash data were obtained from the Shanghai Traffic Information Center. The recorded time of a crash occurring is the time of capturing the scene of a crash with the video surveillance system. Due to the wide-spread use of mobile phones and video surveillance systems, inaccuracy between recorded time and actual time of a crash occurring is minimal. In general, the error about crash reporting time is less than 2 minutes and almost impossible to reach 5 minutes. The locations of crashes were described by roadway segment ID, which is the digital ID consisting of 11 numbers for each roadway segment.

Dual inductive Loops Detectors (LDs) were installed in each lane of a roadway segment within the studied urban expressways. Traffic data including segment-based average speed and total traffic volume at 2-minute interval in each roadway segment from LDs, were obtained from the Shanghai Traffic Information Center. The small aggregation traffic data (e.g. 10-second, 20-second, 30-second raw and 2-minute loop detector data), contain measurement noises and useless traffic fluctuation information (Abdel-Aty et al. 2005; Pande et al. 2011; Xu et al. 2013b). For the purpose of reducing the random noise and obtaining averages and standard deviations (Abdel-Aty et al. 2005; Pande et al. 2011), most of the studies on real-time crash risk evaluation used higher aggregated data, such as 5-minute aggregated data (e.g., Abdel-Aty et al. 2005; Pande et al. 2011; Yu & Abdel-Aty, 2013b; Yu et al., 2016), 6-minute aggregated data (e.g., Ahmed & Abdel-Aty, 2013; Yu & Abdel-Aty, 2014). Thus, 6-minute aggregated data were obtained and used for the crash risk analyses in this study as follow the literature.

Due to the network layout of the expressway system, the function of each urban expressway is different (Shanghai City Comprehensive Transportation Planning Research Institute, 2006). The functions of Yan'an elevated road and North-South elevated road are mostly to take on the cross-border traffic in east–west and north–south directions. The functions of Inner ring elevated road and Middle ring elevated road are essentially to attract cross-border traffic passing through the city center and connect the sub-CBDs. The functions of Yixian elevated road and Humin elevated road are mostly to connect the central urban area with suburbs. Besides, the proportion of the motor vehicle with non-shanghai license plate on every urban expressway is different, ranging from 15% to 25% (i.e., Yan'an elevated road: 15%; North-South elevated road: 17%; Inner ring elevated road: 15%; Middle ring elevated road: 25%; Yixian elevated road: 20%; Humin elevated road: 18%) (Shanghai City Comprehensive Transportation Planning Institute, 2016). Moreover, the proportions of the vehicles with different sizes on every urban expressway are different (see **Table 1**). The proportion of the medium-sized vehicle on Yan'an elevated road takes up to 26.6%, and the proportion of the large-sized vehicle on Humin elevated road takes up to 6.5%. Thus, the driver behaviors on different urban expressways are different due to the different functions and the different socioeconomic characteristics (e.g. education level, job, and driving experience) of driver from other cities or different vehicle types. For instance, different urban expressways have different operation speeds (see **Fig. 2**), which is the core of driver

behaviors (Evans, 2004). Therefore, it may be unrealistic to assume that the effects of variables (e.g. traffic flow, roadway geometrics, and other factors) are the same across all roadway segments due to variations in driver behavior or traffic composition.

Table 1 The proportion of the vehicles with different sizes on Shanghai urban expressways.

Road	Large-sized Vehicle	Medium-sized Vehicle	Small-sized Vehicle
Yan'an elevated road	5.2%	26.6%	68.3%
North-South elevated road	1.4%	2.2%	96.4%
Inner ring elevated road	1.5%	4.6%	93.8%
Middle ring elevated road	1.7%	4.5%	93.8%
Yixian elevated road	2.1%	2.5%	95.4%
Humin elevated road	6.5%	9.8%	83.7%

* Large-sized vehicle: the length >9.5m; medium-sized vehicle: 9.5m≥the length >5.5m; small-sized vehicle: the length ≤5.5m;

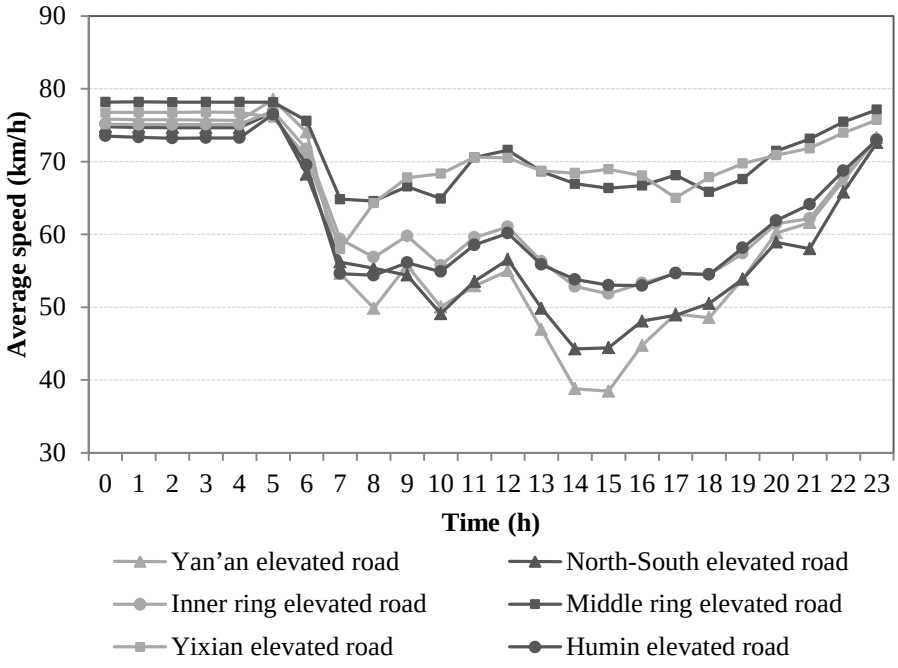

Fig. 2. The temporal distribution of the average speed on different urban expressways.

4 Data preparation

Three datasets from Shanghai Urban Expressway System were used to conduct the analysis: (i) crash data for September 2013; (ii) roadway segment geometry data in ArcMap (ESRI, 2006); (iii) real-time traffic data detected by LDs along the urban expressways. Due to the use of roadway segment ID to describe the location of each crash, roadway segment geometry data in ArcMap were employed to identify the upstream segment and the downstream segment of crash data.

The approach for developing real-time crash risk evaluation models is to analyze historical crashes and traffic surveillance data corresponding to historical crashes and try to detect crash precursor that are often observed before crash occurrence (Abdel-Aty et al., 2010). As we "approach" the time and location of the crash, the statistical significance of variables and the crash risk tend to increase (Abdel-Aty & Pande, 2005). Besides, 30 minutes is enough for detecting a crash prone condition and triggering proactive traffic management strategies to reduce the crash risk. Moreover, since the impacts of the traffic state on the crash occurrence decrease with the increase in the time before the crash occurs (e.g. Xu et al., 2012), a 5-minute time period ending 30 min before the reported crash time had few impacts and was selected as the normal traffic condition (Oh et al., 2001). Thus, most of the studies on real-time crash risk evaluation used the variables for half an hour period prior to the crash occurrence (e.g. Abdel-Aty & Pande, 2005; Abdel-Aty et al. 2005; Pande et al. 2011; Yu et al., 2016). Therefore, three segments (i.e. upstream segment, crash segment and downstream segment) and five time slices with 6-minute interval (30-minute time window in total) prior to the crash occurrence, were identified with the help of ArcMap in this paper for the purpose of extracting traffic data prior to the crash occurrence. This three segments were named as U, C and D respectively, and five time slices were named as TS1, TS2, TS3, TS4, and TS5 with TS1 being the 0-6 minutes just prior to the crash reporting time (**Fig. 3**). For example, for the crash occurred at 09:35 on September 23rd, traffic data from 09:04 to 09:34 (i.e. a 30-minute window) were extracted for each of the three segments and were named as TS1, TS2, TS3, TS4, and TS5 (**Fig. 3**), with TS1 being the time period between 09:28 and 09:34.

