Accident Analysis and Prevention 57 (2013) 17–29

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Understanding crash mechanism on urban expressways using high-resolution traffic data

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 3 February 2011
Received in revised form 14 March 2013
Accepted 21 March 2013

Keywords:
Urban expressway
Crash mechanism
High-resolution traffic data
Random multinomial logit
Classification and regression trees

ABSTRACT

Urban expressways play a vital role in the modern mega cities by serving peak hour traffic alongside reducing travel time for moderate to long distance intra-city trips. Thus, ensuring safety on these roads holds high priority. Little knowledge has been acquired till date regarding crash mechanism on these roads. This study uses high-resolution traffic data collected from the detectors to identify factors influencing crash. It also identifies traffic patterns associated with different types of crashes and explains crash phenomena thereby. Unlike most of the previous studies on conventional expressways, the research separately investigates the basic freeway segments (BFS) and the ramp areas. The study employs random multinomial logit, a random forest of logit models, to rank the variables; expectation maximization clustering algorithm to identify crash prone traffic patterns and classification and regression trees to explain crash phenomena. As accentuated by the study outcome, crash mechanism is not generic throughout the expressway and it varies from the BFS to the ramp vicinities. The level of congestion and speed difference between upstream and downstream traffic best explains crashes and their types for the BFS, whereas, the ramp flow has the highest influence in determining the types of crashes within the ramp vicinities. The paper also discusses about the applicability of different countermeasures, such as, variable speed limits, temporary restriction on lane changing, posting warnings, etc., to attenuate different patterns of hazardous traffic conditions. The study outcome can be utilized in designing location and traffic condition specific proactive road safety management systems for urban expressways.

1. Background

Urban expressways are generally constructed as viaducts running above the surface roads to tackle the heavy traffic during peak periods and reduce travel time of moderate to long distance intra-city trips. They are significantly different from the conventional arterial, collector or local roads in their role and geometric construction. Unlike the later classifications, they trade mobility with reduced accessibility and provide for 100% through movement of traffic in general. Normally they are access controlled, divided and sometimes even elevated. Due to their functional efficacy, they are gaining increasing popularity among the city planners in modern days. At present, it is hard to find a mega city in the developed world without an urban expressway. These urban expressways mainly serve the city traffic and often have few lanes (in many cases only two in each direction) with low lane width. They accommodate higher number of closely spaced ramps as compared to the conventional expressways and often remain heavily congested during the peak hours. Therefore, occurrence of crashes on urban expressways is associated with high consequences. These crashes not only waste valuable time of the travelers but also make the rescue efforts and consequent activities such as road clearance, repair and maintenance quite challenging due to their geometric and traffic constrains. With the burgeoning global trend of urbanization, it is certain that the city authorities will continue constructing many more urban expressways in the mega cities of both developed and developing countries. Hence, it is becoming increasingly important to understand safety related issues concerning and specific to the urban expressways and keep the number of crashes as well as their consequences as low as possible thereby.

Researchers over the past two decades have conducted significant number of studies to identify factors influencing crash (Fridstrom et al., 1995; Miaou and Song, 2005) and developed crash prediction models to calculate the frequency and associated severity of crash on conventional expressways (Khan et al., 1999; Caliendo et al., 2007). Several analogous studies have underscored positive correlations between traffic flow variables and road crashes (Cedar and Livneh, 1982; Cedar, 1982; Frantzeskakis and Iordanis, 1987) that brought long-term safety benefits by improving geometric designs, road side environment and helping in

* Corresponding author. Tel.: +81 080 3509 7060; fax: +81 45 924 5524.
E-mail addresses: moinul048i@yahoo.com, moinul033@gmail.com (M. Hossain), ymuro@enveng.titech.ac.jp (Y. Muromachi).

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

0001-4575/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2013.03.024

decision making for budget allocation, albeit the countermeasures were rather reactive in nature (Oh et al., 2001; Lee et al., 2003). They also ignored the complex interaction among traffic flow variables that may have abetted crashes. This is as they employed highly aggregated traffic data (e.g., hourly, daily or yearly flow) which could not capture the suddenly developed hazardous traffic conditions that could lead to a road crash. Recently, with the enhanced data collection, storage and analysis capabilities, researchers have started paying attention in developing proactive road safety management systems for expressways using high-resolution real-time traffic data. Several real-time crash prediction models have been proposed based on the hypothesis that the probability of a crash on a specific road section can be predicted for a very short time window using the instantaneous traffic flow data (Lee et al., 2002, 2003; Golob et al., 2003; Pande and Abdel-Aty, 2005). This opened the possibility to develop proactive road safety management systems which may even be able to prevent some crashes that would have taken place otherwise (Lee et al., 2002, 2003; Abdel-Aty and Pande, 2004; Abdel-Aty and Abdalla, 2004; Oh et al., 2005a,b; Abdel-Aty et al., 2006a,b; Dias et al., 2009; Hossain and Muromachi, 2010b). Jang et al. (2012) extended the study horizon by introducing a real-time collision warning system for the intersections where conditions related to vulnerable line of site and/or traffic violation can be observed. Christoforou et al. (2012) in their studies have determined crash probability along with associated crash severity. However, these studies were focused on improving the prediction capability rather than providing insight into crash phenomena. Among the studies related to identifying the traffic variables leading to crash, Abdel-Aty et al. (2005) ascertained that crashes occur in high speed and low speed scenarios. While the former is caused by quick formation and subsequent dissipation of queues causing a backward shock wave, the later is due to a disruption in the downstream that propagates a shock wave to the upstream impending driving errors. With a similar approach but including only rear-end crash data, Pande and Abdel-Aty (2006a) affirmed that crashes are related to coefficient of variation in speed and average occupancy under extended congestion. They also found that the high speed crashes were more explainable with average speed and occupancy in a downstream detector. They mentioned that presence of ramp in the downstream have impact on crash but did not shed light on the types of ramps and their relative vicinity. Two simultaneous studies were conducted on the same study area (I-4, Ontario, FL, USA) for lane-changing related collisions and it was found that average speeds at upstream and downstream together with difference in occupancy on adjacent lanes and standard deviation of volume and speed at a location downstream of the crash point are the major contributing factors (Pande and Abdel-Aty, 2006b; Lee et al., 2006). Dias et al. (2009) introduced level of congestion rather than the aggregated speed of vehicles as a predictor and affirmed a positive correlation between congestion and crash risk. Zheng et al. (2010) considered only congested traffic condition and used matched control logistic regression to prove that traffic oscillations contribute to crash. Christoforou et al. (2011) utilized real-time traffic data to associate different traffic parameters with various crash types. Xu et al. (2012) suggested that traffic characteristics leading to crash vary substantially between congested and uncongested situations. The studies existing were more concerned about identifying the factors and placed little or no concentration on why and how these factors contribute to a crash. They in most cases did not verify if the factors vary for the basic freeway segments (BFS) and ramp areas. McCartt et al. (2002) found different crash types and characteristics dominating different types of ramps. Chen et al. (2009, 2010) found significant safety impact even for off ramps of freeways when they had different number and arrangements of lanes. Due to high variation in ramp density between conventional expressways and urban expressways, the relevance and transferability of the findings of these studies to urban expressways may not be justified adequately. Thus, it is important to investigate if the existing findings are generic to all kinds of expressways or if they differ significantly.

