IATSS Research 37 (2013) 68–79

Contents lists available at SciVerse ScienceDirect

IATSS Research

# A real-time crash prediction model for the ramp vicinities of urban expressways ☆

Moinul Hossain *, Yasunori Muromachi [1]

*Department of Built Environment, Tokyo Institute of Technology, Nagatsuta-machi, Midori-ku, Yokohama, Kanagawa 226-8502, Japan*

ARTICLE INFO

*Article history:*
Received 4 October 2012
Received in revised form 23 March 2013
Accepted 14 May 2013

*Keywords:*
Real-time crash prediction model
Random multinomial logit
Bayesian belief net
Ramp vicinities

ABSTRACT

Ramp vicinities are arguably the known black-spots on urban expressways. There, while maintaining high speed, drivers need to respond to several complex events such as maneuvering, reading road signs, route planning and maintaining safe distance from other maneuvering vehicles simultaneously which demand higher level of cognitive response to ensure safety. Therefore, any additional discomfort caused by traffic dynamics may induce driving error resulting in a crash. This manuscript presents a methodology for identifying these dynamically forming hazardous traffic conditions near the ramp vicinities with high resolution real-time traffic flow data. It separates the ramp vicinities into four zones – upstream and downstream of entrance and exit ramps, and builds four separate real-time crash prediction models. Around two year (December 2007 to October 2009) crash data as well as their matching traffic sensor data from Shibuya 3 and Shinjuku 4 expressways under the jurisdiction of Tokyo Metropolitan Expressway Company Limited have been utilized for this research. Random multinomial logit, a forest of multinomial logit models, has been used to identify the most important variables. Finally, a real-time modeling method, Bayesian belief net (BBN), has been employed to build the four models using ramp flow, flow and congestion index in the upstream and flow and speed in the downstream of the ramp location as variables. The newly proposed models could predict 50%, 42%, 43% and 55% of the future crashes with around 10% false alarm for the downstream of entrance, downstream of exit, upstream of entrance and upstream of exit ramps respectively. The models can be utilized in combination with various traffic smoothing measures such as ramp metering, variable speed limit, warning messages through variable message signs, etc. to enhance safety near the ramp vicinities.

## 1. Introduction

Ramp vicinities are arguably the most crash prone locations on urban expressways. There, the drivers need to maintain high speed and yet respond to several complex events, such as, maneuvering, taking decisions regarding routes, reading road signs and not to mention, maintaining safe distance from other maneuvering vehicles simultaneously. Hence, any additional disruption in the traffic condition may force driving error which can eventually lead to a crash. If the formation of a disrupted traffic condition can be spotted early, road authorities can take proactive measures by warning the drivers as well as applying various traffic smoothing methods such as variable speed limits, ramp metering, main line metering, etc. to bring the traffic condition back to normal. Recently, a small group of researchers, mainly from North America, are actively pursuing studies with loop detector data to reveal the interrelationship between crash and traffic flow variables [1,19,20,23–25,27]. They have emphasized that certain traffic conditions can be associated with high crash likelihood. Oh et al. [25] implied causal relationship between crash likelihood and 5 minute standard deviation of speed and average occupancy. Abdel-Aty et al. [3] ascertained that traffic leading to crash differs between high speed and low speed scenarios. At high speed, quick formation and subsequent dissipation of queues cause a backward shock wave and in case of low speed scenario, the impact between a congested downstream and a fast paced upstream impends driving errors. Dias et al. [11] discovered positive correlation between level of congestion and crash occurrence. Zheng et al. [32] underscored that recurring patterns of decelerations followed by acceleration increases crash risk. The findings have also stimulated idea of building models that can eventually forecast the crash potential for a short time window in near future taking the real-time loop detector data as input [4,5,12,13,15,16,24,25]. As

☆ Peer review under responsibility of International Association of Traffic and Safety Sciences.

* Corresponding author. Tel.: +81 80 3509 7060; fax: +81 45 924 5524.
*E-mail addresses:* moinul048i@yahoo.com (M. Hossain), ymuro@enveng.titech.ac.jp (Y. Muromachi).

[1] Tel./fax: +81 45 924 5524.

0386-1112/$ – see front matter © 2013 International Association of Traffic and Safety Sciences. Production and hosting by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.iatssr.2013.05.001

the research field is in its infancy, the existing models are yet theoretical and have several major drawbacks. Considering crash phenomena as generic throughout the road section is one of the major drawbacks of the previous studies. Pande and Abdel-Aty [28] affirmed that presence of ramp in the downstream has impact on crash but did not shed light on the types of ramps and their relative vicinity. Jovanis and Chang [18] implied that traffic conditions vary substantially between the basic freeway segments and the ramp areas. General observation may verify that traffic conditions even vary among different types of ramps and their locations. For example, a high proportion of merging traffic can be observed at the downstream of an entrance ramp whereas diverging traffic is quite predominant near the upstream of an exit ramp. On the contrary, maneuvering frequency is comparatively lower in the basic freeway segments than the ramp vicinities. Therefore, it may not be suitable to use a universal model for both the basic freeway segments and the ramp areas. This may be one major reason behind the low detection and high false alarm rate of the existing models. Hossain and Muromachi [13] have dealt with this issue and developed real-time crash prediction models solely for the basic freeway segments and obtained a high detection rates for future crashes with low false alarm. Also, in a later study, Hossain and Muromachi have demonstrated that underlying phenomena behind crash is substantially different among basic freeway segments as well as different ramp vicinities. Likewise, building separate real-time crash prediction models for the ramp vicinities have another advantage. The existing models were developed assuming that they will be implemented throughout the length of the expressway rather than on specific areas of interest. Nevertheless, this may often deter the expressway authorities as it involves huge initial investments as well as regular maintenance cost. On the contrary, ramp vicinities, which are widely regarded as black-spots, cover only a fraction of the total length of the expressway and it is cost effective for authorities interested in real-time monitoring of hazardous location to implement the models first in these locations.

This study separates the ramp vicinities into four zones — upstream and downstream of the entrance and exit ramps; and develops four different real-time crash prediction models. As per the knowledge of the authors, this may be the first attempt to build real-time crash prediction models specifically for the ramp vicinities. The manuscript is organized into five major sections. The introductory section has stated the motivation of the study, its theoretical background and its objectives. Section 2 describes the activities involving experimental design, data extraction and processing. The third section provides a concise yet adequate introduction to Bayesian belief net (BBN), the real-time modeling method employed in the study. It also briefly explains random multinomial logit (RMNL) model which has been applied for variable selection. The subsequent section discusses the model building process as well as their performance evaluation. The concluding section summarizes the noble findings, mentions the limitations and lays out the future directions.

## 2. Study area and the data

### 2.1. Study area

A research of this nature demands large sample size as well as a detector layout that is to some extent uniform. The Shibuya 3 and Shinjuku 4 are arguably the two busiest urban expressways in Japan and every year they sustain a substantial number of crashes. Moreover, they have a relatively uniform (approximately 250 m center to center) detector spacing which makes them highly suitable for this research. The Shibuya 3 and Shinjuku 4 routes have two lanes in each direction and are respectively 11.9 and 13.5 km long. The expressways all together harbor 14 entrance and 15 exit ramps and 210 detectors. A detailed map has also been provided by the Tokyo Metropolitan Expressway Company Limited to identify the location of ramps (see Fig. 1). They have also provided access to a separate dataset containing the location of detectors in nearest 10 m to facilitate the research. The detectors installed in the expressways yield data of speed, vehicle count, occupancy and number of heavy vehicles for each 8 ms round the clock (24 h a day, 365 days a year) for each lane. However, the Tokyo Metropolitan Expressway Company Limited later archives the data by aggregating for all lanes for every 5 min. Hence, the supplied dataset contains 5 minute vehicle count, 5 minute vehicle count for heavy vehicles only, 5 minute average speed as well as occupancy for each detector location. The crash data contain information on date, time, location (in nearest 10 m), route number, direction (in-bound or out-bound), lane, and number of vehicles involved along with their types and type of crash. The data have been collected in two phases spanning over two different time frames. The first dataset contains both crash and detector data from December 2007 to March 2008 for Shibuya 3 route and from December 2007 to October 2008 for Shinjuku 4 route (Phase I). The second phase data collection encompassed data from May 2008 to October 2009 for Shibuya 3 route and November 2008 to October 2009 for Shinjuku 4 route. Thus, the final dataset contains 22 and 23 month detectors as well as crash data for these routes respectively. Another crucial point in the study is the accuracy of the reported time of crash as real-time crash prediction models are supposed to identify hazard risk for a very short time window and error in reported crash time in data will make the newly built models highly erroneous. In an interview the authority responsible for identifying the time of crash has confirmed that the reported crash time can be considered to be within a minute of its occurrence for various reasons. These two expressways are located in the heart of Tokyo, one of the busiest mega cities in the world and serve substantial number of traffic even during night time on the weekends. A portion of the routes are under constant camera surveillance. Safety cars are in operation round the clock on these routes. Moreover, as they have only two lanes in each direction, any incident on road creates high impact and gets detected very quickly. Interestingly, the crash type recorded in the crash database included — rear end, side swipe, hitting road furniture, tipping over along with some other types which may not be directly related to crash. They include vehicles catching fire, hitting objects accidentally fallen from other vehicles or objects falling from vehicles but not hitting any other vehicle, etc. As these incidents are not directly related to crash which might have taken place due to hazardous traffic conditions, they were excluded from the crash samples under consideration. The final crash dataset contains 3018 crash cases (1141 for Phase I and 1877 for Phase II).