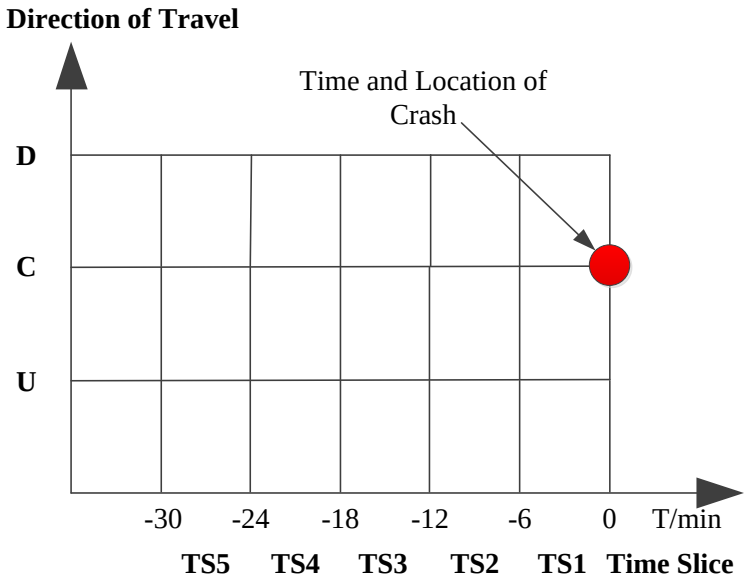

Fig. 3. Nomenclature for defining stations and time slices relative to the location and time of a crash.

In addition, for the purpose of comparing normal traffic conditions with traffic characteristics prior to the crash occurrence, matched case-control study was designed, and traffic data of non-crash cases were also extracted. It should be noted that traffic data of a non-crash scenario were the "normal traffic conditions" where no crashes were observed within the 1-hour window, given the same time of the day, day of the week, month and location but in different weeks. The matched case-control analysis allows us to control the external factors, such as time of day, season, geometric, roadway features and driver population on the freeway (e.g. more commuters on weekday peak hours, indicating more young to middle age drivers, etc.) (Abdel-Aty & Pande, 2005; Yu & Abdel-Aty, 2013b). It was frequently utilized in the disaggregate crash occurrence studies since the confounding external factors can be controlled for by matching (Yu & Abdel-Aty, 2013b). Abdel-Aty et al. (2004) found that no significant differences had been observed when changing the number of non-crashes (m), by analyzing separately each time duration data for each matched data set (1:m, m=1, 2, 3, 4 and 5). As for this study, a 1:4 ratio of crashes to non-crashes was used which followed the suggestion provided by most of the previous studies (e.g. Rothman and Greenland, 1998; Yu & Abdel-Aty, 2013b; Xu et al., 2012).

For example, for the crash that occurred at 09:35 on September 23rd (Monday), the traffic conditions at the same segment and time in September 30th (Monday), September 16th (Monday), September 9th (Monday) and September 2nd (Monday) were collected as non-crash cases. The final dataset has 1,425 matched strata with 1,425 crashes and 3,974 non-crashes. It should be notable that the final dataset cannot meet exact 1:4 ratio of crashes to non-crashes due to the unavailability of LDs data. But it was still available and used in this research since no

significant differences had been observed when changing the number of non-crashes (Abdel-Aty et al., 2004) and the scale of the dataset might be a bit small if only a 1:2 or 1:4 ratio was used.

On the basis of the collected traffic data for crash and non-crash cases, average speed, total volume, standard deviation of speed, and standard deviation of volume between three 2-minute intervals were calculated at the 6-minute interval for every different road segments and time slices. This resulted in a total of 60 explanatory variables (i.e. 4 Variables x 3 Segments x 5 Time Slices), which can be used in the crash risk evaluation model.

The final dataset was used to train the crash risk evaluation model based on mixed logit model. And then, it was used to select the thresholds by several methods, whose performances were validated by several evaluation criteria with the help of K-fold cross-validation.

5 Methodology

In this methodology section, the modeling techniques for crash risk evaluation, threshold selection methods and several evaluation criteria are introduced. The flowchart of introducing methods in the methodology section is shown as **Fig. 4**.

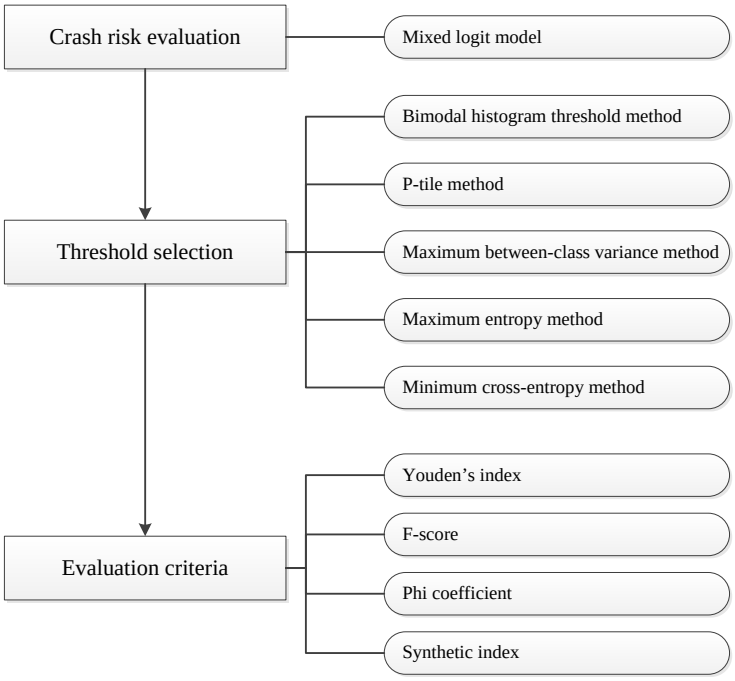

Fig. 4. A flowchart of introducing methods in the methodology section.

5.1 Crash risk evaluation

Mixed logit model was employed to develop the crash risk evaluation model so as to evaluate the crash risk (i.e. the probability of a crash occurring) given a specific traffic condition. It is important to apply a methodological approach that allows for the possibility that the influence of

variables affecting crash risk may vary across roadway segments. This is an important consideration because, due to probable variations in driver behavior or traffic composition for example, it may be unrealistic to assume that the effects of variables (e.g., traffic flow, roadway geometrics, and other factors) are the same across all roadway segments.

Mixed logit models assume a utility function U_{in} conformed by a deterministic component V_{in}, a random component ε_{in} independent and identically distributed, and one or more additional random terms. These additional error terms can be grouped together in an additive term η_{in}, that can be function of the data (attributes of alternatives), and that potentially models the presence of correlation and heteroscedasticity. So, the utility function (Munizaga & Alvarez-Daziano, 2001) is defined as

$$U_{in} = V_{in} + \eta_{in} + \varepsilon_{in} \quad (1)$$

where $V_{in} = \beta_n' X_{in}$, indicating that the deterministic component of the utility is linear in the β parameters that multiply the attributes X_{in}; η_{in} is a random term with zero mean whose distribution over individuals and alternatives depends in general on underlying parameters and observed data relating to alternative i and individual n; and ε_{in} is a random term with zero mean that is IID over alternatives and does not depend on underlying parameters or data.