This study employs high resolution detector data to identify the traffic patterns impending hazardous driving conditions. Unlike the previous studies, this study separates the road sections of the urban expressways into five groups – the basic freeway segments (BFS) and areas near downstream (d/s) and upstream (u/s) of the on (entrance) and off (exit) ramps and attempts to identify generic crash prone traffic patterns for each of these groups. The major objectives of this study are to – (i) identify and rank factors leading to crash, (ii) apply latest data mining techniques to identify traffic patterns associated with crash, and (iii) explain crash mechanism based on the research findings.

The manuscript is organized into five sections. The first section has explained the background, rationale and stated the objectives of the research. The second section discusses the study area, experimental designs for different road sections, and data extraction and processing steps. The third section explains the scientific methods used. Section 4 presents the analysis and results. The manuscript concludes with the explanation of the findings, limitations and future scopes.

2. Data preparation

2.1. Study area

The study requires data from urban expressways having closely spaced detectors with moderate standard deviation. Additionally, it expects sufficient number of corresponding crash samples as it separately investigates crash patterns for the basic freeway segments (BFS) and ramp areas. Shibuya 3 and Shinjuku 4 routes of Tokyo Metropolitan Expressway are chosen as the study area as they have relatively uniformly spaced detectors (approximately 250 m), making them two of the most sophisticatedly instrumented urban expressways in the world. They are respectively 11.9 km and 13.5 km in length and they harbor altogether 210 detectors, 14 on and 15 off ramps in both the directions. The detectors store data of speed, vehicle count, occupancy and number of heavy vehicles for each 8 ms round the clock. The expressway authority later aggregates the data of each station for every 5 min. Therefore, the supplied dataset contains 5-min vehicle count, heavy vehicle count, average speed and occupancy data. The study has collected detector data from December, 2007 to October, 2009 for both the routes. However, the supplied crash data encompass a time period from December, 2007 to March, 2008 for Shibuya 3 and from December, 2007 to November, 2008 for Shinjuku 4 route. The crash data contain information on date, time in minutes, location in nearest 10 m, vehicles involved, type of crash, etc. However, the severity data are not provided. The dataset contains 1141 crash samples. The expressway authority has emphasized that the reliability and precision of the data are high as major parts of the routes are under constant surveillance through cameras and an array of safety vehicles are patrolling round the clock. Likewise, being two of the busiest urban expressways with only two lanes in each direction ensures that any crash occurrence on the road gets detected quite fast.

2.2. Experimental design

The study area has been classified into five groups: BFS, u/s of off ramp, d/s of off ramp, u/s of on ramp and d/s of on ramp. Areas near ramps have been demarcated by zones encompassing 375 m u/s and d/s from the gore area. Rest of the road sections of the

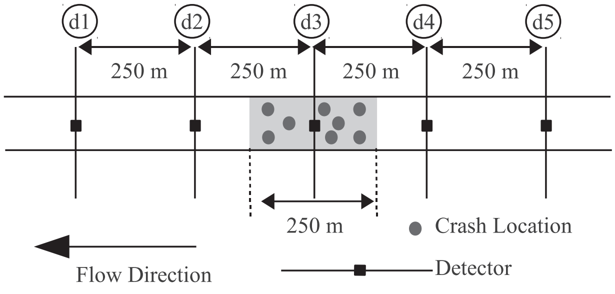

Fig. 1. Experimental design for data extraction (BFS).

expressways have been considered as the BFS. The value was decided by separating the whole road section into 25 m small sections and calculating the number of crashes for each section. It was found that the crash frequency is significantly higher in the areas 375 m u/s or d/s from the gore area, i.e., around 55% of the crashes in the study area took place within 375 m from the ramp locations. At this point, the question rises about which detectors best capture the pre-crash traffic condition for a crash data instance. Literature review implies that variation in traffic flow variables in the upstream and downstream, mainly when a fast moving upstream confronts a congested downstream, may build up hazardous traffic condition (Abdel-Aty et al., 2005; Pande and Abdel-Aty, 2006a,b; Lee et al., 2006). Most of these studies were conducted in North America where the inter detector spacing was too high and in many cases, not so uniform either. Likewise, this drawback bounded them to refer to the terms 'upstream' and 'downstream' detectors in a relative spatial relationship based on order of occurrence rather than in the unit of distance. Abdel-Aty et al. (2008) in their paper using Dutch freeway data, where inter-detector spacing was always less than 800 m, identified the detector locations with distance. However, they did not consider the influence of ramps. Hence, it may not be appropriate to directly adopt a detector or a combination of detectors endorsed by previous studies for extracting pre-crash traffic data. To address this problem, this paper has subdivided the BFS into 250-m sections which is also approximately equal to the inter-detector spacing on the study area. Thus, almost every section harbors one detector within it. Locations of all the crash points are then identified on the road layout and each crash point is associated with its corresponding 250-m section. Further, for every section, two upstream, two downstream and the detector within the section have been identified (d1, d2, d3, d4 and d5 in Fig. 1). Pre-crash condition data for each crash instance is collected by extracting traffic flow data from the five detectors associated with the section where the crash took place.

In case of the ramp areas, the detector arrangements selected for data extraction are illustrated by Fig. 2.