### 2.2. Experimental design

Preliminary analysis on the crash dataset suggests that approximately 55% of the crashes took place within 375 m from the ramp vicinities and crash concentration reduces beyond 375 m from the ramp locations. As mentioned earlier, this manuscript develops crash prediction models for the upstream and downstream of entrance and exit ramps separately. Therefore, these four models are built with crashes that had occurred within 375 m upstream and downstream of entrance and exit ramps. The underlying concept of the model is to treat the possibility of a crash occurrence as a classification problem, associate a dataset with hazardous traffic condition and identify its corresponding normal traffic condition data, build a classification based model with them and calculate the probability of a future traffic condition data belonging to any of these two traffic conditions. Hence, it is important to define hazardous and normal traffic conditions. Oh et al. [23] selected a 5 minute time period ending at the time of crash as hazardous traffic condition. They retrieved the corresponding normal traffic condition by taking another 5 minute time period ending 30 min before the reported crash time. Zheng et al. [32] seconded the approach but considered a larger interval (10 min). Abdel-Aty et al. [6] argued that the objective of a real-time crash prediction model is to identify the evolving risk of a crash so that countermeasures can be taken to pacify the traffic. They emphasized that the model must allow adequate

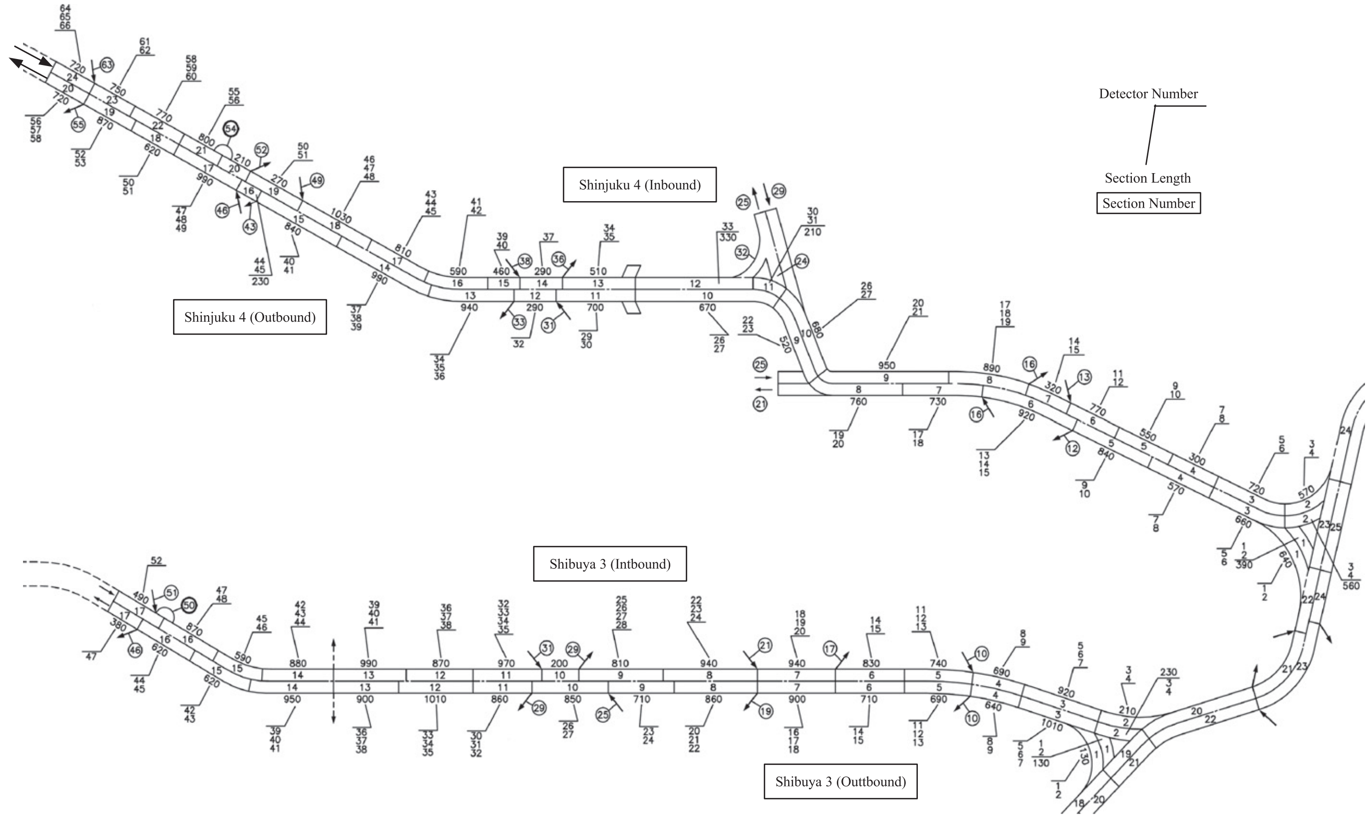


**Fig. 1.** The study area.

time for an intervention to demonstrate positive results. Hence, they also recommended a 5 minute time period but that ends 10 min before the reported time of crash as the hazardous traffic condition. [27] investigated both a 3 minute and a 5 minute aggregation and concluded that the later produces better result for crash prediction. For normal traffic conditions, apart from the definition of Oh et al. [23–25], Abdel-Aty et al. [6] defined it as a 5 minute time period occurring at the same time and same day of the week within the dataset but the crash date. Hossain and Muromachi [15] further refined the definition by excluding those data when a crash had occurred within 1 h from the normal traffic condition data. This study follows the definition by Hossain and Muromach [15] for extracting data of normal traffic condition. To elaborate more, assume that a crash took place on the in-bound direction of Shinjuku 4 route at 8:15 am on 16 October, 2008 (Thursday). Hence, the hazardous traffic condition for this crash data will be a time period between 8:05 am and 8:10 am on that day. Its corresponding normal traffic condition should be the traffic flow data from 8:05 am to 8:10 am for all the Thursdays in the dataset. To refine the data further, if any crash had occurred within 1 h from any of the data points of the normal traffic condition then those data points will be discarded to ensure purity of the normal traffic condition data. The last stage of the data extraction process involves selecting the detector combination from which the hazardous and normal traffic condition data will be extracted. Fig. 2 demonstrates the chosen detector locations for data extraction for all four conditions — upstream and downstream of entrance and exit ramps. Hence, for each condition, data have been extracted from the detector on the ramp (d2) as well as from one detector each in the downstream (Location 1: detector d1) and upstream (Location 2:

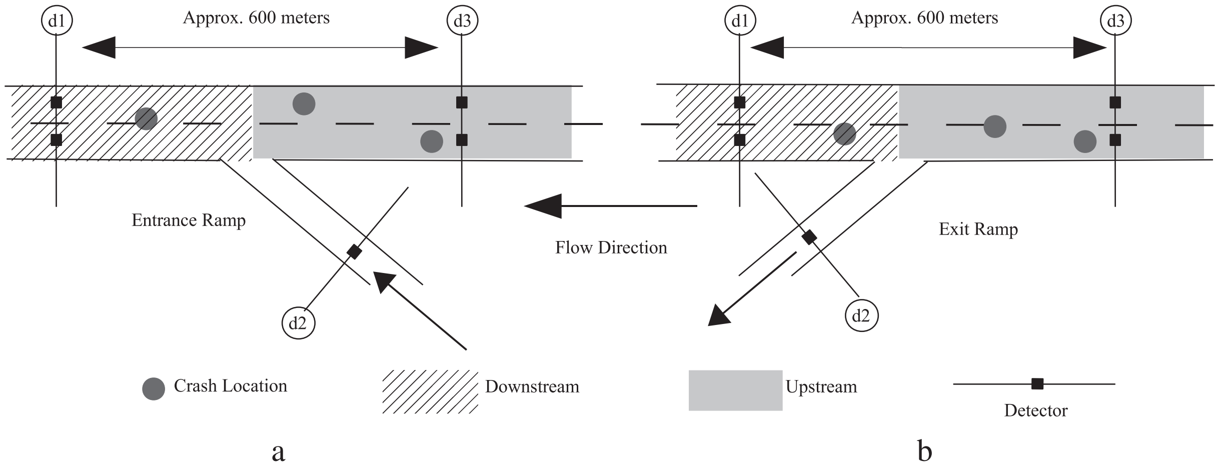


**Fig. 2.** Selected positions of detectors for data extraction.

detector d3) with respect to the ramp location. The detectors d1 and d3 are selected in such way that they are placed at least 300 m away from the ramp location creating a minimum gap of 600 m between them. However, there are some locations where two ramps are very closely spaced disallowing a 300 m gap without overlapping with the adjacent ramp. In those cases, one detector in between the two ramps has been chosen as d1 or d3 (depending on the location) and the other detector on the mainstream has been chosen in such a way that it is at least 500 m away from that already selected detector.

Now, as mentioned earlier, the detectors in location d1 and d3 yield data on 5 minute cumulative vehicle count, 5 minute cumulative heavy vehicle count, 5 minute average speed and occupancy. However, the detector on ramp, i.e., d2 only yields data on 5 minute aggregated vehicle count. These information are stored in variables: d1q, d1p, d1v, d1o, d3q, d3p, d3v, d3o and d2q where the second digit represents the detector number and the third digit represents the variable, i.e., ‘q’ for flow, ‘p’ for heavy vehicle flow, ‘v’ for speed and ‘o’ for occupancy. These are the most common variables that can be found in the previous studies as well and most of the modern detectors can yield data for these variables. Zheng et al. [32] have demonstrated that crash risk is directly related to oscillating (stop and go) traffic and measured its impact. Previously, Dias et al. [11] had introduced congestion index, a measure that standardizes the level of congestion (CI), as a new variable and demonstrated decent results in finding their interrelationship with crash. They calculated CI at each detector location as:

$$\text{Congestion Index (CI)} = (\text{Free Flow Speed}-\text{Speed})/\text{Free Flow Speed}; \\ \text{when CI} > 0 = 0; \text{when CIb} = 0. \qquad (1)$$

This study also includes CI at location d1 and d3 as new variables calculated based on Eq. (1) and represents them as d1i and d3i. The free flow speed at each detector location has been calculated from the speed-flow and speed-occupancy plots. Apart from these, the study also considers the longitudinal variation of the traffic flow variables between detectors 1 and 3 and they are represented as: d13q, d13p, d13v, d13o and d13i. Hence, every data point associated with a hazardous or normal traffic condition is represented with 16 variables ($2 \times 5 = 10$ variables directly from d1 and d3, 1 from d2 and 5 more representing the longitudinal differences).