The mixed logit class of models assumes a general distribution for η_{in} and an IID extreme value type 1 distribution for η_{in}. Denote $\eta_{in} \sim f(\eta_{in}|\boldsymbol{\Omega})$ where $\boldsymbol{\Omega}$ are the fixed parameters of the distribution. For a given value of η_{in}, the conditional probability for the choice i is logit, since the remaining error term is IID extreme value (Hensher & Greene, 2003):

$$L_{in}(\beta_n|\eta_{in}) = \frac{\exp(\beta_n' X_{in} + \eta_{in})}{\sum_j \exp(\beta_n' X_{jn} + \eta_{jn})} \quad (2)$$

The unconditional choice probability would be this logit formula integrated over all values of η_{in} weighted by the density of η_{in} as shown (Hensher & Greene, 2003):

$$P_{in}(\beta_n|\boldsymbol{\Omega}) = \int_{\eta_{in}} L_{in}(\beta_n|\eta_{in}) f(\eta_{in}|\boldsymbol{\Omega})\, \eta_{in} \quad (3)$$

Models of this form are called *mixed logit* because the choice probability P_{in} is a mixture of logits with f as the mixing distribution. Besides, the mixed logit model was developed in Version 13.0 of STATA (StataCorp, 2013).

5.2 Threshold selection

Threshold selection is a key step of real-time crash prediction and its application in Active Traffic Management. It could be defined as:

$$\Pr edictedCrash = \begin{cases} 1 & if\ CrashRisk \geq Threshold \\ 0 & if\ CrashRisk < Threshold \end{cases} \quad (4)$$

where $PredictedCrash$ denotes the predictive result of a case; it is predicted as a crash when $PredictedCrash = 1$, while it is predicted as a non-crash when $PredictedCrash = 0$. $CrashRisk$ denotes the crash risk (probability) evaluated by the crash risk evaluation model, and it ranges from 0 to 1.

Let a histogram of the crash risk in a dataset be represented by $n_1, n_2, \cdots, n_i, \cdots, n_L$, where n_i is the number of cases at level i, and L is the number of distinct levels. And the levels are $T_1, T_2, \cdots, T_k, \cdots, T_L$, and $0 \leq T_1 < T_2 < \cdots < T_k < \cdots < T_L \leq 1$. From the histogram, the probability of occurrence of the level i is defined as follows:

$$p_i = \frac{n_i}{N}$$

where N is the total number of cases in a dataset, that is, $N = \sum_{i=1}^{L} n_i$.

A total of five methods were used for the threshold selection, and introduced as follows.

A) Bimodal histogram threshold method

Similar to histogram shape-based methods in image segmentation (Weszka et al., 1974), in an ideal case, the crash risk histogram of dataset has a deep and sharp valley between two peaks representing crash and non-crash scenarios (**Fig. 5**), respectively. The histogram exhibits a bimodal distribution in which the peak on the right represents crash events while the peak on the left represents non-crash events. Therefore, the cut-off point of crash risk at the bottom of the valley is an acceptable threshold to separate the two distinct histograms maximally completely, which is known as bimodal histogram threshold method. In this paper, the parametric technique using the curve fitting was firstly used to fit the histogram of the crash risk, and then the optimal threshold with the minimum (i.e., bottom) between the two local maxima (i.e., peaks) was selected as the threshold.

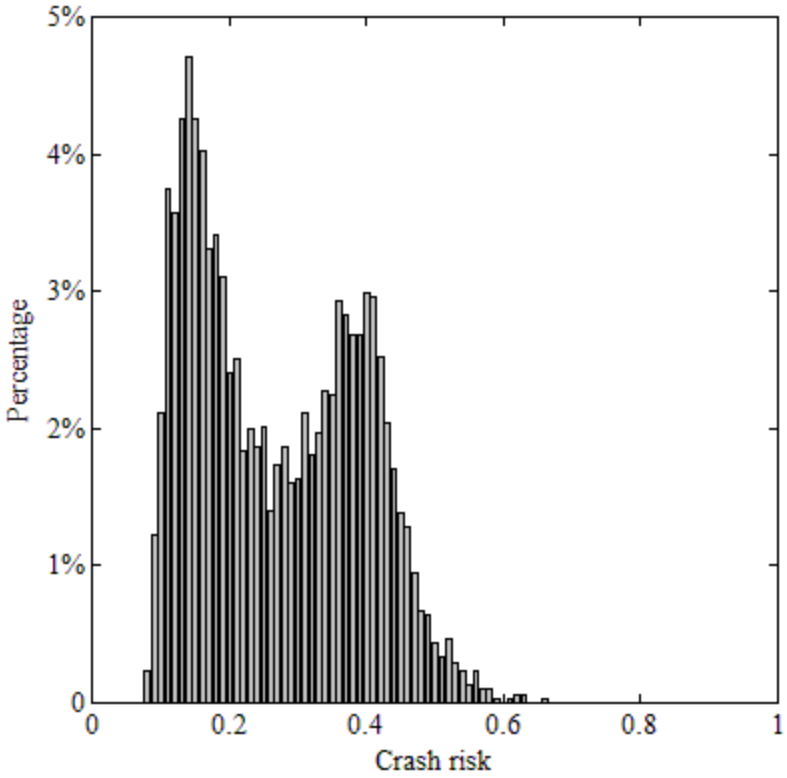

Fig. 5. Crash risk histogram.

B) P-tile method

P-tile method was a premature threshold selection method based on the gray-level histogram for image segmentation techniques (Doyle, 1962; Samopa & Asano, 2009). The method requires the proportion of the object would not be less than that in the original dataset. Similarly, if the proportion of the object (crash or non-crash) is known as P%, the cut-off point of crash risk where the cumulative proportion of the object is equal to P% could be selected as the threshold.

C) Maximum between-class variance method

The maximum between-class variance method, which is also known as Otsu method (Otsu, 1979), selects an optimal threshold by maximizing the separability of the resultant classes in discriminant analysis. Let a dataset with the probability resulting in a crash be portioned into two classes (C_0 (non-crash) and C_1 (crash)) by a threshold T_k. The between-class variance σ_B^2 is shown as

$$\sigma_B^2 = \rho_0(\mu_0 - \overline{\mu})^2 + \rho_1(\mu_1 - \overline{\mu})^2 = \rho_0\rho_1(\mu_0 - \mu_1)^2 \quad (5)$$

where μ_0 denotes the expectation of crash risk with levels between T_1 and T_k (class C_0), and μ_1 denotes the expectation of crash risk with levels between $T_{k+1}$and T_L (class C_1); $\overline{\mu}$ denotes the expectation of crash risk with levels between $T_1$and T_L (total dataset). $\rho_0 = \sum_{i=1}^{k} p_i$ denotes the proportion of cases in class C_0, and $\rho_1 = \sum_{i=k+1}^{L} p_i$ denotes the proportion of cases in class C_1.

Then the method seeks an optimal threshold T_{k^*} by maximizing σ_B^2 as follows:

$$T_{k^*} = \arg\max_{1 \le k \le L} \left\{ \sigma_B^2 \right\} \tag{6}$$

where argmax is an operator that finds a maximum point for a function and its complementary operator is argmin.