Normally off and on ramps in urban expressways are placed in pairs. Therefore, it was difficult to consider detectors beyond 375 m upstream or downstream from the location of the ramp to collect pre-crash traffic data. For each of the four scenarios (u/s-off, d/s-on, u/s-on and d/s-off), data have been extracted from three detectors (d1, d2 and d3) whose special arrangement is shown in Fig. 2. To expound more, d1 is the downstream detector placed more than 300 m downstream from the ramp; d2 is the detector on the ramp and d3 is the upstream detector places more than 300 m upstream from the ramp and all the crash locations within 375 m upstream or downstream of the ramp are considered to be crashes within the ramp vicinity. It is important to mention here that there were some locations where the off and on ramps are too close to each other. In those locations, the distance between the ramps has been divided into half and crashes taking place in any half are associated with the nearest ramp from them. It is possible that in such cases the impact of other ramp may also influence the crash characteristics. However, this study does not consider the combined impact of closely spaced ramps on crashes.

2.3. Data extraction and filtration

The data extraction process is commenced by defining pre-crash and normal traffic condition. The normal traffic condition data are extracted to check the exposure for each crash pattern so that a safe traffic condition does not wrongly get classified as dangerous. The pre-crash traffic conditions are defined as a 5-min time period ending at least 4 min before the recorded crash time. Previous studies have demonstrated that traffic condition varies significantly within just before and after the crash. Therefore, overlapping of post crash traffic condition with the pre-crash traffic condition is not desirable. For this, the 4-min gap from the reported crash time has been maintained intentionally to keep the pre-crash condition pure. Also, a major use of the knowledge gained about crash mechanism from high resolution detector data is to develop real-time crash prediction models that can be coupled with appropriate interventions to bring the hazardous traffic condition back to normal. Hence, it is desirable to maintain a gap between detection and the time of occurrence of the crash to let the intervention yield positive improvement of the traffic condition. Normal traffic conditions for the crash instance are represented with the same 5 min traffic flow data collected from the same set of detectors and the same day of week for throughout the year when no crash occurred. For example, if a crash had occurred on the 28th August, 2008 at 1:54:00 pm, then a time period on the same day starting from 1:45:00 pm to 1:50:00 pm will be classified as pre-crash traffic condition. The 28th August, 2008 is a Thursday. Continuing with the same example, for the normal traffic condition corresponding to the crash, all the data for the corresponding detectors from 1:45:00 pm till 1:50:00 pm for all Thursdays within the study period will be classified as normal

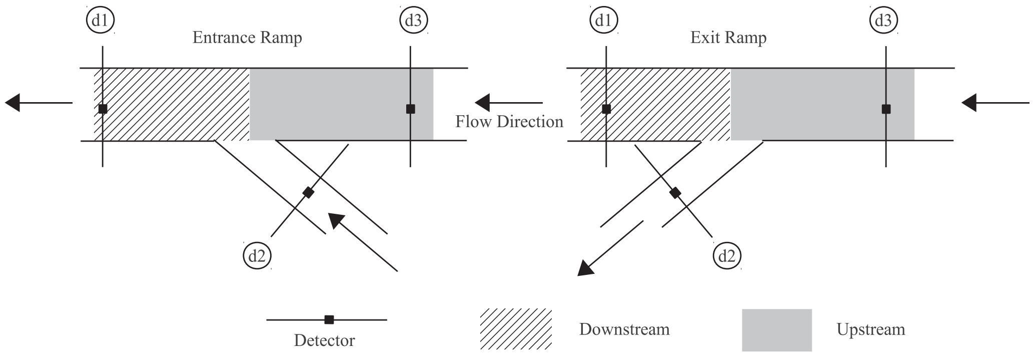

Fig. 2. Experimental design for data extraction (ramp areas).

Table 1
Sample size of crash and normal traffic conditions for different groups.

Group	Crash	Normal
Basic freeway segments	189	6478
u/s-off ramp	86	3699
d/s-off ramp	98	4066
u/s-on ramp	110	3628
d/s-on ramp	120	4148

traffic condition. Thus, the list will include 21st, 14th, August in the year 2008, etc. To ensure high integrity of the normal traffic condition data, all those records falling within 1 h before or after a crash occurring on the same route are also discarded. Each detector on the main stream yields data regarding speed, flow, occupancy and number of heavy vehicles. The detectors on the ramp keep only flow data. Thus, for each crash on the BFS, data are extracted for 20 variables (5 detectors × 4 variables = 20 variables). Similarly, for every crash in the ramp area, data of 9 variables are amassed (2 mainstream detectors × 4 variables + 1 detector on ramp × 1 variable = 9 variables). Lastly, detectors in many occasions fail to yield data for all the variables. It is important for the study that the data considered are complete and do not have any missing values. For this, records with missing values are further discarded from the dataset. Table 1 presents the final dataset used in the study.

2.4. Data preparation

Dias et al. (2009) in their study introduced Congestion Index (CI) as a variable instead of directly using the speed of the stream. They also employed the data from Tokyo Metropolitan Expressways. However, their study encompassed only one week detector data which are freely distributed by the expressway authority every year. Although their primary reason to use CI was to find the relationship between congestion and crash risk, it actually holds more significance. As compared to the conventional expressways, urban expressways have a relatively lower speed limit and it varies from section to section as they needs to accommodate difficult geometric configurations and harbor more frequent ramps. In case of Tokyo Metropolitan Expressways, the speed limit varies from 80 km/h to as low as 50 km/h. It is clearly evident that when a speed of 70 km/h is considered as low for a section with speed limit of 80 km/h, it is substantially high for a section that permits maximum 50 km/h. Therefore, CI becomes a more relative term in case of urban expressways as compared to the conventional expressways. Dias et al. (2009) calculated CI as:

$$\text{Congestion Index (CI)} = \frac{\text{Free Flow Speed} - \text{Speed}}{\text{Free Flow Speed}};\quad \text{when CI} > 0 = 0;\ \text{when CI} \leq 0 \tag{1}$$

Thus, CI is introduced as an extra variable along side flow, speed, heavy vehicle count and occupancy for each detector in this study. The free flow speed is calculated for each detector by visual observation of speed-flow, flow-occupancy and speed-occupancy diagrams (Fig. 3). Speed histograms have been prepared for detectors where the free flow speed cannot be easily ascertained from the aforementioned diagrams.