The initial dataset contains 680 (89 on Shibuya 3 and 591 on Shinjuku 4) and 970 (411 on Shibuya 3 and 559 on Shinjuku 4) crash samples for Phases I and II respectively for the ramp vicinities. Table 1 presents the distribution of crash types associated with the four ramp vicinities. However, after retrieving the corresponding detector data for all the crash samples, only 619 cases from both the phases had complete information of all the 16 variables. Of these, the last two months' crash data have been kept for evaluating the performance of the models and the rest of the data will be used to train the models. Table 2 presents the details regarding the sample size for the datasets for modeling and evaluation.

## 3. Methodology

The study on real-time crash prediction deals with a classification problem of a scenario with dichotomous outcome. It also has some of its own problem specific requirements. As it can be observed from the previous section, the study deals with a problem having a small sample size for one of the two outcomes but employs a large variable space to make predictions. The predictors are highly correlated in nature as well. Moreover, a real-time crash prediction model is supposed to assess the risk of a traffic condition within a very short time window and thus, it is better for it to have a small variable space. For this, a mechanism is needed to identify and rank the most important variables. The study employs random multinomial logit (RMNL), a recently introduced method that combines the benefits of random forest (RF) [9] and multinomial logit (MNL) models, for variable selection. For the modeling purpose, the research faces some major challenges. The model uses detector data corresponding to crash as the input data. It is always difficult to have a large sample size in most of the research areas concerning road crashes. Crash is a highly diverse and complex phenomenon and depends on a series of variables. Moreover, when new data are available, it may not contain information regarding all the variables present in the model. Hence, the method used for real-time crash prediction is expected to be flexible enough to update itself in course of time with partially available new data, accommodate correlated variables, as well as incorporate new variables easily into the existing model when data are available in future. Considering these, Bayesian belief net (BBN) has been chosen in this study as the method for real-time crash prediction model building. This itself is a real-time modeling method which also ensures shorter calculation time under real-time operation. The following sections provide a brief but self containing introduction to RMNL and BBN.

**Table 1**
Distribution of crash types near ramp vicinities.

| Crash type | Location | | | |
|---|---|---|---|---|
| | Downstream entrance (%) | Downstream exit (%) | Upstream entrance (%) | Upstream exit (%) |
| Rear-end | 44.91 | 55.46 | 31.73 | 51.09 |
| Side-swipe | 31.02 | 26.05 | 36.06 | 38.69 |
| Road furnitures | 23.61 | 16.81 | 32.21 | 8.76 |
| Tip over | 0.46 | 1.68 | 0 | 1.46 |

**Table 2**
Sample size for model building and evaluation.

| Location | Model | | Evaluation | |
|---|---|---|---|---|
| | Hazardous | Normal | Hazardous | Normal |
| Downstream − entrance | 143 | 6182 | 22 | 978 |
| Downstream − exit | 102 | 5148 | 19 | 793 |
| Upstream − entrance | 136 | 5534 | 21 | 916 |
| Upstream − exit | 147 | 7006 | 29 | 1318 |

### 3.1. Random multinomial logit (RMNL)

The approaches followed by most of the notable previous studies for selecting the variables for modeling can be classified into three groups: engineering judgment, logistic regression and random forest. Among these, logistic regression has so far been the most widely used [2,3,27,28]. It has a robust theoretical background and the method is well known in a wide variety of research fields. However, logistic regression can be vulnerable in handling a large variable space with too many highly correlated variables. Abdel-Aty et al. [6] addressed this issue by applying random forest which uses the boosting [30] and bagging [8] method of ensemble learning coupled with random variable selection to overcome the problems associated with multi-collinearity. Random forest also has an in-built mechanism to associate numerical value to the relative importance of the variables in the model. However, according to Strobl et al. [31], although the method is considered stable, unbiased and capable of handling large variable space with small sample size, it can be biased when any or a group of variables have relatively larger number of classes as compared to the other variables under consideration. Prinzie and Poel [29] addressed the shortcomings of both these models by introducing the boosting and bagging method as well as the random variable selection technique of random forest into multinomial logit model and named the new method as random multinomial logit (RMNL). The basic difference between random forest and RMNL is in the method used for generating trees. In case of random forest the trees generated are classification and regression trees (CART) [7] whereas the trees in RMNL are individual multinomial logit models. The procedure of calculating the relative variable importance can be explained in four steps:

i) Let $L$ be the dataset under consideration with $N$ records, $M$ variables and $B$ trees. The $b$-th bootstrap sample $L_b$ is created by randomly selecting $n$ samples (approximately 2/3rd of $N$) with replacement from $L$. Next, the out of bag (OOB) of $L_b$ is created extracting $L - L_b$.
ii) Let $T_b$ be one of the $B$ trees of logit model generated by randomly selecting $m$ out of $M$ variables from $L_b$.
iii) The OOB error rate $r_b$ of tree $T_b$ is obtained by comparing the predicted outcome of the $L - L_b$ dataset with the actual outcome using the logit model developed in $T_b$.
iv) Relative variable importance of each variable is calculated by permuting the values of all the variables one by one for each of the $B$ trees, recalculating the difference in misclassification rate for each permuted variable and averaging the misclassification differences for each variable for the $B$ trees. To elaborate more, for each $j$-th variable of $T_b$ permute all its data points in $L_b$ and recalculate the misclassification rate $r^{j}_{b}$. The value $|r_b - r^{j}_{b}|$ is called the importance of the $j$-th variable $V_j$ for $T_b$. The process is repeated for all $B$ trees. The final importance of the $j$-th variable is obtained by averaging all the values of $V_j$. Hence, the term 'variable importance' in RMNL reflects how much misclassification error is associated with wrongly calculating the value of the variable under concern. A higher value indicates that the variable is more important.

At present there are no commercial or open source software products available to apply RMNL. Therefore, a program has been developed for this study using R scripting language [10] applying the detailed algorithm provided by [29].

### 3.2. Bayesian belief net (BBN)

#### 3.2.1. The fundamentals

A BBN $N = (G = (V, E), P)$ consists of a set of probability distributions $P$ and an acyclic directed graph (DAG) $G$ of a set of unique random variables represented by nodes (or vertices) $V$ and connected with directed links (or edges) $E$ where there is a conditional probability distribution $P(X_i|parents(X_i))$ for each random variable $X_i$ $V$. The joint probability distribution of the universe of unique random variable $U = \{X_1, \ldots., X_n\}$ for a system factorizes according to the structure of $G$ as [17,22]:

$$P(U) = P[X_1, X_2, \ldots, X_n] = \prod_{i=1}^{n} P(X_i|\, parents(X_i)). \tag{2}$$

In general, Eq. (2) represents a system or an expert's perspective regarding the system represented with the interrelationship among a set of random variables $[X_1, X_2, \ldots, X_n]$. Now, Eq. (2) can be utilized to obtain answer regarding any probabilistic query on the system when knowledge about state of one or a set of random variables will be available. For example, let us assume that evidence regarding the state of $m$ random variables $e1, \ldots., em$ becomes available ($m < n$). Plugging this new information into Eq. (2) we can obtain Eq. (3):

$$P(U, e) = \prod_{i=1}^{n} P(X_i|\, parents(X_i)) \prod_{j=1}^{m} e_j. \tag{3}$$

Now, if we would like to draw inference regarding different states of a random variable $X$ $U$ then we will need to marginalize the left hand side of Eq. (3) as Eq. (4):

$$P(X, e) = \sum_{U \setminus \{X\}} P(U, e). \tag{4}$$

Finally, $P(X|e)$ can be calculated by using the Bayes' theorem [17] as illustrated by Eq. (5):

$$P(X, e) = \frac{\sum P(X, e)}{P(e)} = \frac{\sum P(X, e)}{\sum_{X} P(X, e)}. \tag{5}$$

#### 3.2.2. An example BBN

To illustrate the concept, let Fig. 3a represent a system explained by an expert with four variables $U = \{X_1, \ldots., X_4\}$. Here, $X_2$ and $X_4$ are called the 'parent nodes' as they have no incoming nodes and thus they are conditionally independent from rest of the variables in the system. $X_1$ is a 'child node' of $X_2$ and $X_1$ and $X_4$ are the parents of $X_3$. Applying Eq. (2), this system can be presented mathematically as:

$$P(U) = P[X_1, X_2, \ldots, X_4] = P(X_2)P(X_4)P(X_1|X_2)P(X_3|X_1, X_4). \tag{6}$$

Let $X_1$ have $k$ states and the probability distribution of different states of $X_1$ be $P(X_1) = (x_{11}, x_{12}, \ldots, x_{1j}, \ldots, x_{1k})$. If evidence $e$ from one instance of data suggest that $X_1$ is in $j$ state then the probability distribution can be represented as $P(X_1) = (0, 0, \ldots, x_{1j}, \ldots, 0)$. Now, the posterior belief associated with the state of any other variable in $U$ can be calculated by plugging this evidence into Eq. (6) and then using Eqs. (4) and (5).