D) Maximum entropy method

The concept of entropy in information theory was first applied to picture segmentation by Pun (1980), and was improved by Kapur et al. (1985). Divide the distribution of crash risk into two probability distributions (D_0 and D_1) by a threshold T_k, and the entropy associated with each distribution are shown as

$$E_0(T_k) = -\sum_{i=0}^{k} \frac{p_i}{\rho_0} \ln \frac{p_i}{\rho_0} \tag{7}$$

$$E_1(T_k) = -\sum_{i=k+1}^{L} \frac{p_i}{1-\rho_0} \ln \frac{p_i}{1-\rho_0} \tag{8}$$

where $\rho_0 = \sum_{i=1}^{k} p_i$ denotes the probability of crash risk with levels between T_1 and T_k (distribution D_0), and $1 - \rho_0$ denotes the probability of crash risk with levels between $T_{k+1}$and T_L (distribution D_1).

Then the optimal threshold T_{k^*} is achieved through the following formulae:

$$T_{k^*} = \arg\max_{1 \le k \le L} \left\{ E_0(T_k) + E_1(T_k) \right\} \tag{9}$$

E) Minimum cross-entropy method

Minimum cross-entropy method, namely minimum cross-entropy thresholding by Li and Lee (1993), selects the threshold by minimizing the cross-entropy between the original dataset and its thresholded dataset. The cross-entropy $CE(T_k)$ is defined as

$$CE(T_k) = \sum_{i=1}^{k} T_i p_i \ln(\frac{T_i}{\mu_0}) + \sum_{i=k+1}^{L} T_i p_i \ln(\frac{T_i}{\mu_1}) \tag{10}$$

where μ_0 and μ_1 are the same with those in Equation (5) respectively. μ_0 denotes the expectation of crash risk with levels between T_1 and T_k, and $\mu_1$denotes the expectation of crash risk with levels between $T_{k+1}$and T_L.

Therefore, the optimal threshold T_{k^*} is

$$T_{k^*} = \min_{1 \le k < L} \{CE(T_k)\} \tag{11}$$

5.3 Evaluation criteria

Youden's index, F-score, Phi coefficient and their synthetic index were used as criteria to evaluate the performances of five proposed threshold selection methods. The confusion matrix is firstly shown as **Table 2** before introducing the criteria.

Table 2 Confusion Matrix.

Actual input	Predicted outcome	
	Crash(1)	Non-crash(0)
Crash(1)	True Positives (TP)	False Negatives (FN)
Non-crash(0)	False Positives (FP)	True Negatives (TN)

In the confusion matrix, TP denotes the number of crashes correctly predicted by the model as crashes, FP denotes the number of non-crashes wrongly predicted by the model as crashes, FN denotes the number of crashes wrongly predicted by the model as non-crashes, and finally TN denotes the number of non-crash correctly predicted by the model as non-crashes.

A) Youden's index

Youden's index was suggested by Youden (1950) as a way of summarizing the performance of a diagnostic test. And its definition can be shown as:

$$J = Sencitivity + Specificity - 1 \tag{12}$$

where $Sencitivity$ denotes the proportion of crashes correctly predicted by the model as crashes ($i.e. Sencitivity = TP/(TP + FN)$), and $Specificity$ denotes the proportion of non-crashes correctly predicted by the model as non-crashes ($i.e. Specificity = TN/(TN + FP)$).

Youden's index ranges from 0 to 1. A value of 0 indicates that the threshold is useless, and a value of 1 indicates that all crashes are correctly predicted as crashes and all non-crashes are correctly predicted as non-crashes. Thus, the higher Youden's index indicates that the threshold could achieve better predictive performance.

B) F-score

F-score is often used in the field of information retrieval (van Rijsbergen, 1979), machine learning and the natural language processing literature for classification performance. Its definition is shown as:

$$Fscore = \frac{2 * \mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \quad (13)$$

where $Precision$ denotes the proportion of crashes predicted as crashes correctly (i. e. $Precision = TP/(TP + FP)$), and $Recall$ denotes the proportion of non-crashes predicted as non-crashes correctly (i. e. $Recall = TN/(TN + FP)$). The bigger the F-score is, the better predictive performance is.

C) Phi coefficient

Phi coefficient was introduced by Karl Pearson (Cramir, 1946), and described as:

$$\varphi = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(FN + TN)(TP + FN)(FP + TN)}} \quad (14)$$

where TP, FP, FN, TN are the same as shown in Table 2. The bigger the Phi coefficient is, the better the predictive performance is.

D) Synthetic index

In order to identify the best method on the basis of three different indexes, a synthetic index (SI) was proposed. It is the average value of these three indexes with the same weight after zero-mean normalization. Zero-mean normalization is to fit the data within unity (one) for each evaluation criteria, through making the difference between original value and mean of variable be divided by standard deviation. Its definition is shown as:

$$SI = \frac{1}{3}\left(\frac{J - \bar{J}}{SD(J)} + \frac{Fscore - \overline{Fscore}}{SD(Fscore)} + \frac{\varphi - \bar{\varphi}}{SD(\varphi)}\right) \quad (15)$$

where $\bar{J}, \overline{Fscore}, \bar{\varphi}$ are, respectively, the mean values of the three indexes, and SD(J), $SD(Fscore), SD(\varphi)$ are, respectively, standard deviation of the three indexes. The bigger the synthetic index is, the better the predictive performance of the method is.

6 Analysis results

6.1 Crash risk evaluation model

Similar to most of the previous studies (e.g. Abdel-Aty & Pande, 2005; Abdel-Aty et al., 2005; Ahmed & Abdel-Aty, 2012; Yu et al., 2016), the variables from TS1 were not considered as inputs. 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; Xu et al., 2013a). Five variables were found to be significantly associated with the crash occurrence, and the correlation effects among the five variables were checked. **Table 3** presents summary statistics for the variables.

Table 3 Summary Statistics of the Variables included in the Final Model.

Variable	Definition	Mean	Std. Dev.	Min	Max
TVU2	Total Volume for U section during 6-12min prior to crash occurrence (pcu/6min)	371.48	152.56	0	967.00
SVU2	Std. Dev. of Volume for U section during 6-12min prior to crash occurrence (pcu/6min)	10.57	7.50	0	63.89
ASC2	Average Speed for C section during 6-12min prior to crash occurrence (km/h)	44.78	20.98	1.00	85.33
SSC2	Std. Dev. of Speed for C section during 6-12min prior to crash occurrence (km/h)	2.89	2.62	0	26.10
SVD2	Std. Dev. of Volume for D section during 6-12min prior to crash occurrence (pcu/6min)	10.38	8.24	0	96.77

For the purpose to consider that the effects of variables (e.g., traffic flow, roadway geometrics, and other factors) are not the same across all roadway segments, mixed logit model is employed to develop a crash risk evaluation model. The mixed logit specification shown in Equation (3) was estimated with simulation-based maximum likelihood. And the model results are listed in **Table 4**. Variables are all from TS2, which is the time period of 6-12 min prior to the recorded time of the crash. Similarly, some previous studies just only employed the variables during 5-10 minutes (e.g. Xu et al., 2013b; Xu et al., 2014; Xu et al., 2015; Yu & Abdel-Aty, 2013a; Yu & Abdel-Aty, 2013b) or 6-12 minutes (e.g. Ahmed & Abdel-Aty, 2013; Yu & Abdel-Aty, 2014) prior to crash occurrence to develop the crash risk evaluation model for the purpose of identifying hazardous traffic condition ahead of the crash occurrence time to make preemptive measures possible (Pande et al., 2011; Xu et al., 2013a). It is noticeable that all variables are statistically significant, which shows they are important predictors.

Table 4 Parameters Estimates of Mixed Logit Model.