Additionally, relative differences in traffic flow parameters between the upstream and the downstream traffic holds special interest as several previous studies accentuated the generation of shock waves as antecedents of hazardous traffic condition. Therefore, the differences between the parameters of the upstream and downstream detectors have also been captured by introducing them as new variables. To illustrate more, for the BFS, speed difference between detector 1 and detector 4 creates a new variable; flow difference between detector 2 and detector 4 creates another new variable and so on. At the end the BFS contained information on 65 variables for each pre-crash and normal traffic condition record in the database. Among these 25 have been generated independently by the 5 detectors (5 detectors × 5 variables in each detector = 25) and the rest 40 have been generated by calculating their longitudinal differences. Likewise, the number of variables is 16 for each of the 4 groups in the ramp areas. Among these 16 variables, 10 are generated independently by the upstream and downstream detectors (2 detectors × 5 variables in each detector = 10), 5 are generated by their longitudinal differences and the last one is the ramp flow. Table 2 enumerates the variables considered for the study along with their short description. For easier understanding, the variables directly being yielded by the detectors have been coded as ‘**dxz**’ where ‘d’ represents detector, ‘**x**’ represents the detector number (here, **x** within {1, 2, 3, 4, 5} for the BFS and {1, 2, 3} for the ramp vicinities) and ‘**z**’ represents the type of data indicated with ‘q’, ‘p’, ‘v’, ‘o’, ‘i’ in short such as q = vehicle count, p = heavy vehicle count, v = speed, o = occupancy and i = congestion index. Variables representing the longitudinal differences are coded as ‘**dxyz**’ where ‘**x**’ and ‘**y**’ are the location codes of the downstream and upstream detectors respectively. To elaborate more, the term ‘d5i’ refers to the congestion index in the location of detector 5 (Fig. 1) and ‘d24o’ stands for the difference in occupancy between the location of detectors 2 and 4 (Fig. 1), and so on. It can be noticed that the codes

Table 2
List of variables with description.

Variables	Location	Description
d1q, d1p, d1v, d1o, d1i; d2q, d2p, d2v, d2o, d2i; d3q, d3p, d3v, d3o, d3i; d4q, d4p, d4v, d4o, d4i; d5q, d5p, d5v, d5o, d5i (Total 25)	BFS	5-min cumulative vehicle count, 5-min cumulative heavy vehicle count, 5-min average speed, 5-min average occupancy and 5-min congestion index calculated for each of the five detectors for the basic freeway segments
d13q, d13p, d13v, d13o, d13i; d23q, d23p, d23v, d23o, d23i; d34q, d34p, d34v, d34o, d34i; d35q, d35p, d35v, d35o, d35i; d14q, d14p, d14v, d14o, d14i; d24q, d24p, d24v, d24o, d24i; d15q, d15p, d15v, d15o, d15i; d25q, d25p, d25v, d25o, d25i (Total 40)	BFS	The longitudinal differences of the above mentioned variables between the u/s and d/s detector combinations for the basic freeway segments
d1q, d1p, d1v, d1o, d1i; d3q, d3p, d3v, d3o, d3i; (Total 10)	Ramp	Similar to those of the basic freeway segments but represents the ramp vicinities
d2q; d13q, d13p, d13v, d13o, d13i (Total 5)	Ramp Ramp	Ramp flow The longitudinal differences of the above mentioned variables between the u/s and d/s detector combinations for the ramp vicinities

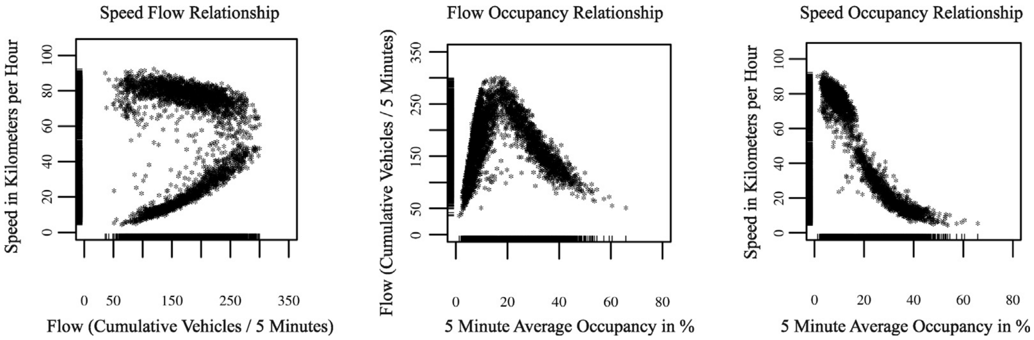

Fig. 3. Basic traffic flow relationships for detector no. 030119 (Shibuya 3, inbound).

used to refer some of the variables in the BFS and the ramp vicinities are same. To avoid confusions, the variables should always be realized based on the context of the discussion (e.g., if it is within the sections discussing about the BFS, d1q should be identified as the 5-min cumulative vehicle count yielded by the detector in d1 position in the BFS). In cases where the discussion about the BFS and the ramp areas overlap, the variables have been addressed with their specific location (e.g., BFS, d/s of on ramp, etc.).

3. Methodology

Random multinomial logit (RMNL) is used to fathom the importance of variables and clustering is used to detect traffic patterns with high crash rate association. Besides, the clusters with high crash rate are investigated with classification and regression tree (CART) considering the types of crashes to clearly understand the phenomena that might have indulged the crash prone situation.

3.1. *Random multinomial logit (RMNL)*

The variable space for the BFS consists of 65 predictors. This is too large for easily interpreting the outcomes of data mining methods such as clustering or CART. Therefore, it is important to devise a way to choose the most important variables in a scientific manner. Most of the previous studies used engineering judgments, literature review and experience for this. Some applied random forest (RF) (Abdel-Aty et al., 2008) and logistic regression (LR) (Hossain and Muromachi, 2010a) for variable selection. Both methods have their own merits and demerits. LR is a robust and popular classifier with sound statistical background. However, it suffers from issues concerning dimensionality when a large feature space is being addressed. On the contrary, RF (Breiman, 2001) is quite capable of handling the multicollinearity issue of large feature spaces by using bootstrap sampling and selecting variables randomly. RF also incorporates an algorithm to measure the variable importance, albeit can be biased toward the variables having larger number of categories (Strobl et al., 2007). This study amalgamates the benefits of both the methods by employing random multinomial logit (RMNL), a RF of LR trees instead of classification trees, and calculates the variable importance following the algorithm of RF. The method was first introduced by Prinzie and Poel (2008) and was found outperforming both RF and LR in the customer relationship management domain. This may be the first attempt to use RMNL in transportation research. For this, a moderately detailed explanation is provided in the following subsections.