Now, let us assume a situation where the model represented with Fig. 3a and Eq. (6) is being transferred to represent a similar system but in a different environment. The DAG for the new environment has been updated as illustrated by Fig. 3b based on local expert opinion and higher level of data availability ($X_5$). The new model can be explained mathematically as:

$$P(U) = P[X_1, X_2, \ldots, X_5] = P(X_2)P(X_5)P(X_1|X_2)P(X_3|X_1, X_4)P(X_4|X_5). \tag{7}$$

Hence, the model can be customized for the new environment only by building the probability distribution $P(X_5)$ and re-building the conditional probability table $P(X_4|X_5)$.

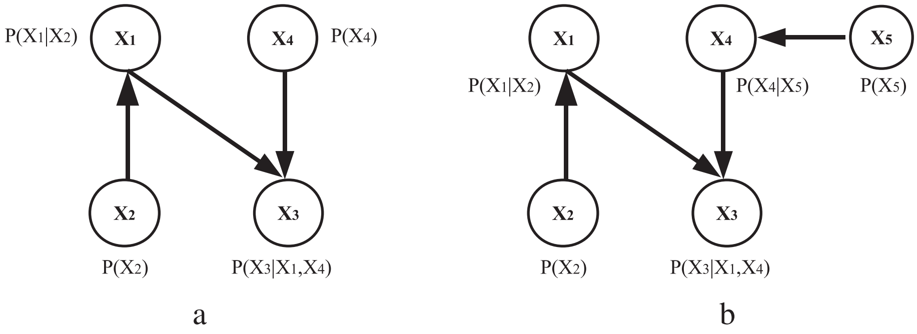


**Fig. 3.** An example Bayesian belief net (BBN).

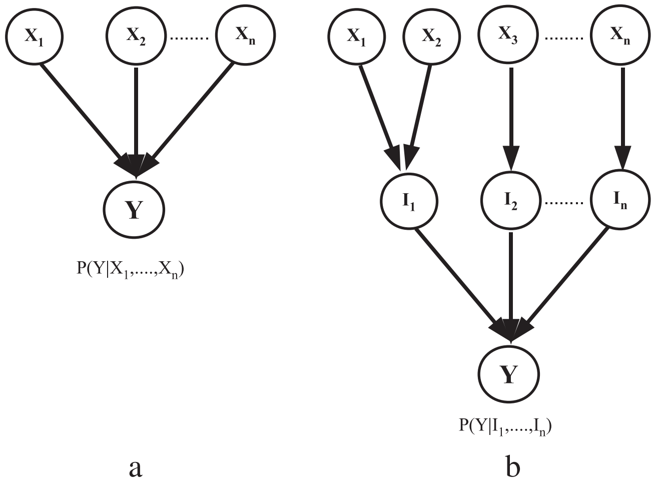


**Fig. 4.** Parent divorcing mechanism in BBN.

### 3.2.3. Parent divorcing in BBN

Complexity of the conditional probability table of a node in a BBN depends on its number of parents and their states. For example, if a node $Y$ in a BBN has set of parents with $n$ elements, i.e., $S=\{X_1,....,X_n\}$ and if $Y$ has $M$ states and $X_i$  $S$ has $Q$ states each then the conditional probability table of $Y$ will have $MxN^n$ number of cells in it (see Fig. 4a). In such situations, the parent divorcing method by Olesen et al. [26] can be employed to reduce the model complexity. Parent divorcing reduces the complexity of a child node by introducing 'intermediate' nodes in between the child node and the parent node(s). Each intermediate node becomes a function of one or more parent nodes and gets represented as their child. These intermediate nodes then become the parent nodes of the original child node $Y$ as illustrated by Fig. 4b. Normally these intermediate nodes have states fewer than $M$. Hence, after 'parent divorcing', the new conditional probability table of $Y$ will have much fewer number of cells.

### 3.2.4. Suitability of BBN for real-time crash prediction models

BBN is a relatively newer probabilistic graphical modeling method that is gaining popularity for its flexibility and efficiency in reasoning under uncertainty. Unlike conventional statistical modeling methods, the BBN models the complete system (limited by the knowledge, availability of data and pre-decided complexity of the system) rather than focusing on only the problem. Therefore, it does not separate the independent and dependent variables. Inference about the state of any variable can be made based on the available evidence about other variables. When information about a variable is sought, it is called the 'outcome variable'. When evidence about a variable or a set of variables is available they are together called 'information variables'. The method has some inherent benefits that are highly suitable for predicting crash in real-time. An acceptable sample size consisting of matching crash and detector data is a prerequisite to develop such models. As crash is a rare event, many a time it is difficult to amass a sufficiently large sample for model building. Moreover, sometimes data on all the variables may not share the same time period. A BBN based model can overcome both these hindrances. When new data are available, the conditional probability tables can be updated to re-calibrate the model. In case of partial data availability, for example, if data for $X_1$ and $X_2$ become available in a later time period for the model in Fig. 3a then the model can be partially re-calibrated by updating probability tables $P(X_2)$ and $P(X_1|X_2)$. Moreover, crash is a highly complex phenomena and it is perceived to be the result of interaction of a wide range of variables. As demonstrated with Fig. 3b and Eq. (7), when the model needs to

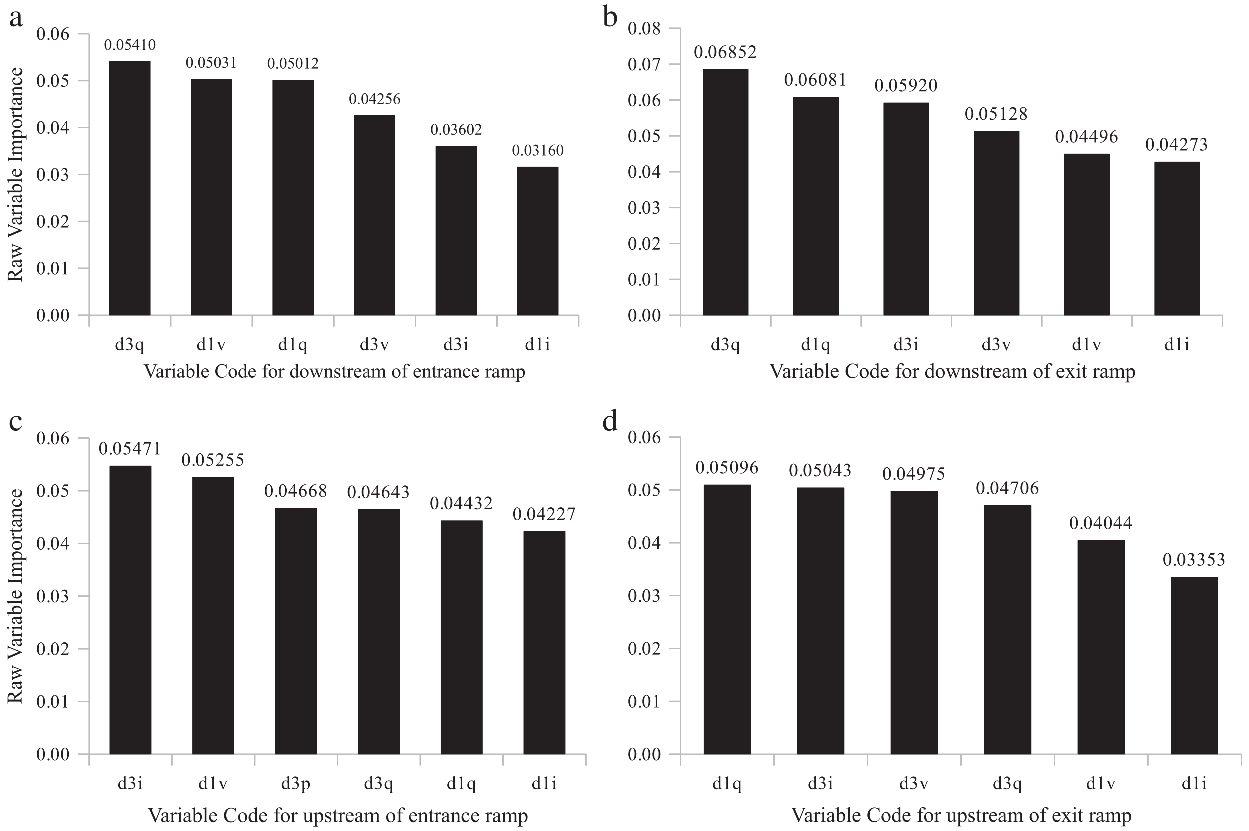


**Fig. 5.** Variable importance (top 6) for (a) downstream of entrance ramp, (b) downstream of exit ramp, (c) upstream of entrance ramp and (d) upstream of exit ramp.

accommodate new variables, it can be achieved without requiring rebuilding the whole model from the scratch. Traffic flow variables are highly correlated in nature and unlike many of the available statistical methods, e.g., binomial logit models, BBN is capable of handling multicollinearity issues. Likewise, BBN is more relaxed towards having statistical assumptions, such as linearity and additivity. Besides these, BBN based models also have the capability to sequentially update itself when new data becomes available provided that its structure and initial specifications of the conditional probability distributions are given in advance. The process is known as 'adaptation'. This is outside the scope of this manuscript and the interested readers are requested to consult [17,22] for further details. For a real-time crash prediction model, adaptation can benefit in two ways. Firstly, an initial model built with low sample size can be updated in real-time once it has been implemented in real life scenario. Secondly, it addresses the issue of model transferability as a model built for one expressway can be customized for another expressway by further training it with new data. This way, the model can benefit from the existing domain knowledge and update itself based on the scenario in the new environment.