Variable	Parameter	Parameter Estimate	Standard Error	z	Pr>|z|
TVU2					
	Mean of coefficient	0.0015697	0.0006441	2.44	0.015
SVU2					
	Mean of coefficient	0.014429	0.0061267	2.36	0.019
	Std. dev. of coefficient*	0.0590554	0.0206252	2.86	0.004
ASC2					
	Mean of coefficient	-0.0649187	0.0052563	-12.35	0.000
	Std. dev. of coefficient*	0.0201141	0.0094824	2.12	0.034
SSC2					
	Mean of coefficient	0.0599981	0.01404	4.27	0.000
SVD2					
	Mean of coefficient	0.0112019	0.004974	2.25	0.024
	Iteration	5			
	Number of obs	5,399			
	LR chi2 (2)	5.44			

	Log likelihood	-1543.9057

*standard deviation

Among the five variables, TVU2 and SVU2 have a positive sign, which means that larger volume and traffic volume variation at upstream segment would increase the crash risk. ASC2 is significant with a negative sign, indicating that lower speed at the crash segment would increase the crash likelihood, and crashes are more likely to occur within congested traffic flow. This result is consistent with previous study (e.g., Yu & Abdel-Aty, 2013a; Yu & Abdel-Aty, 2013b; Yu et al., 2016; Xu et al., 2013a), where the signs of the average speed are all negative. SSC2 holds a positive coefficient, which indicates that larger speed variation between three time-intervals at crash segment would increase the crash risk. Similar to SVU2, SVD2 has a positive sign, which means that larger volume variation at downstream segment would increase the crash risk. This can be understood as that congested traffic flow spreads to upstream and downstream, which causes the enlargement of the volume variation, increasing the likelihood of rear-end crashes; the large speed variation at crash segments increases the lane changes, which may increase the likelihood of sideswipe crashes.

Moreover, if their estimated standard errors were not statistically different from 0, the parameters were fixed to be constant across the roadway-segment population, thus, SVU2 and ASC2 are random. Looking at the specific results in **Table 4**, SVU2 is normally distributed with mean 0.014429 and standard deviation 0.0590554. Given these estimates, the constant term is less than 0 on 40.3% of the segments and greater than 0 on 59.7% of the segments. This implies that in slightly less than half of the roadway segments result in a decrease in SVU2 (i.e. Std. Dev. of Volume for U section during 6-12min prior to crash occurrence) and slightly more than half result in an increase in SVU2. This result is likely picking up a complex interaction among traffic volume variation, driver behavior and crash risk. This finding has important implications in that it suggests that the effect of the volume variation on crash risk outcomes cannot be assumed to be uniform across geographic locations.

ASC2 results in a parameter that is normally distributed with a mean -0.0649187 and standard deviation 0.0201141. Again, both the mean and standard deviation are statistically significant indicating that the parameter effect varies over the sample of roadway segments. With the estimated parameters, 99.9% of the distribution is less than 0 and 0.1% is greater than 0. This implies that almost all of the roadway segments result in a decrease in ASC2 (i.e. Average Speed for C section during 6-12min prior to crash occurrence).

The AUC (Area under the ROC Curve) (Hand, 2009) is employed to test the predictive performance of the mixed logit model. The AUC value is 0.762, and is better than most previous studies (e.g. Xu et al., 2015), which indicates that the goodness-of-fit of mixed logit model is very good.

6.2 Threshold selection

The crash risk of cases in the dataset was evaluated by the mixed logit model. And then, several methods were employed to select the thresholds on the basis of the crash risk of cases. Besides, the performances of the thresholds from different methods were identified by several evaluation criteria. The threshold values from other different methods are presented in **Fig. 6**.

As shown in **Fig. 6**-(a), the histogram of the crash risk has a deep and sharp valley between two peaks. The percentage at the bottom of this valley is 0.21%, and the cut-off point is 0.2425. Thus, the threshold is 0.2425 by the bimodal histogram threshold method.

The proportion of crash and non-crash in training data is 26.39% and 73.61%, respectively. From the cumulative proportion curve (CPC) of crash risk (**Fig. 6**-(b)), the cut-off point of crash risk for P-tile method is 0.3436 when the proportion is equal to 73.61% (1-26.39%).

Similarly, the threshold is 0.25 for the maximum between-class variance method (**Fig. 6**-(c)); the threshold is 0.129 for maximum entropy method (**Fig. 6**-(d)); the threshold is 0.177 for minimum cross-entropy method (**Fig. 6**-(e)).

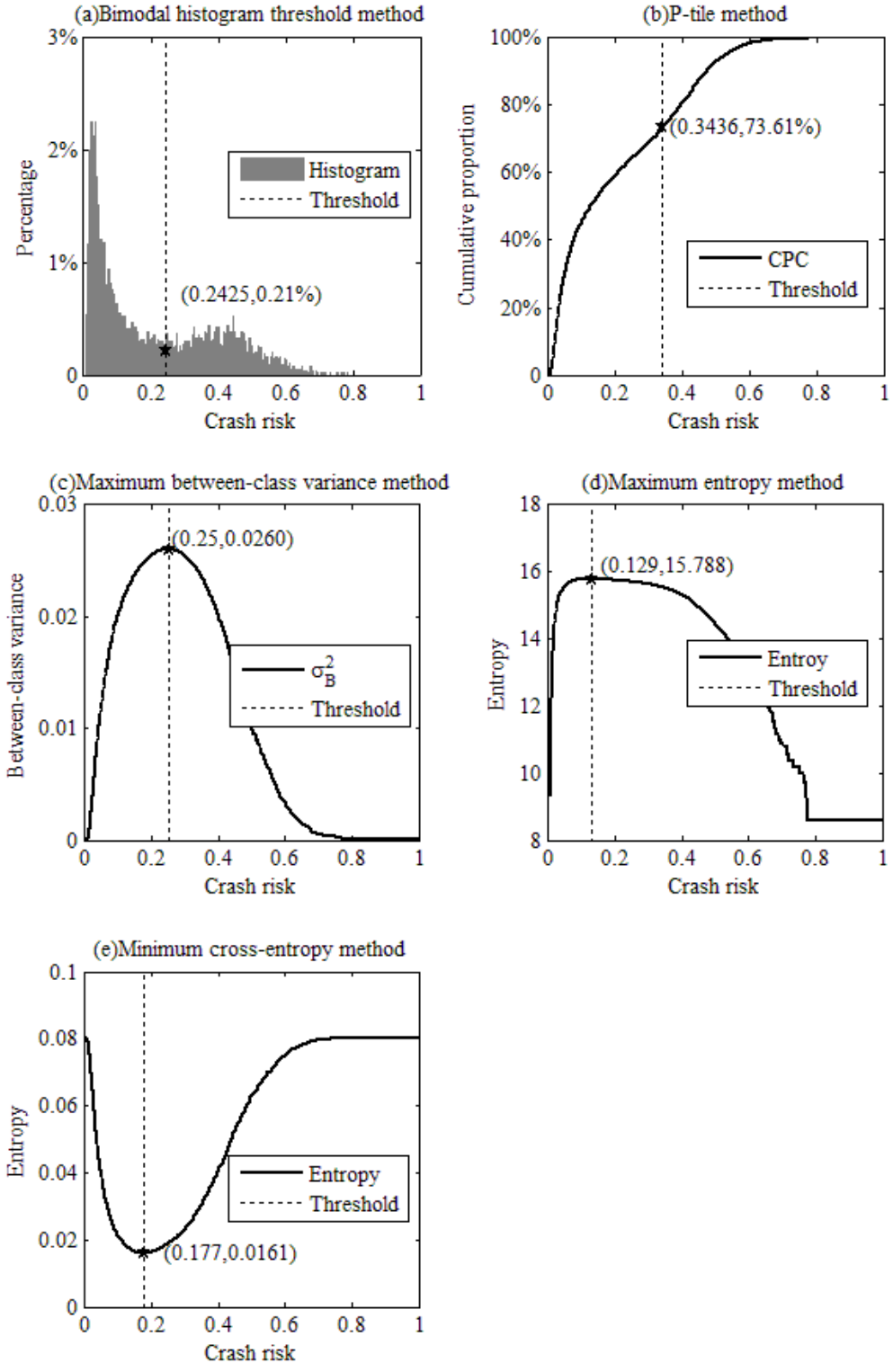

Fig. 6. Threshold selection of five methods.