Currently, RF is considered as one of the latest and most efficient methods in evaluating and ranking variable importance (Harb et al., 2009). It generates an array of datasets from the original data using two well-known methods in ensemble learning – boosting (Shapire et al., 1998) and bagging (Breiman, 1996), which are in the next step used for growing CART trees (Breiman et al., 1984) in an innovative way to rank the variables as well as make predictions. In case of RMNL, the CART trees are replaced with LR models. To illustrate the concept, let L be the complete dataset with M predictors and N records and B the total number of LR trees in the RMNL. Now, let L_b be the bth bootstrap sample created by randomly selecting n samples with replacement from L. The data that are not selected, i.e., $L-L_b$, are called the out of bag data (OOB) of bth bootstrap sample. Now, for the bth tree, instead of growing a CART tree, m predictors are randomly selected from M predictor space and a LR model is built. Let the new OOB dataset with m predictors only be noted as $(L-L_b)'$. Next, the $(L-L_b)'$ is used to calculate the misclassification rate r_b of tree T_b. Subsequently, the values of the jth predictor of m predictors in $(L-L_b)'$ are permuted and the new dataset is used to calculate the misclassification rate r_b^j. Here, $|r_b - r_b^j|$ is the variable importance V_j of the jth variable in the bth tree. The process is repeated for B trees and the final variable importance is calculated by averaging the V_j of each variable ($j = 1$ to M). The concept underlying variable importance is, if the value of a variable is miscalculated and it creates the highest error in prediction then it must be the most important variable.

RMNL is still a very young concept in ensemble learning and at this moment there is no commercial, open source or free package available to perform the modeling. Therefore, a new script has been written using R program (Dalgaard, 2008) to implement RMNL for this study.

3.2. *Clustering (expectation maximization algorithm)*

Clustering techniques are employed when the objective is to divide the instances into natural groups rather than predict their classes. The mechanism finds instances that bear strong resemblance and groups them together. However, it does not answer why the instances are similar in nature. It is a standard practice in data mining to divide the complete data into clusters and then investigate the clusters of interest with more micro level techniques such as decision trees to discover the phenomena of the domain. There are several types of clustering with each type having many algorithms to perform the activity. This research applies expectation maximization, also known as the EM algorithm to perform clustering. It is a probability based technique built on the idea of the popular k-means clustering method. In k-means clustering, k random points are selected within the dataset. Afterwards, the distance of each data point from each k points are calculated using the Euclidean method and the instances in the dataset are assigned to k clusters based on their proximity. Next, the centroids of the k clusters are calculated and those become the new k points. The process continues until two successive iterations yield the same result. K-means clustering method has some drawbacks, albeit being widely adopted. The success of the clustering vastly depends on how the initial k points are chosen. Besides, it is a time consuming process and it assigns each point to a cluster deterministically. These shortcomings can be overcome to some extent by adopting a more principled statistical approach. EM clustering is a probability based clustering method where the goal is to find the most likely set of clusters given the data. It suggests that as no finite amount of evidences can suffice the requirements to make an absolute decision on the matter. The instances (including the training instances) should not be categorized in one cluster or the other. However, they can be associated with a set of probabilities representing their probability of belonging to each cluster. The foundation of statistical clustering is based on a statistical model namely 'finite mixture' where the 'mixture' is a set of k probability distributions that represent k clusters governing the attribute values of members of that cluster. Hence, each cluster has a different distribution and although any particular instance belong to only one cluster, its probability to belong to different clusters is represented by a distribution (unlike discretely as in the case of K-means clustering). EM clustering is also such a method. Rather than randomly selecting k points, it assumes their mean, standard deviation and probability of the data being in that cluster. The term 'expectation' is associated with the expected value assumption and 'maximization' is related to the maximum likelihood of the distributions given the data. Another alteration of EM algorithm from k-means clustering is, each instance knows only its probability to be associated with a cluster, not the actual cluster where it belongs and this is performed by introducing the idea of 'weight'. To elaborate more, for a dataset of N instances, if w_i is the probability of instance i belonging to cluster A then its mean μ_A is $(w_1x_1+w_2x_2+\cdots+w_nx_n)/(w_1+w_2+\cdots+w_n)$ and the standard deviation σ_A is $(w_1(x_1-\mu_A)^2+w_2(x_2-\mu_A)^2+\cdots+w_n(x_n-\mu_A)^2)/(w_1+w_2+\cdots+w_n)$; where x_i is such that $i=1$ to N. The final major difference between the classical K-means clustering and EM clustering methods is in the terminating condition. Being probabilistic by nature, two successive runs in the EM algorithm are not likely to yield identical results. For this, the overall likelihood that the data comes from this dataset given the value of mean, standard deviation and the likelihood that the data belongs to the cluster under consideration are calculated. In theory, this can be obtained by multiplying the probabilities of the individual instances of i as shown in Eq. (2):

$$\prod(p_A Pr[x_i|A] + p_B Pr[x_i|B]) \tag{2}$$

where the normal distribution function $f(x;\mu,\sigma)$ is used to derive the probabilities given the cluster A and B. Here, p_A and p_B are the probabilities of any data belonging to cluster A or B respectively. In the same way, $Pr[x_i|A]$ and $Pr[x_i|B]$ stand for probability of a data point x_i conditional to cluster A or B respectively. In practice, a log-likelihood is calculated and higher values of likelihood imply higher goodness of the cluster. The algorithm of EM clustering guarantees to converge to a maximum. However, it is normally recommended to repeat the whole process several times with different initial guesses for the parameter values as in some cases EM algorithm may reach a local maximum rather than the global maximum. The study uses WEKA, an open source tool for machine learning to perform EM clustering and a detailed description of the algorithm and WEKA can be found in Witten and Frank (2005).