**Table 3**
Correlation among information variables.

| Pearson correlation (ρ) | | | | | | |
|---|---|---|---|---|---|---|
| Location | Variable | d1q | d1v | d3q | d3v | d3i |
| Downstream entrance | d1q | 1.0 | −0.120 | 0.949 | −0.071 | 0.126 |
| | d1v | | 1.0 | −0.109 | 0.896 | −0.912 |
| | d3q | | | 1.0 | −0.041 | 0.0901 |
| | d3v | | | | 1.0 | −0.980 |
| Downstream exit | d1q | 1.0 | 0.0204 | 0.972 | −0.192 | 0.073 |
| | d1v | | 1.0 | 0.004 | 0.854 | −0.914 |
| | d3q | | | 1.0 | −0.199 | 0.079 |
| | d3v | | | | 1.0 | −0.969 |
| Upstream entrance | d1q | 1.0 | −0.249 | 0.946 | −0.155 | 0.241 |
| | d1v | | 1.0 | −0.210 | 0.879 | −0.907 |
| | d3q | | | 1.0 | −0.081 | 0.171 |
| | d3v | | | | 1.0 | −0.977 |
| Upstream exit | d1q | 1.0 | −0.026 | 0.948 | −0.091 | 0.040 |
| | d1v | | 1.0 | −0.023 | 0.879 | −0.898 |
| | d3q | | | 1.0 | −0.075 | 0.024 |
| | d3v | | | | 1.0 | −0.988 |

## 4. Model building

The model building process has been subdivided into three parts — variable selection, building the BBN and performance evaluation.

### 4.1. Variable selection

The data collected during the Phase I of the research have been applied to identify and rank the most important variables for modeling. The variable representing ramp flow (d2q) has not been included to be ranked as sufficient literatures are available in support of its importance in influencing crash [28]. Thus, the study avoids redundancy and directly incorporates ramp flow as a variable for the final BBN models. The Phase I data contain 120, 86, 110 and 98 samples for hazardous traffic conditions respectively for the downstream and upstream of entrance and exit ramps and their corresponding 4268, 3785, 3738 and 4164 samples for normal traffic conditions. RMNL has been applied for all four ramp vicinities separately through growing 500 trees of logit model by randomly selecting 4 variables at a time and the results for the average variable importance have been extracted for every 100 trees. The difference in average variable importance for the most important variables are found to be negligible (up to 4 decimal points) between 400 and 500 trees and the results yielded after 500 trees have been finalized as the relative average variable importance. The outcomes for the top six most important variables have been illustrated by Fig. 5. It can be observed that the 5 minute cumulative vehicle count and average speed are the two most important variables for the detector in the downstream (d1q and d1v) for all four ramp vicinities. The 5 minute cumulative vehicle count is also always among the top two variables yielded by the upstream detector (d3q). It is interesting to observe that none of the variables representing spatial variation of traffic flow variables (d13q, d13p, d13v and d13i) are found to be within the top 6 positions for any of the ramp vicinities. This suggests that the hazardous traffic condition in this study can be better distinguished from the normal traffic condition based on the values of the individual detectors rather than their spatial variation. The study intends to choose two most important variables from both the detectors in the upstream (Location 2) and the downstream (Location 1). Hence, along with d1q, d1v and d3q the fourth variable under consideration is the one between d3i and d3v. To choose the fourth variable the study investigates the correlation among these 5 variables by performing a Pearson correlation test and the results are shown in Table 3. It can be observed that d1q is highly correlated ($\rho > |0.7|$) with d3q; d1v is highly correlated with d3v and d3i; and d3v is highly correlated with d3i in all four cases. As d3i has relatively lower correlation with other information variables, it has been selected along with d1q, d1v and d3q for the final model building with BBN. Lastly, to provide an overall idea about the variables, Table 4 illustrates their descriptive statistics based on their association with hazardous and normal traffic conditions.

### 4.2. The BBN models

As mentioned earlier, a BBN model consists of two parts — the graphical and the numerical parts. The outcome variable in the models is 'crash' with two categories representing the hazardous traffic condition and the normal traffic condition. Alongside the ramp flow, the

**Table 4**
Outcome of the logistic regression models.

| | Estimate | Std. err. | Z-value | Pr(>\|z\|) |
|---|---|---|---|---|
| *(a) Model 1: downstream of entrance ramp* | | | | |
| (Intercept) | −0.6312 | 0.2903 | −2.175 | 0.0296* |
| d1q | −0.0103 | 0.0015 | −7.028 | 2.09e−12*** |
| d1v | −0.0257 | 0.0032 | −7.999 | 1.25e−15*** |
| *(b) Model 2: downstream of entrance ramp* | | | | |
| (Intercept) | −2.7283 | 0.2847 | −9.582 | <2e−16*** |
| d3q | −0.0105 | 0.0017 | −6.290 | 3.16e−10*** |
| d3i | 1.83401 | 0.2346 | 7.818 | 5.35e−15*** |
| *(c) Model 1: downstream of exit ramp* | | | | |
| (Intercept) | −1.7589 | 0.3359 | −5.237 | 1.63e−07*** |
| d1q | −0.0060 | 0.0019 | −3.254 | 0.00114** |
| d1v | −0.0229 | 0.0037 | −6.202 | 5.57e−10*** |
| *(d) Model 2: downstream of exit ramp* | | | | |
| (Intercept) | −3.7835 | 0.3493 | −10.832 | <2e−16*** |
| d3q | −0.0044 | 0.0018 | −2.480 | 0.0131* |
| d3i | 1.9458 | 0.3101 | 6.276 | 3.48e−10*** |
| *(e) Model 1: upstream of entrance ramp* | | | | |
| (Intercept) | −1.2433 | 0.3162 | 3.932 | 8.44e−05*** |
| d1q | −0.0080 | 0.0014 | −5.861 | 4.60e−09*** |
| d1v | −0.0188 | 0.0034 | −5.536 | 3.10e−08*** |
| *(f) Model 2: upstream of entrance ramp* | | | | |
| (Intercept) | −2.7013 | 0.2683 | −10.069 | <2e−16*** |
| d3q | −0.0099 | 0.0016 | −6.179 | 6.43e−10*** |
| d3i | 1.7037 | 0.2490 | 6.843 | 7.77e−12*** |
| *(g) Model 1: upstream of exit ramp* | | | | |
| (Intercept) | −1.2298 | 0.3055 | −4.025 | 5.70e−05*** |
| d1q | −0.0093 | 0.0016 | −5.798 | 6.73e−09*** |
| d1v | −0.0185 | 0.0030 | −6.258 | 3.91e−10*** |
| *(h) Model 2: upstream of exit ramp* | | | | |
| (Intercept) | −3.4129 | 0.3738 | −9.130 | <2e−16*** |
| d3q | −0.0076 | 0.0017 | −4.561 | 5.09e−06*** |
| d3i | 2.5759 | 0.2796 | 9.214 | <2e−16*** |

Signif. codes: 0, '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1 and '' 1.

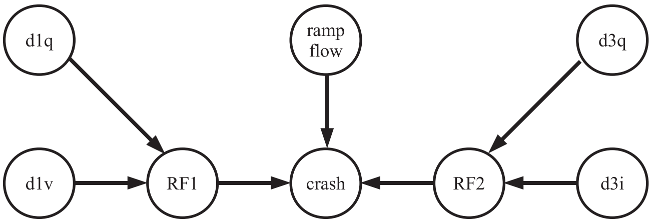
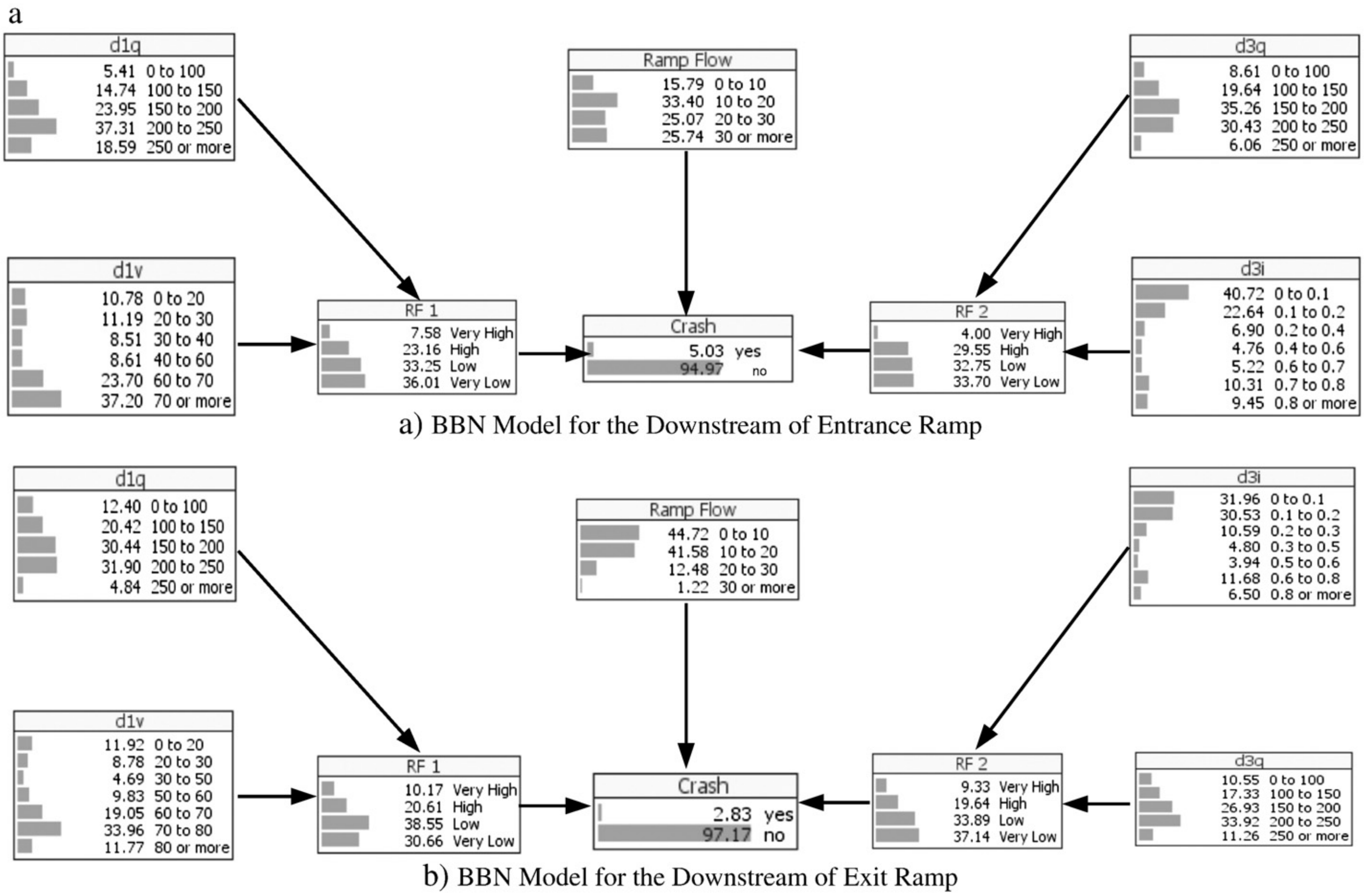