6.3 Comparison of predictive performance

On the basis of the selected thresholds, predictive accuracy of crash and non-crash, and three evaluation criteria, were calculated. The values and rankings are summarized in **Table 5**, and the bigger value and the higher ranking indicate the better predictive performance.

For the individual index, maximum entropy method is the best in terms of the accuracy of crash (the proportion of crashes predicted correctly in crashes), while P-tile method is the best on the basis of the accuracy of non-crash (the proportion of non-crashes predicted correctly in non-crashes). Besides, Youden's index, F-score, and Phi coefficient conclude that minimum cross-entropy method provides the best predictive performance. Finally, the mean value of Youden's index, F-score, and Phi coefficient were synthesized with the same weight after zero-mean normalization, also showing that minimum cross-entropy method has the best predictive performance.

Table 5 Comparison Results of Predictive Performance by Evaluation Criterion.

Method	Accuracy of crash	Accuracy of non-crash	Youden's index	F-score	Phi coefficient	Synthetic index
Bimodal histogram threshold method	3* (55.89%)**	3 (70.39%)	4 (0.26281)	3 (0.46875)	3 (0.24054)	4 (0.21470)
P-tile method	5 (42.08%)	**1 (79.09%)**	5 (0.21170)	5 (0.41997)	5 (0.21145)	5 (-1.72912)
Maximum between-class variance method	4 (55.12%)	2 (71.19%)	3 (0.26314)	4 (0.46820)	2 (0.24200)	3 (0.24620)
Maximum entropy method	**1 (69.92%)**	5 (57.26%)	2 (0.27174)	2 (0.48366)	4 (0.23955)	2 (0.49309)
Minimum cross-entropy method	2 (64.24%)	4 (63.65%)	**1 (0.27883)**	**1 (0.48368)**	**1 (0.24777)**	**1 (0.77512)**

*Ranking of each method from the evaluation criteria.

**The value of the evaluation criteria for each method.

For the purpose of obtaining a more accurate estimate of prediction performance (Seni & Elder, 2010), k-fold cross-validation (Rodriguez et al., 2010) was used in this study. Since the common advice to take k=5 is sufficient when the computational power is limited (Arlot and Lerasle, 2016), k=5 was chosen. In the 5-fold cross-validation, the dataset was firstly partitioned into five equally (or nearly equally) sized segments or folds. Secondly, five iterations of training and validation were conducted. Within each of the iterations, a different fold of the dataset was chosen for validation while the four other folds were used for selecting thresholds. Finally, the estimation result of several criteria was the average value of the estimations made in each fold. And the results are shown as **Table 6**.

Table 6 Comparison Results by 5-fold cross-validation.

Method	Threshold	Accuracy of crash	Accuracy of non-crash	Youden's index	F-score	Phi coefficient	Synthetic index
Bimodal histogram threshold method	(0.263, 0.02019)	3* (53.36%, 2.79%)**	3 (71.99%, 1.31%)	4 (0.25350, 0.02779)	4 (0.46099, 0.02697)	4 (0.23490, 0.02767)	4 (-0.06310)***
P-tile method	(0.344, 0.00436)	5 (42.00%, 2.61%)	**1 (79.27%, 1.08%)**	5 (0.21273, 0.01553)	5 (0.41972, 0.00906)	5 (0.21266, 0.00986)	5 (-1.66875) ***
Maximum between-class variance method	(0.250, 0.00130)	4 (55.09%, 1.85%)	2 (71.14%, 1.97%)	3 (0.26232, 0.01922)	3 (0.46758, 0.01982)	2 (0.24145, 0.02107)	3 (0.29766) ***
Maximum entropy method	(0.129, 0.00217)	**1 (69.89%, 2.51%)**	5 (57.28%, 1.39%)	2 (0.27169, 0.02093)	2 (0.48341, 0.01923)	3 (0.23962, 0.01970)	2 (0.57701) ***
Minimum cross-entropy method	(0.176, 0.00089)	2 (64.19%, 1.68%)	4 (63.64%, 1.55%)	**1 (0.27837, 0.01588)**	**1 (0.48359, 0.01893)**	**1 (0.24750, 0.01687)**	**1 (0.85717)** ***

*Ranking of each method from the evaluation criteria.

**Mean and standard deviation of performance from 5-fold cross-validation.

***The average value of three criteria (i.e. Youden's index, F-score and Phi coefficient) with the same weight after zero-mean normalization.

The thresholds and evaluation values of several criteria from 5-fold cross-validation (**Table 6**) are similar to those in **Table 5**. On the basis of Youden's index, F-score, Phi coefficient and their synthetic index, minimum cross-entropy method provides the best predictive performance. Cross entropy measures a theoretic information distance between two distributions. The smaller the cross entropy is, the more similar the distributions of the two variables are. Thus, it can be well-behaved to identify thresholds in the bilevel thresholding case by minimizing the cross entropy between the original and binarized dataset (Yin, 2007), such as crash risk, the gray-level of pixel in image. However, it could be very time-consuming in the multilevel thresholding scenario for more complex dataset analysis.

In terms of Youden's index, F-score and synthetic index, the performances of maximum between-class variance method and bimodal histogram threshold method are at the second best levels. The maximum between-class variance method assumes that the crash risk level of crashes and non-crashes in dataset is Gaussian distribution with equal variances (Kurita et al., 1992), thus it is also considered as one of the top threshold selection methods for thresholding a histogram with bimodal. Nevertheless, the formulation of between-class variance is inefficient in

case of multilevel thresholding. With the growth of the number of levels, the computational time scales exponentially, and its accuracy decreases with each new threshold point (Sathya & Kayalvizhi, 2011). In addition, it fails if the histogram is unimodal or approximately unimodal. On the other hand, bimodal histogram threshold method is also considered as one of the threshold selection methods for thresholding a histogram with bimodal distribution but fails if the histogram is unimodal or approximately unimodal.

6.4 Implementation discussion

The crash risk evaluation model in this study has the potential to be used in the ATM to improve traffic safety on urban expressways. **Fig. 7** illustrates a possible real-time implementation of the crash risk evaluation model and thresholds to identify the hazardous traffic conditions. Traffic flow parameters are collected from all the segments on urban expressways in real time, and are aggregated into traffic variables. The crash risk (i.e. crash probability) of each segments are then estimated by the crash risk evaluation model with the input of traffic variables, and further are compared with the predetermined thresholds. If the crash risk exceeds the threshold, the segment will be flagged as a *potential crash scenario*, and then a crash warning is alerted, and further control strategies are triggered (Abdel-Aty et al., 2010). For instance, 0.176 was identified as a threshold by minimum cross-entropy method. The segment will be flagged as the hazardous traffic condition and the ATM can take measures (e.g. Variable Speed Limit, Queue Warning and Ramp Metering) to reduce the crash risk, if the crash probability is greater than 0.176. The threshold should be automatically and theoretically identified by the threshold selection methods (e.g. minimum cross-entropy method in this study) with the help of a matched case control dataset. Additionally, different thresholds for different traffic conditions at different roads or during different periods can be automatically identified by this method, based on the different matched case control samples.