3.3. Classification and regression trees (CARTs)

Classification and regression tree (CART) is a method of generating decision trees developed by Breiman et al. (1984) that can be applied for knowledge discovery and classifying new data. In case of problem domains with large feature space, it may not be wise to opt for a global single predictive linear or polynomial regression model for the entire data space. On the contrary, CART is non-parametric by nature and partitions the data space into subdivisions in a recursive manner and brings it down to small manageable chunks containing data of only one dominant class. Its tree type structure is specially helpful to gain insight about the problem domain and facilitates identifying the most important predictors, too. The methodology has three major activities. First, it grows a decision tree of maximum depth in such way that each end node, often referred as leaf, contains data of a pure class. The second step prunes the tree to an appropriate size and obtains a sequence of nested sub-trees. Lastly, the best classification tree is chosen and the model is ready for classifying new data. The subsequent paragraphs explain the basics of the splitting and pruning rule used by CART. Although there are many algorithms available for the job, this study explains Gini splitting rule to split the nodes and cross validation to prune the trees as the software used in this study uses these methods (rpart package of R program).

Let the learning dataset have M number of predictors x_i, where $i=1$ to M. Let t_p be a parent node and t_l, t_r the left and the right child nodes after splitting. In CART, the splitting rule aims to separate the data into two chunks with maximum homogeneity. The algorithm ascertains the splitting value x_i^R in such way that for all splitting values of all the variables, x_i^R ensures maximum homogeneity of the child nodes. This is calculated by defining an impurity function $I(t)$. The idea accents that x_i^R will maximize the difference between the impurity of the parent node and the child nodes as presented in Eq. (3):

$$arg\max[\Delta I(t) = I(t_p) - P_l \times I(t_l) - P_r \times I(t_r)] \tag{3}$$

where P_l and P_r are the proportions of data in left and right nodes. Several algorithms are available for defining the impurity functions that can satisfy Eq. (3) to find the appropriate value of x_i^R. However, it has been ascertained that the final tree is insensitive to the algorithm selected. This study adopts Gini index based splitting algorithm for node splitting. If the outcome variable has K number of categories then the Gini index will vary between zero and $(1-1/K)$. The minimum value is observed when a node is pure, i.e., data of one class only and the maximum value is yielded when the outcome classes are equally distributed in the node. Gini index at any node t can be defined as:

$$I(t) = \sum_{j\neq l} p(j|t)p(l|t) = \sum_j p(j|t)(1-p(j|t)) = \sum_j p(j|t) - \sum_j p(j|t)^2 = 1 - \sum_j p(j|t)^2 \tag{4}$$

where j and l are the categories of the outcome variable and $p(j|t)$ is the proportion of outcome class j in node t. Now, the change in impurity can be calculated by plugging Eqs. (3.3) into (3.2). The change in impurity can be maximized by minimizing $[P_l \times I(t_l) - P_r \times I(t_r)]$. Using this splitting algorithm, tree is grown up to the maximum depth through recursive splitting until every node contains a pure class. Subsequently, the tree is pruned through a trade off between the complexity of the tree and the misclassification error. It is achieved by minimizing a compound function called cost-complexity (cp) function as shown in Eq. (5).

$$\min R_\alpha(T) = R(T) + \alpha(T') \tag{5}$$

where $R(T)$ is the misclassification error of tree T; T' is the total sum of terminal nodes in the tree T and $\alpha(T')$ is the complexity measure. The cross-validation method calculates the value of α by repeatedly taking a part of the data as learning sample to build the tree and using the other part to test the classification accuracy.

Apart from visualizing the problem domain in a graphical form, the final tree can be used to make inference for new data, too. Every

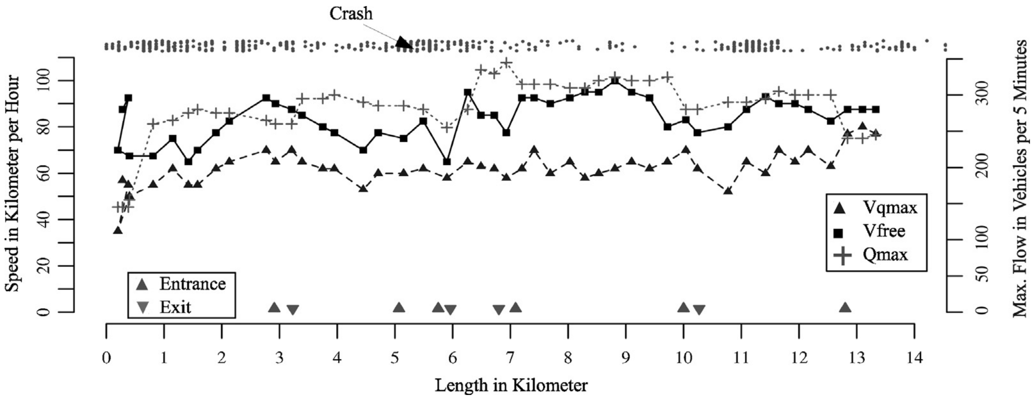

Fig. 4. Distribution of crashes, free flow speed (V_{free}), maximum flow (Q_{max}), speed at maximum flow (V_{qmax}) and ramp locations on Shinjuku 4 (inbound), Tokyo Metropolitan Expressway.

data point can be run down the tree using the splitting criteria and the class of the data will be the dominating class of the node where it ends up. This study uses rpart package of the R program (Dalgaard, 2008) to conduct the activities related to CART. Interested readers are requested to consult Soman et al. (2006) for further details.