**Fig. 6.** Proposed graph for BBN models.

previous section has selected d1q, d1v, d3q and d3i as the other information variables. One possible way of designing the graphical model can be by drawing an arc from all the information variables to the outcome variable 'crash'. Though logically sound, it will have some major demerits in its numerical part. To exemplify, if each of the information variables has even only five categories each then the conditional probability table for the outcome variable 'crash' will have $5^5 = 3125$ cells for each of its outcomes (hazardous or normal) and thus increase the complexity of the model geometrically. Moreover, as the sample size is not large specially for the hazardous traffic conditions, very few cells of the hazardous traffic condition category in the conditional probability table of 'crash' will contain values. This problem can be addressed by applying the parent divorcing method (Section 3.2). For that, this study introduces two intermediate variables; risk factor 1 (RF1) and risk factor 2 (RF2) in between the information variables and the outcome variable as shown in Fig. 6. The RF1 and RF2 both have four categories to explain the risk of crash — very high, high, low and very low. Hence, the conditional probability table for crash now will have $4 \times 4 \times A = 64A$ cells for each of the two outcomes (A is the number of categories in d2q) and both RF1 and RF2 will have $B \times C \times 4$ cells where B and C are the number of categories for d1q and d1v for RF1 and d3q and d3i for RF2. Thus, the complexity of the conditional probability table for the outcome variable 'crash' can be substantially reduced.

The break points for these categories of RF1 and RF2 are calculated by first developing two binary logit models with d1q, d1v, crash (Model 1) and d3q, d3i, crash (Model 2) as the variables. Subsequently, the probability of crash for each data point has been calculated. The break points for the two intermediate variables have then been derived from these two models by reclassifying the continuous probability values into categories. The results of the two models for all four ramp vicinities are presented with Table 5. The calculation for the categorization of the intermediate variables is demonstrated briefly in Appendix A.

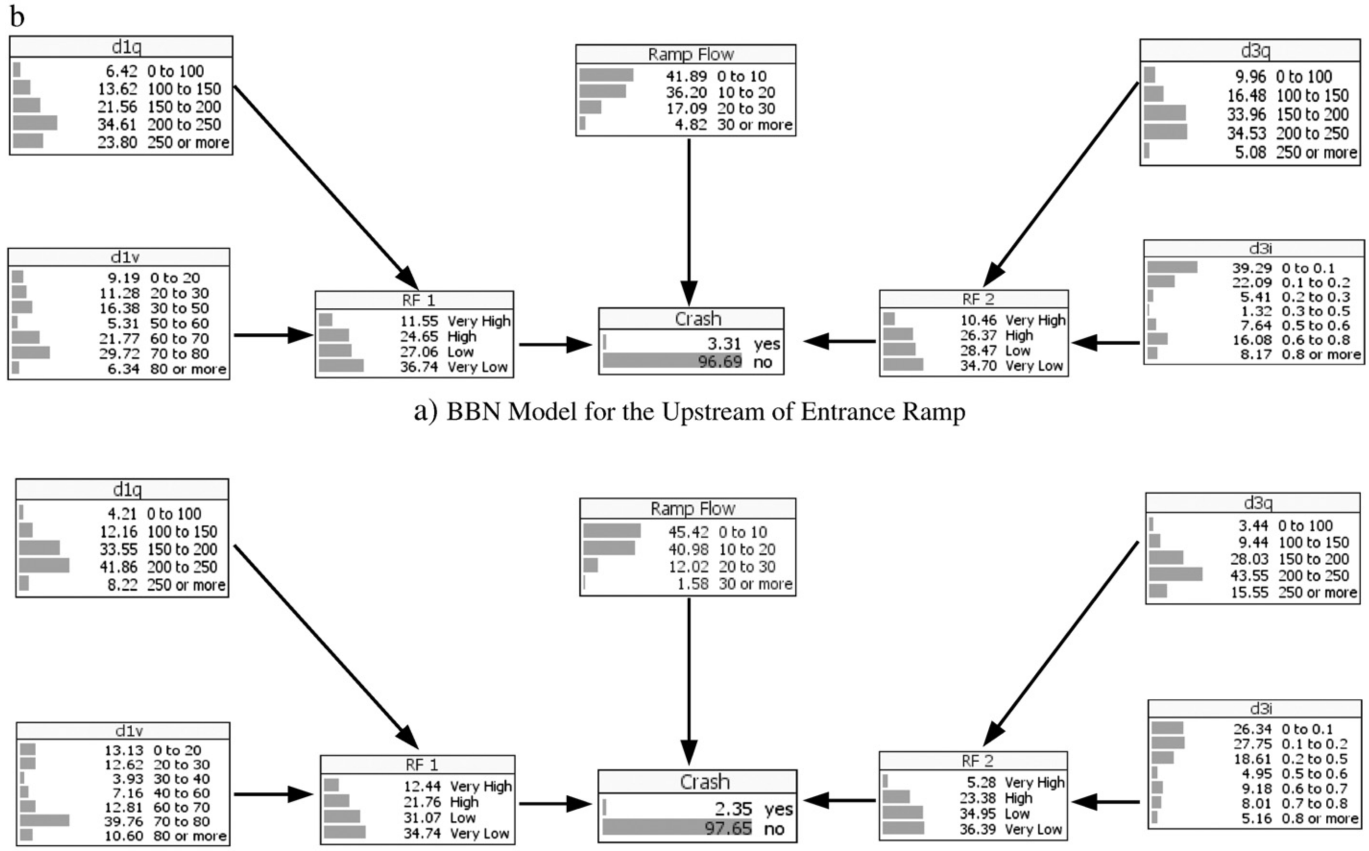
The next step in BBN involves generating the probability tables for the information variables without parents (P(d1q), P(d1v), P(d3q), P(d3i) and P(d2q)) and the conditional probability tables for those having parent nodes (here, P(RF1|d1q, d1v), P(RF2|d3q, d3i) and P(crash|RF1,RF2, d2q)). For this, it is necessary to categorize the continuous variables d1q, d1v, d3q, d3v and d2q. Histograms have been produced for all these variables considering their association both with hazardous and normal traffic condition to decide upon their final categories. Special care has been taken to ensure that the conditional probability tables for the child nodes contain at least one value for each of the conditions. Lastly, BBN models for the four ramp vicinities have been made using Hugin Expert [22] as the software tool following the methodology mentioned in Section 3.2. The final models are illustrated by Fig. 7. Although the four models use the same variable set, Fig. 7 suggests that they do not share the same categories for all



a) BBN Model for the Downstream of Entrance Ramp

b) BBN Model for the Downstream of Exit Ramp

**Fig. 7.** a. The final BBN models (a) downstream of entrance ramp and (b) downstream of exit ramp. b. The final BBN models (a) upstream of entrance ramp and (b) upstream of exit ramp.



a) BBN Model for the Upstream of Entrance Ramp

b) BBN Model for the Upstream of Exit Ramp

**Fig. 7** (*continued*).

the variables. The values in the conditional probability tables also vary substantially. Moreover, the overall probability of any traffic condition belonging to hazardous or normal traffic condition based on the given probabilities of different information variables (e.g., 5.03 in the model for the downstream of entrance ramp as illustrated by Fig. 7(a)) also differs among the models. These differences justify the importance of developing different models for the four ramp vicinities — downstream and upstream of the entrance and exit ramps.

### *4.3. Performance evaluation*

The performances of the newly built models have been evaluated with their corresponding datasets containing hazardous traffic conditions resulting from crashes that took place during the last two months in the study area. However, their corresponding normal traffic condition data have been extracted from throughout the study area (Table 2). Therefore, for every crash data, their two normal traffic condition data points from within the last two months of the study period were re-extracted to form the evaluation dataset. The final evaluation dataset contains 22, 19, 21 and 29 data points representing hazardous traffic conditions in downstream and upstream of entrance and exit ramps and their corresponding 44, 38, 42 and 58 normal traffic condition data points respectively. The prediction success has been presented in Table 6. The probability value that represents the chance of a data belonging to hazardous traffic condition when no other information is available (e.g., 5.03 for the downstream of entrance ramp as illustrated by Fig. 7(a)) has been used as the minimum baseline threshold. Later, higher threshold values have also been used to investigate the overall prediction performance. This measure ensures a very conservative approach in which maintaining sufficient accuracy in reducing false alarms (misclassifying normal traffic condition) has high priority. In general detection of hazardous traffic formation will bring benefit when it will be coupled with appropriate evasive measures. Hence, it is important to identify those conditions which pose abnormally high level of threat. Moreover, the research field dealing with real-time intervention designing is still in its infancy and several of the previous studies have recommended warning the road users through variable message signs [21]. Therefore, a low false alarm rate will mean fewer numbers of unnecessary warning signs which will eventually help in maintaining the trust of the road users.