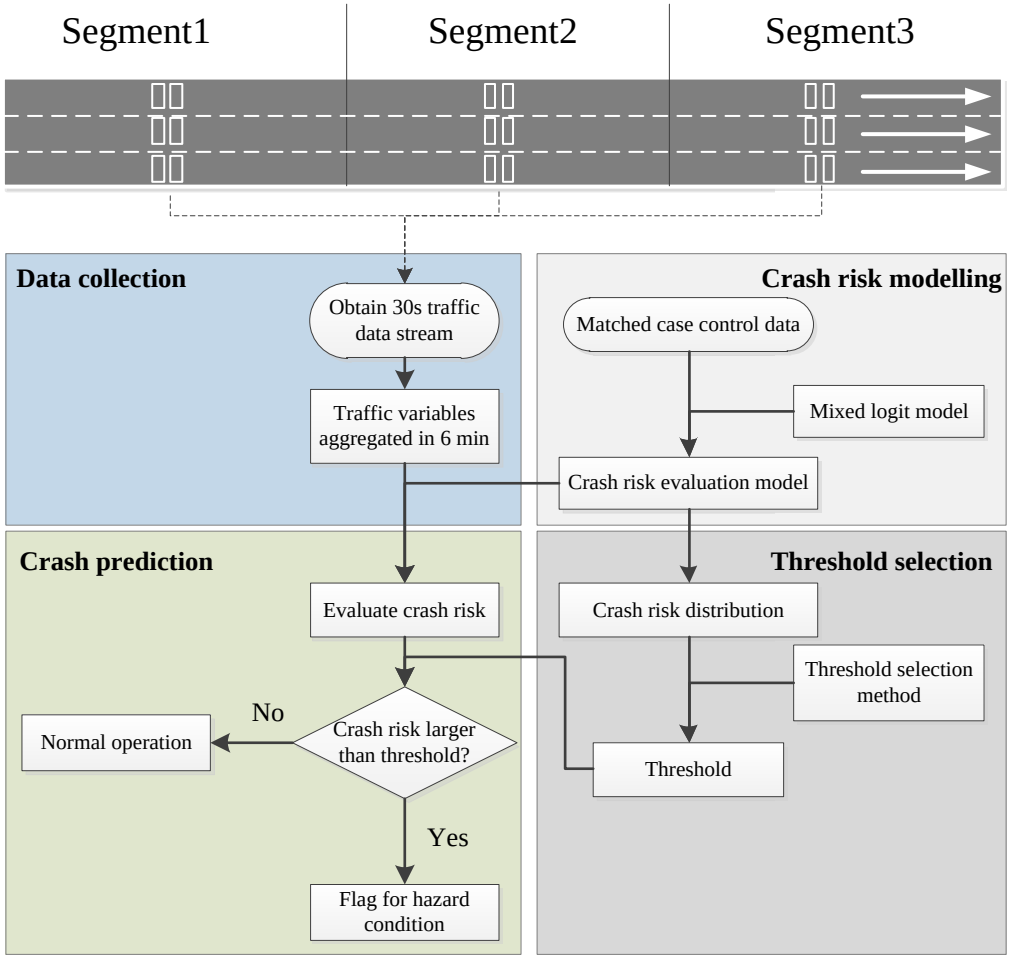

Fig. 7. Real-time implementation of the crash risk crash models and threshold in ATMs

In deriving the probability outcomes for the mixed logit model we have to recognize that some explanatory variables are a composite function of a mean parameter, a distribution around the mean and decomposition of the mean by some contextual effect. Each individual case is "located" in parameter space on the normal distribution for two traffic variables in our research. For each individual we randomly draw a location on the distribution given the mean and standard deviation and derive their overall contribution to "relative utility". This is derived a repeated number of times and averaged per case (Jones & Hensher, 2004). Given that the focus is on a sample drawn from a population of cases, the parameterization used to establish the probabilistic outcomes is a representation of the preference profile of a sample (Jones & Hensher, 2004). Thus, it is necessary to aggregate the probabilities associated with each outcome across entire sample to obtain the predicted outcome values (i.e., the absolute number predicted to be in each outcome category or outcome category shares). Thus, the history traffic data from each segment on the urban expressways should be stored and used as matched case-control data to help predict the crash risk from new sample in real time. Similarly, more and more researchers (e.g. Lovreglio et

al., 2016; Tamura & Giampaoli, 2013; Trabelsi et al., 2015) employed this technique to predict the probability outcomes.

7 Discussions and conclusions

In this study, in order to automatically identify the optimal thresholds in different traffic and weather conditions and avoid subjective judgments, a total of five theoretical methods were proposed. The mixed logit model was chosen to develop the crash risk evaluation model and further evaluate the crash risk. Different thresholds were selected by five threshold selection methods, and their predictive performances were further evaluated through different evaluation criteria. Besides, 5-fold cross-validation was used to test the thresholds for obtaining more accurate evaluation results.

Considering the random effect of variable factors across all roadway segments, the mixed logit model can obtain a good goodness-of-fit. Both the mean and standard deviation of SVU2 and ASC2 are found to be statistically significant, indicating that the parameter effect varies over the sample of roadway segments. This result is likely picking up a complex interaction among traffic volume variation, driver behavior and crash risk. On the basis of the results of the crash risk evaluation model, crashes are more likely to occur within congested traffic flow, which was consistent with previous study (e.g. Yu & Abdel-Aty, 2013a; Yu & Abdel-Aty, 2013b; Yu et al., 2016; Xu et al., 2013a). The congested traffic flow spreads to the upstream and downstream, which causes the increase of the volume variation, increasing the likelihood of rear-end crashes; the large speed variation at crash segments increases the lane changes, which may increase the likelihood of sideswipe crashes.

On the basis of the several criteria, the threshold by the maximum entropy method achieved the highest predictive accuracy of crash, and it could be the best choice if we expect to identify as many crashes as possible. But if we expect to gain as few false alarms as possible, which helps reduce the cost of the ATM, P-tile method could be the best choice. However, it is not applicable if the prior knowledge (i.e. the proportion of crashes or non-crashes) is not known or various from different road segments or road networks. In terms of the Youden's index, F-score and synthetic index, the related entropy methods (i.e. maximum entropy method, minimum cross-entropy method) are the first choices for thresholding a histogram with bimodal, and then the maximum between-class variance method and bimodal histogram threshold method are the second choices.

This study firstly provides thorough investigations of threshold selection methods for binary classification in urban expressway real-time crash prediction with the help of microscopic traffic data, which is an important step in the application of the ATM. On the other hand, there are some other issues needed to be studied further. Similar to the spatial and temporal transferability of the developed crash risk evaluation models (e.g. Pande et al., 2011; Xu et al., 2014), for instance,

there might be the transferability issue of the threshold values due to probable variations in traffic flow, roadway geometrics, traffic management, weather, and other factors. It might be unrealistic to directly transfer the threshold value from one freeway or one time period to another. It was not tested because of the limitation of the length of the manuscript and dataset. The authors recommend that future studies may focus on this issue.

ACKNOWLEDGEMENT

This study was sponsored by National Science Foundation of China (NSFC no. 71771174, 71531011, 51138003, 51522810); and supported by the 111 Project (B17032) and the Science and Technology Commission of Shanghai Municipality (15DZ1204800).