4. Analysis and results

4.1. Preliminary data analysis

The dataset employed in this chapter contains 1141 crash cases of which 162, 39, 593 and 347 took place respectively in inbound and outbound of Shibuya 3 and Shinjuku 4 routes of Tokyo Metropolitan Expressway. Shibuya 3 has fewer numbers of crashes as the exposure for data are only for four months as compared to that of Shinjuku 4 which covers a duration of twelve months. Of these, 599, 283, 252 and 7 are respectively rear-end, sideswipe, hitting road furniture and tip over crashes. A further analysis shows that rear-end, sideswipe, hitting road furniture crashes account for 60%, 17% and 22% of the crash data in the BFS. The share varies for the ramp areas. It is 44%, 31% and 24% for the d/s of on ramps; 31%, 36% and 32% for the u/s of on ramps; 56%, 26% and 17% for the d/s of off ramps and 51%, 39% and 8% for the u/s of the off ramps. The outcome accentuates that the occurrence of sideswipe crashes are much higher in the ramp areas as compared to that of the BFS. Even within the ramp areas, rear-end and single vehicle crashes hitting the road furniture are less frequent in on ramps than that in the off ramps. This implies that crash mechanism may not be generic for throughout the length of the expressway and may differ in different types of ramp areas and BFS. Fig. 4 demonstrates the spatial variation of crashes, free flow speed (V_{free}), maximum flow (Q_{max}) and speed at maximum flow (V_{qmax}) along with the location of different types of ramps for the inbound traffic of Shinjuku 4 route. It has been prepared by selecting the required data from three random weeks of three different months. It can be observed that the crashes concentrate near the ramp areas and are less frequently scattered for long BFS.

4.2. Basic freeway segments (BFS)

The dataset prepared for the BFS contains information on 65 predictors, 183 crash cases and their corresponding 6478 non-crash situations. The problem under consideration involves large sample space and low instances of a certain class of the outcome variable. Hence, RMNL has been used to identify the most important variables (Section 3.1). A total of 500 trees of logit models are grown by randomly choosing 4 variables at a time. Selecting only 4 variables for each tree reduces the probability of picking up two highly correlated variables at a time for a tree. The results yield how many times a variable is picked up randomly and its average raw importance calculated based on Section 3.1. Outcome shows that each variable got picked up at least 20 times and maximum 41 times during the process. The change in raw variable importances is found to be insignificant after growing 500 trees and for this the tree growing process has been terminated at that point. The top 10 most important variables with their relative importance are presented in Fig. 5. Interestingly, none of the top ten most important variables are associated with traffic flow and most of the variables are yielded by detectors 2, 4 or the difference in variables between these two locations. Likewise, the level of congestion, represented by the congestion index, has more importance in explaining traffic conditions influencing crash than the speed, as hypothesized in the study. This holds high significance from the intervention designing perspective as any certain speed may not be classified as high or low for throughout the road sections.

This study is concerned about how a crash prone traffic condition is formed. Therefore, it is important to inspect the spatial variation of the variables. To ensure that, the differences in congestion index, speed and occupancy between detectors 2 and 4 are subjected to further investigation through clustering. A negative value of the variable suggests that the corresponding value in the upstream is higher than that in the downstream. Now, the sample size increases as a total of 311 crash cases and their corresponding 9158 non-crash situations have complete information on variables yielded by detectors 2 and 4. EM clustering algorithm has been applied by sequentially generating 2, 3, 4 and 5 clusters out of the dataset. The log-likelihood values obtained respectively are −4.873, −4.615, −4.529 and −4.441; exhibiting gradual improvements in results. The outcomes for the analysis with 5 clusters are presented in Table 3. It is important to mention here that the summation of number of crashes is slightly higher than 311 as the algorithm

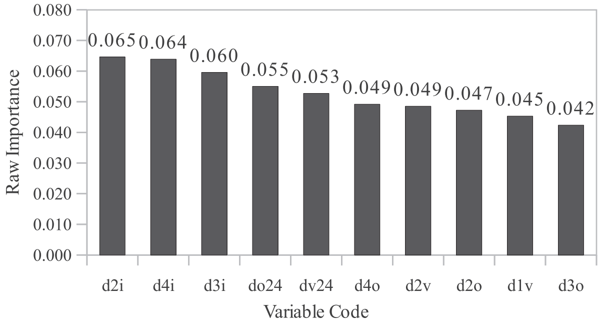

Fig. 5. Most important variables identified by RMNL for the BFS.

Table 3
Clustering analysis for the basic freeway segments.

Mean	Cluster no.				
	0	1	2	3	4
Speed difference	−2.820	32.960	9.130	−18.358	9.457
Occupancy difference	1.060	−19.194	−1.748	9.477	−3.843
CI difference	0.015	−0.420	−0.004	0.246	−0.092
Crash	82	91	36	51	54
Non-crash	4965	79	2813	604	699
Risk ratio	0.498	**16.254**	0.389	**2.369**	**2.209**

considers the expected values of the means and standard deviation rather than a discrete value from the instances of the dataset.

It can be ascertained from Table 3 that the Clusters 1 and 3 have the highest risk of crash. The average risk (no. of crash cases/no. of non-crash situations) in the dataset is 311/(311 + 9158) = 0.0328 and the risk of Cluster 1 (no. of crash cases in the cluster/no. of non-crash cases in the cluster) is 91/(91 + 79) = 0.5339. The risk ratio for Cluster 1 (risk of the cluster/average risk) is fathomed by the ratio of these two values, i.e., 16.254 in this example. The situation manifests when the average speed in the downstream is around 33 km/h higher than the upstream section along with a pact traffic condition in the upstream. Clusters 2 and 4 have similar speed difference, however, the later has much higher occupancy in the upstream than the former. Like the previous studies, this study also accentuates that shock waves created by a congested downstream and a fast approaching upstream can impend to potentially danger (Cluster 3). However, this study underscores that a substantially fast moving downstream trailed by a slow and congested upstream may pose a more conspicuous threat of crash. Next, it will be interesting to research which types of traffic conditions are accompanied with which types of crashes. For this, classification and regression trees have been grown with the dataset containing the crash data of Clusters 1, 3 and 4. Only data of traffic conditions leading to crash are considered as the objective is to investigate if certain crash types have high affiliation with specific traffic conditions. The new dataset contains 189 data points of which 66.7% are rear-end collisions, 17.9% are sideswipe collisions and 15.4% are hitting the road furniture – suggesting rear-end crashes to be the most predominant type. Instead of growing the tree up to maximum depth where each terminal node containing a pure class, a minimum split rule has been used in such way that a node gets split only when it contains at least 20 data points and subsequent minimum child node has at least 1/3rd of those data points. This facilitates in reducing the calculation complexity as well as the tree size substantially. The resulting tree contains 9 nodes as illustrated by Fig. 6. The node numbers are encircled next to the nodes. If the parent node is ‘*n*’ then the left and the right child nodes are numbered as ‘2*n*’ and ‘2*n* + 1’ respectively.