Table 6 suggests that all the four models can predict more around 60% or more formation of hazardous traffic conditions for the initial threshold value. However, apart from the upstream of the exit ramps the other three ramp vicinities generated more than 25% false alarms for this threshold value. This study investigates how much prediction accuracy for the hazardous condition formation can be achieved by maintaining a false alarm rate less than 10–12%. The results suggest that the downstream of entrance ramps and the upstream of the exit ramps can still maintain classification accuracy over 50% under this boundary condition. However, the accuracy of the other two ramp vicinities for identifying hazardous traffic formation with the increased threshold value drops to just over 42%. There can be at least two probable reasons for this. Maneuvering in the ramps mainly takes place in the downstream of the entrance ramp (merging traffic) and the upstream of exit ramp (diverging traffic) causing more detectable interruptions in the mainline traffic flow. This may also be because a larger number of crash samples were available for modeling for these two locations.

## 5. Conclusion

The idea of predicting crash for a very short time window in near future by observing the traffic condition in real-time is one of the

**Table 5**
Descriptive statistics of the information variables.

| Detector: variable | Min | 1stQ | Median | Mean | 3rdQ | Max. |
|---|---|---|---|---|---|---|
| de: d1q (Crash = 1)[a] | 54.0 | 113.5 | 153.0 | 162.9 | 206.0 | 309.0 |
| de: d1q (Crash = 0) | 27.0 | 158.0 | 207.0 | 197.2 | 240.0 | 355.0 |
| de: d1v (Crash = 1) | 4.00 | 17.30 | 40.10 | 40.91 | 62.60 | 82.30 |
| de: d1v (Crash = 0) | 4.80 | 33.42 | 65.80 | 56.05 | 72.70 | 91.20 |
| de: d3q (Crash = 1) | 52.0 | 104.0 | 144.0 | 148.3 | 191.0 | 271.0 |
| de: d3q (Crash = 0) | 25.0 | 140.0 | 182.0 | 175.3 | 212.0 | 302.0 |
| de: d3i (Crash = 1) | 0.0000 | 0.1100 | 0.5400 | 0.4726 | 0.8400 | 0.9600 |
| de: d3i (Crash = 0) | 0.0000 | 0.0600 | 0.1100 | 0.2695 | 0.5500 | 0.9500 |
| dx: d1q (Crash = 1) | 29.0 | 115.2 | 153.0 | 153.2 | 198.8 | 274.0 |
| dx: d1q (Crash = 0) | 15.0 | 135.0 | 181.0 | 172.6 | 214.0 | 301.0 |
| dx: d1v (Crash = 1) | 4.10 | 20.30 | 43.50 | 43.29 | 66.88 | 88.10 |
| dx: d1v (Crash = 0) | 4.80 | 50.80 | 68.00 | 58.99 | 76.10 | 95.10 |
| dx: d3q (Crash = 1) | 33.0 | 125.8 | 171.0 | 169.8 | 213.2 | 291.0 |
| dx: d3q (Crash = 0) | 17.0 | 144.0 | 192.0 | 183.6 | 228.0 | 318.0 |
| dx: d3i (Crash = 1) | 0.0000 | 0.1400 | 0.3900 | 0.4251 | 0.7150 | 0.9300 |
| dx: d3i (Crash = 0) | 0.0000 | 0.0700 | 0.1400 | 0.2532 | 0.3400 | 0.9400 |
| ue: d1q (Crash = 1) | 31.0 | 119.5 | 173.0 | 173.1 | 227.0 | 292.0 |
| ue: d1q (Crash = 0) | 22.0 | 159.0 | 212.0 | 200.2 | 246.0 | 355.0 |
| ue: d1v (Crash = 1) | 6.20 | 25.50 | 53.35 | 46.49 | 68.05 | 83.00 |
| ue: d1v (Crash = 0) | 4.70 | 32.30 | 64.80 | 55.16 | 72.70 | 94.60 |
| ue: d3q (Crash = 1) | 29.0 | 108.8 | 155.0 | 151.9 | 198.0 | 257.0 |
| ue: d3q (Crash = 0) | 15.0 | 144.0 | 187.0 | 176.3 | 214.0 | 312.0 |
| ue: d3i (Crash = 1) | 0.0000 | 0.1100 | 0.4650 | 0.4546 | 0.7750 | 0.9400 |
| ue: d3i (Crash = 0) | 0.0000 | 0.0600 | 0.1200 | 0.2918 | 0.6700 | 0.9600 |
| ux: d1q (Crash = 1) | 34.0 | 127.0 | 176.0 | 169.9 | 206.5 | 292.0 |
| ux: d1q (Crash = 0) | 19.0 | 170.0 | 200.0 | 193.7 | 225.0 | 332.0 |
| ux: d1v (Crash = 1) | 4.50 | 18.65 | 37.40 | 43.44 | 71.85 | 87.70 |
| ux: d1v (Crash = 0) | 3.50 | 29.42 | 70.20 | 57.38 | 76.20 | 97.00 |
| ux: d3q (Crash = 1) | 36.0 | 144.5 | 186.0 | 181.1 | 221.5 | 307.0 |
| ux: d3q (Crash = 0) | 17.0 | 180.0 | 210.0 | 204.4 | 237.0 | 339.0 |
| ux: d3i (Crash = 1) | 0.0000 | 0.2100 | 0.6000 | 0.5289 | 0.8000 | 0.9500 |
| ux: d3i (Crash = 0) | 0.0000 | 0.0900 | 0.1700 | 0.2881 | 0.5400 | 0.9400 |

de = downstream entrance; dx = downstream exit; ue = upstream entrance; ux =upstream exit.

[a] Crash = 1 and 0 stand for hazardous and normal traffic conditions respectively.

**Table 6**
Evaluation of model performance.

| Location | Threshold | Hazardous | Normal | Hazardous (%) | Normal (%) |
|---|---|---|---|---|---|
| Downstream (entrance) | 5.03 | 13 | 26 | 59.09 | 59.09 |
| | 5.25 | 12 | 36 | 54.55 | 81.82 |
| | 5.50 | 11 | 41 | 50 | 93.18 |
| Downstream (exit) | 2.83 | 12 | 31 | 63.16 | 81.58 |
| | 3.00 | 11 | 32 | 57.89 | 84.21 |
| | 3.50 | 8 | 34 | 42.11 | 89.47 |
| Upstream (entrance) | 3.31 | 13 | 29 | 61.90 | 69.05 |
| | 3.50 | 9 | 36 | 42.86 | 85.71 |
| | 3.75 | 9 | 37 | 42.86 | 88.10 |
| Upstream (exit) | 2.35 | 22 | 43 | 75.86 | 74.14 |
| | 3.25 | 20 | 47 | 68.97 | 81.03 |
| | 3.75 | 16 | 52 | 55.17 | 89.66 |

emerging fields in proactive road safety management system. The application of such a system may become highly useful especially for urban expressways as they serve the peak hour demand of mega city traffic. Any crash during these peak hours normally has huge consequences and reducing even a fraction of it is highly desirable. The recent development enjoyed in the field of information technology has opened up the possibility of actual implementation of such a system in near future. The currently available models are yet theoretical and need to address several issues related to i) the variables to be used, ii) detectors to be used to extract data, iii) updating the model in course of time, iv) transferring existing models to other expressways, etc. To isolate these problems, the manuscripts identify the ramp vicinities as one of the most hazardous road sections on the urban expressways and present models that can predict the formation of hazardous traffic conditions in these areas in real-time. This can be considered more appropriate from the implementation point of view as ramp vicinities generally occupy a fraction of the whole expressway and it demands more attention from the drivers to drive in these areas as the number of decisions needed to be taken by them are higher. Thus, implementing a system like that for such a small but vulnerable area will be more productive from the point of view of the expressway authorities. Moreover, the manuscript also acknowledged that traffic conditions in the downstream and upstream of entrance and exit ramps may not be identical and thus, instead of making one universal model, it develops four separate models following the same methodology for these four ramp vicinities. Shibuya 3 and Shinjuku 4 routes under the jurisdiction of Tokyo Metropolitan Expressway Company Limited were chosen as the study area as they are two of the busiest expressways in Tokyo, they experience high number of crashes throughout the year and they are densely packed with detectors.