REFERENCE

Abdel-Aty, M., Dilmore, J., Dhindsa, A., 2006. Evaluation of variable speed limits for real-time freeway safety improvement. Accident Analysis & Prevention 38(2), 335-345.

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., Pande A., Hsia L., 2010. The concept of proactive traffic management for enhancing freeway safety and operation. ITE Journal 80(3), 34.

Abdel-Aty, M., Uddin, N., Pande, A., 2005. Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. Transportation Research Record: Journal of the Transportation Research Board 1908, 51-58.

Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., 2004. Predicting freeway crashes based on loop detector data using matched case-control logistic regression. In: Compendium of Papers CD-ROM, Transportation Research Board 2004 Annual Meeting. Washington, DC.

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., 2013. A data fusion framework for real-time risk assessment on freeways. Transportation Research Part C 26(1), 203-213.

Arlot, S., M. Lerasle., 2016. Choice of V for V-fold cross-validation in least-squares density estimation. Journal of Machine Learning Research 17(208), 1-50.

Cramir, H., 1946. Mathematical methods of statistics. Princeton University Press, Princeton.

Doyle, W., 1962. Operations useful for similarity-invariant pattern recognition. Journal of the ACM (JACM) 9(2), 259-267.

ESRI (Environmental Systems Resource Institute), 2009. ArcMap 9.2.

Evans, L., 2004. Traffic Safety. Bloomfield Hills, MI: Science Serving Society.

Hand D J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine learning 77(1): 103-123.

Hensher, D. A., Greene, W. H., 2003. The mixed logit model: the state of practice. Transportation 30(2), 133-176.

Jones, S., Hensher, D. A., 2004. Predicting firm financial distress: a mixed logit model. The Accounting Review 79(4), 1011-1038.

Kapur, J., Sahoo, P., Wong, A., 1985. A new method for gray-level picture thresholding using the entropy of the histogram. Computer Vision, Graphics, and Image Processing 29(3), 273-285.

Kurita, T., Otsu, N., Abdelmalek, N., 1992. Maximum likelihood thresholding based on population mixture models. Pattern Recognition 25(10), 1231-1240.

Li, C., Lee, C., 1993. Minimum cross entropy thresholding. Pattern Recognition 26(4), 617-625.

Li, C. W., Tzeng, G. H., 2009. Identification of a threshold value for the DEMATEL method using the maximum mean de-entropy algorithm to find critical services provided by a semiconductor intellectual property mall. Expert Systems with Applications 36(6), 9891-9898.

Lovreglio, R., Fonzone, A., Olio, L. D., 2016. A mixed logit model for predicting exit choice during building evacuations. Transportation Research Part A-policy and Practice 92, 59-75.

Munizaga, M. A., Alvarez-Daziano, R., 2001. Mixed logit vs. nested logit and probit models. In 5th tri-annual Invitational Choice Symposium: Hybrid Choice Models, Formulation and Practical Issues. Available from http://www. cec. uchile. cl/~ dicidet/mmunizaga/mixed_logit. pdf.

Oh, C., Oh, J. S., Ritchie, S., & Chang, M., 2001. Real-time estimation of freeway accident likelihood. In: Compendium of Papers CD-ROM, Transportation Research Board 2001 Annual Meeting. Washington, DC.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics 9(1), 62-66.

Pande, A., Abdel-Aty, M., 2006a. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transportation Research Record: Journal of the Transportation Research Board 1953, 31-40.

Pande, A., Abdel-Aty, M., 2006b. Assessment of freeway traffic parameters leading to lane-change related collisions. Accident Analysis & Prevention 38(5), 936-948.

Pande, A., Das, A., Abdel-Aty, M., Hassan, H., 2011. Real-time crash risk estimation: Are all freeways created equal? In: Compendium of Papers CD-ROM, Transportation Research Board 2011 Annual Meeting. Washington, DC.

Pun, T., 1980. A new method for grey-level picture thresholding using the entropy of the histogram. Signal processing 2(3), 223-237.

Rodriguez, J. D., Perez, A., Lozano. J. A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3), 569-575.

Rothman, K., Greenland, S., Lash, T., 1998. Modern epidemiology. Modern Epidemiology.

Samopa, F., Asano, A., 2009. Hybrid image thresholding method using edge detection. International Journal of Computer Science and Network Security 9(4), 292-299.

Sathya, P. D., Kayalvizhi, R., 2011. Modified bacterial foraging algorithm based multilevel thresholding for image segmentation. Engineering Applications of Artificial Intelligence 24(4), 595-615.

Seni G., Elder J., 2010. Ensemble methods in data mining: improving accuracy through combining predictions. Synthesis Lectures on Data Mining and Knowledge Discovery 2(1): 1-126. (https://doi.org/10.2200/S00240ED1V01Y200912DMK002)

Shanghai City Comprehensive Transportation Planning Institute, 2016. Shanghai Comprehensive Transportation 2015 Annual Report (in Chinese).

Shanghai City Comprehensive Transportation Planning Research Institute, 2006. Study on Shanghai urban expressway network (in Chinese).

StataCorp, L. P., 2013. Stata: release 13-statistical software. College Station, TX.

Tamura, K. A., Giampaoli, V., 2013. New prediction method for the mixed logistic model applied in a marketing problem. Computational Statistics & Data Analysis 66, 202-216.

Trabelsi, S., He, R., He, L., Kusy, M., 2015. A comparison of Bayesian, Hazard, and Mixed Logit model of bankruptcy prediction. Computational Management Science 12(1), 81-97.

van Rijsbergen, C., 1979. Information retrieval (2nd edit.) butterworths. London, UK.

Weszka, J., Nagel, R., Rosenfeld, A., 1974. A threshold selection technique. IEEE Transactions on Computers 100(12), 1322-1326.

Xu C., Liu P., Wang W., Li Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accident Analysis & Prevention 47, 162-171

Xu, C., Wang, W., Liu, P., 2013a. A genetic programming model for real-time crash prediction on freeways. IEEE Transactions on Intelligent Transportation Systems 14(2), 574-586.

Xu, C., Wang, W., Liu, P., 2013b. Identifying crash-prone traffic conditions under different weather on freeways. Journal of Safety Research 46, 135-144.

Xu, C., Wang, W., Liu, P., Guo, R., & Li, Z., 2014. Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models. Transportation Research Part C 38(1), 167-176.

Xu, C., Wang, W., Liu, P., Li, Z., 2015. Calibration of crash risk models on freeways with limited real-time traffic data using Bayesian meta-analysis and Bayesian inference approach. Accident Analysis & Prevention 85, 207-218.

Yin, P., 2007. Multilevel minimum cross entropy threshold selection based on particle swarm optimization. Applied Mathematics and Computation 184(2), 503-513.

Youden, W., 1950. Index for rating diagnostic tests. Cancer 3(1), 32-35.

Yu, R., Abdel-Aty, M., 2013a. Investigating the different characteristics of weekday and weekend crashes. Journal of Safety Research 46, 91-97.

Yu, R., Abdel-Aty, M., 2013b. Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis & Prevention 51, 252-259.

Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety Science 63(4), 50-56.

Yu, R., Wang, X., Yang, K., Abdel-Aty, M., 2016. Crash risk analysis for Shanghai urban expressways: a Bayesian semi-parametric modeling approach. Accident Analysis and Prevention 95(Pt B), 495-502.

Zhang, Y., 2014. Half century for image segmentation //Mehdi. Khosrow-Pour. Encyclopedia of Information Science and Technology. 3rd ed. Hershey: Information Resources Management Association, 5906-5915