The study finds that approximately 55% of the reported crashes in the study area took place within 375 m from the ramp locations. The downstream of the entrance ramps and the upstream of the exit ramps experienced more crashes, too as compared to the other two ramp areas. This is also expected as substantial amount of maneuvering take place in these locations. For every ramp location, the study collected data from 3 detectors, one placed more than 300 m downstream, one placed more than 300 m upstream from the ramp location and the third one placed on the ramp. Apart from the 5 minute cumulative vehicle count, 5 minute heavy vehicle count, 5 minute average speed and occupancy the study has also introduced congestion index as one of the potential predictors of hazardous traffic condition. For the detector on ramp only the 5 minute cumulative vehicle count was available. The initial variable space contained 16 variables for model building. They were highly correlated in nature as well. RMNL, was introduced as a method to identify the top 4 most important variables from the complete variable space. RMNL makes a fusion of multinomial logit and random forest to come up with a modeling method that is free from the multi-collinearity issues of multinomial logit and can associate relative numeric values to express the importance of variables. The final variable space used for developing four separate models for the four ramp vicinities included the 5 minute vehicle count yielded by both the upstream and downstream detectors, the speed in downstream, congestion index in the upstream and the ramp flows. Next, BBN was applied to develop the final models. BBN is a popular real-time probabilistic prediction method in the field of artificial intelligence. It has some inherent properties that make it highly suitable for predicting crash in real-time. Road crashes are normally influenced by an assorted collection of factors. However, being a rare event, it is always difficult to find a large sample to develop the model with a large variable space. Moreover, data for all the potential variables may not be available at a time. Hence, the modeling method is required to have inherent capabilities to add new variables as well as partially update itself when new data become available. BBN has both the capabilities to update itself by introducing new variables as well as updating the model with partial data that will be available in future without needing to recalibrate the complete model. The newly developed BBN models could predict 50%, 42%, 43% and 55% of all the crashes from the evaluation dataset for the downstream of entrance ramp, downstream of exit ramp, upstream of entrance ramp and upstream of exit ramp respectively maintaining a false alarm rate less than or around 10%. At this moment, it is very difficult to compare the performance of the models presented in this manuscript with the previously proposed models. As far as the authors know this may be the first attempt in which real-time crash prediction models have been exclusively built for the areas near the ramps only. Although some previous models included presence of ramp as a variable they did not consider the varying traffic conditions in the different parts of the entrance and exit ramps. As the performance of the models was calculated based on data from the same expressway and that is also, for a short period of time, in actual scenario, initially, a variation of performance is expected. However, BBN has inherent capability to train the model in real-time as real-time data are being fed into it on regular basis. Hence, it can be expected that in course of time, the models will get accustomed to the condition of the location where it is being implemented and once it reaches accepted level of accuracy, expressway authorities can take the decision of applying it to administer various crash preventive interventions.

The proposed models have been specifically made for the ramp vicinities and may not be directly used for the long basic freeway segments of expressway. The model also did not consider variables related to weather. Moreover, the manuscript has kept its scope limited within predicting crashes in real-time and did not explain the underlying phenomena of crash. The study also did not propose countermeasures to bring the hazardous traffic condition back to normal. At present, the study only predicts the probability of future crash but does not indicate its type. It is recommended that future studies also incorporate crash reports and investigate if the crashes can be predicted along with their types. It is expected that the outcome of the study will reduce the gap between theory and practice for predicting crash in real-time and existing as well as future studies focusing on designing appropriate countermeasures, such as Hossain and Muromachi [14], will be able to utilize the proposed models to evaluate their efficacy.

## Acknowledgment

The authors would like to express their gratitude to Tokyo Metropolitan Expressway Company Limited for facilitating the study.

## Appendix A. Sample calculation demonstrating the selection of the break points for the intermediate variables

The procedure followed to calculate the risk factor 1 (RF1) and 2 (RF2) are same for all four ramp vicinities. Here a sample calculation has been presented to explain the steps involved in selecting the break points to classify RF1 for the downstream of entrance ramp.

Let a randomly selected data point associated with hazardous traffic condition in the dataset used for modeling has values:

d1q = 113 (5 minute cumulative vehicle count yielded by the downstream detector).

**Table A1**
Descriptive statistics for excess probability after applying Rule 1 and Rule 2 for Model 1: downstream of entrance ramp.

| Rule | Min | 1stQ | Median | Mean | 3rdQ | Max. |
|---|---|---|---|---|---|---|
| Rule 1 | 0.0003 | 0.0099 | 0.0298 | 0.0435 | 0.0620 | 0.1856 |
| Rule 2[a] | 0.0198 | 0.0146 | 0.0121 | 0.0110 | 0.0078 | 0.0000 |

[a] All the values are negative.

d1v = 9.3 (5 minute average speed in kilometers per hour yielded by the downstream detector).

Now, from the parameters presented by Table 4(a), the probability of a traffic condition belonging to hazardous category can be calculated as $= 1 / (1 + \exp(-(-0.6312 - 0.0103 * 113 - 0.0257 * 19.3))) = 1 / (1 + \exp(-2.03411)) = 0.11567$.

Now, the dataset used for modeling has 143 and 6182 data points associated with hazardous and normal traffic conditions respectively. Thus, the average probability of a data point belonging to the hazardous traffic condition is $= (143)/(143 + 6182) = 0.0226$.

Thus, the excess probability for the data point is $= 0.11567 - 0.0226 = 0.093$.

The same procedure has been repeated for the complete dataset to calculate the excess probability of every data point. Subsequently, the descriptive statistics (mean, median, 1st quartile, 3rd quartile, minimum, maximum) of all these excess probabilities have been calculated with these rules:

Rule 1: for crash = 'yes' and excess probability > 0.
Rule 2: for crash = 'no' and excess probability < 0.

The results are presented in Table A1. It can be observed that, the median value 0.0298 (median value for Rule 1) is used as the break point between 'very high' risk and 'high' risk, 0.0 is used as the break point between 'high' risk and 'low' risk. Accordingly −0.0121 is used as break point between 'low risk' and 'very low' risk for RF1 of Model 1 for the downstream of entrance ramp. This conservative approach ensures that a data point gets classified as 'high risk' only when its probability to belong to hazardous traffic condition is substantially higher than the average probability and vice versa. Continuing with the illustration, the calculated excess probability for the data point is 0.093 which is higher than 0.0298 and thus gets classified as 'very high'.

## References

[1] M. Abdel-Aty, A. Pande, Classification of real-time traffic speed patterns to predict crashes on freeways, Proceedings of the 83rd Annual Meeting of Transportation Research Board, Washington, DC, 2004.

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

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

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

[5] M. Abdel-Aty, R. Pemmanaboina, L. Hsia, Assessing crash occurrence on urban freeways by applying a system of interrelated equations, Transportation Research Record 1953 (2006) 1–9.

[6] M. Abdel-Aty, A. Pande, A. Das, W.J. Knibbe, Assessing safety on dutch freeways with data from infrastructural-based intelligent transportation systems, Transportation Research Record 2083 (2008) 153–161.

[7] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth, Inc., Pacific Grove, CA, 1984.

[8] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.

[9] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[10] P. Dalgaard, Introductory Statistics with R, Springer, NY, 2008.

[11] C. Dias, M. Miska, M. Kuwahara, H. Warita, Relationship between congestion and traffic accidents on expressways: an investigation with bayesian belief networks, Proceedings of 40th Annual Meeting of Infrastructure Planning (JSCE), Japan, 2009.

[12] M. Hossain, Y. Muromachi, Development of real-time crash prediction model by Bayesian network, Journal of Traffic Engineering, Japan Society of Traffic Engineering 42 (2) (2012) 39–44.

[13] M. Hossain, Y. Muromachi, A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways, Accident Analysis and Prevention 45 (2012) 373–381.

[14] M. Hossain, Y. Muromachi, Understanding crash mechanism and selecting appropriate interventions for real-time hazard mitigation on urban expressways, Transportation Research Record 2213 (2011) 53–62.

[15] M. Hossain, Y. Muromachi, Optimum detector spacing for real-time monioring of hazardous locations on urban expressways, Journal of Infrastructure Planning Review (JSCE) 27 (2010) 1045–1054.

[16] M. Hossain, Y. Muromachi, Development of a real-time crash prediction model for urban expressway, Journal of the Eastern Asia Society for Transportation Studies 8 (2010) 2092–2107.

[17] F.V. Jensen, T.D. Nielsen, Bayesian Networks and Decision Graphs, Springer, NY, 2007.

[18] P.P. Jovanis, H.L. Chang, Modeling the relationship of accidents to miles traveled, Transportation Research Record 1068 (1986) 42–51.

[19] C. Lee, F. Szccomanno, B. Hellinga, Analysis of crash precursors on instrumented freeways, Transportation Research Record 1784 (2002) 1–8.

[20] C. Lee, B. Hellinga, F. Saccomanno, Real-time crash prediction model for the application to crash prevention in freeway traffic, Transportation Research Record 1840 (2003) 67–77.

[21] C. Lee, M. Abdel-Aty, Testing effects of warning messages and variable speed limits on driver behavior using driving simulator, Transportation Research Record 2069 (2008) 55–64.

[22] A.L. Madsen, M. Lang, U.B. Kjaerulff, F. Jensen, Hugin tool for learning Bayesian networks, Lecture Notes in Computer Science 2711 (2004) 594–605.

[23] C. Oh, J. Oh, S. Ritchie, M. Chang, Real-time estimation of freeway accident likelihood, Proceedings of the 80th Annual Meeting of Transportation Research Board, Washington, DC, 2001.

[24] J. Oh, C. Oh, S. Ritchie, M. Chang, Real time estimation of accident likelihood for safety enhancement, Journal of Transportation Engineering (ASCE) 131 (5) (2005) 358–363.

[25] C. Oh, J. Oh, S. Ritchie, M. Chang, Real time hazardous traffic condition warning system: framework and evaluation, IEEE Transactions on Intelligent Transportation Systems 6 (3) (2005) 265–272.

[26] K.G. Olesen, U. Kjaerulff, F. Jensen, F.V. Jensen, B. Falck, S. Andreassen, S.K. Andersen, A MUNIN network for the median nerve — a case study on loops, Applied Artificial Intelligence 3 (1989).

[27] A. Pande, M. Abdel-Aty, A freeway safety strategy for advanced proactive traffic management, Journal of Intelligent Transportation Systems 9 (3) (2005) 145–158.

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

[29] A. Prinzie, D.V. Poel, Random forests for multiclass classification: random multinomial logit, Expert Systems with Applications 34 (3) (2008) 1721–1732.

[30] R. Shapire, Y. Freund, P. Bartlett, W. Lee, Boosting the margin: a new explanation for the effectiveness of voting methods, The Annals of Statistics 26 (5) (1998) 1651–1686.

[31] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinformatics 8 (25) (2007).

[32] Z. Zheng, S. Ahn, C.M. Monsere, Impact of traffic oscillations on freeway crash occurrences, Accident Analysis and Prevention 42 (2) (2010) 626–636.