The occupancy difference emerges as the main splitting variable of the root node. It is ascertained that when the occupancy in the upstream is higher than the downstream by 7.45% or more, most of the resulting crashes are rear-end collisions. A further investigation on these 101 crash samples in Node 2 implies that most of those ensued when both the upstream and the downstream were congested, however, the downstream just started clearing. The crashes may have resulted in as the upstream drivers were adjusting to a higher speed and different drivers had different perception about the desired speed demanding drivers to exhibit multiple cognitive abilities to avoid collision. For the rest of the data, if the occupancy difference is between 2.2% and −7.45% and the speed difference is less than −1.35 kph then majority of the crashes are either sideswipe or hitting the road furniture (Node 27). However, for the same speed condition, if the occupancy difference is higher than 2.2%, i.e., suggesting a more congested downstream then rear-end crashes are more likely than the sideswipe crashes (Node 26). This implies that when drivers from high occupancy locations move toward a much lower occupancy location, they mainly focus on adjusting the speed and thus are more susceptible to rear-end crashes. However, when the difference is lower, they look for maneuvering options to progress forward. Some of the possible countermeasures that can be opted for are discussed in the concluding section (Section 5). Every terminal node except for Node 27 contains more rear-end crash cases, suggesting this crash type as the most common and predictable among all the types for the basic

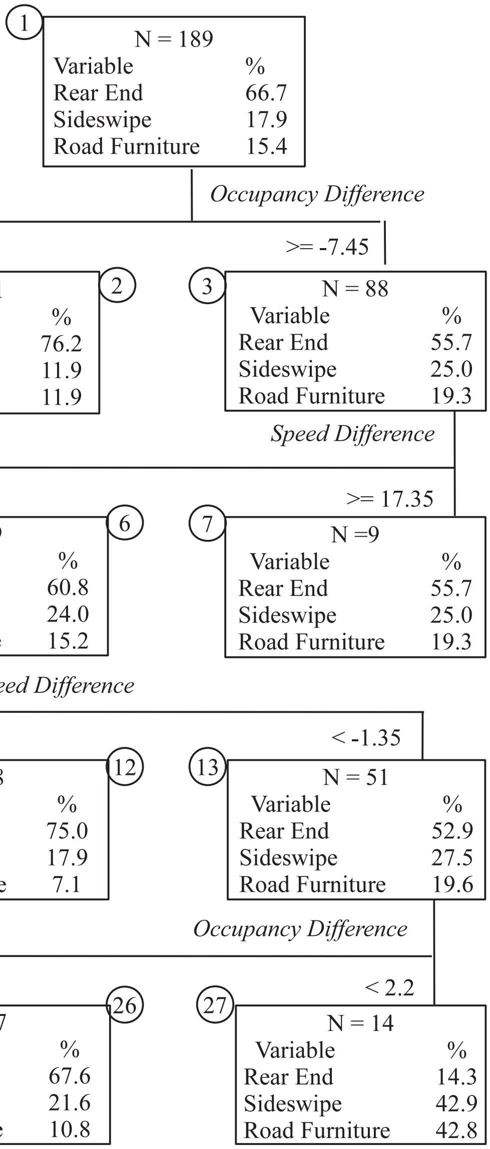

Fig. 6. Classification tree of crash data in Clusters 1 and 3 for the BFS.

Table 4
Clustering analysis for the ramp areas.

Mean	Cluster no.				
	0	1	2	3	4
d/s-on ramp					
Ramp flow	14.483	27.343	26.367	34.694	19.396
Flow difference	13.890	24.675	30.201	34.793	17.210
Speed difference	−1.595	12.592	6.979	−5.402	−16.541
Occupancy difference	0.714	−0.034	−3.132	2.090	11.548
CI difference	0.043	0.025	0.114	0.042	0.175
Crash	34	8	43	14	26
Non-crash	2011	597	419	796	329
Risk ratio	0.591	0.479	**3.315**	0.618	**2.575**
d/s-off ramp					
Ramp flow	14.593	9.164	16.345	13.464	10.370
Flow difference	−14.801	−10.895	−16.296	−15.567	8.933
Speed difference	8.894	−19.493	−6.309	−17.263	1.136
Occupancy difference	−1.946	1.120	4.203	7.257	−0.495
CI difference	−0.006	−0.084	0.080	0.195	0.007
Crash	10	15	16	30	18
Non-crash	861	876	368	331	1266
Risk ratio	0.530	0.785	**1.851**	**3.654**	0.631
u/s-on ramp					
Ramp flow	38.036	4.190	21.379	49.877	20.871
Flow difference	37.781	3.270	20.328	53.705	21.023
Speed difference	−6.352	−4.199	−10.544	11.116	4.325
Occupancy difference	2.303	0.755	6.453	−9.392	0.304
CI difference	0.043	0.005	0.127	−0.140	−0.007
Crash	9	13	44	17	30
Non-crash	633	604	282	265	1847
Risk ratio	0.518	0.737	**4.570**	**2.073**	0.555
u/s-off ramp					
Ramp flow	12.556	12.759	6.029	12.284	11.206
Flow difference	−11.942	−13.398	−3.778	−19.101	−9.140
Speed difference	14.397	−9.919	8.915	−23.644	3.317
Occupancy difference	−4.680	3.402	−1.638	11.718	−0.796
CI difference	−0.127	0.032	−0.075	0.297	0.012
Crash	25	23	14	19	22
Non-crash	509	749	762	265	1783
Risk ratio	**1.964**	**1.262**	0.795	**2.798**	0.520

freeway segments. Unlike the previous studies, this study ascertains that detecting a crash when the downstream is faster than the upstream is more conspicuous than the opposite scenario. The relatively lower significance of flow as a variable underscores the importance of using high-resolution traffic flow data for detecting and effective countervailing crash prone situations as compared to classical approaches depending on daily, monthly or yearly flow data and fixed speed limits of the roads.

4.3. Ramp areas

The dataset for the ramp area accumulated for this study consists of 414 crash samples with complete information on all 16 chosen variables. Of these, 120, 110, 86 and 98 took place in downstream and u/s of on and off ramps respectively. They have corresponding 4148, 3628, 3699 and 4066 non-crash traffic condition data. Following the steps of clustering analogous to the BFS, the data are separated into 2–5 clusters considering the difference in flow, number of heavy vehicles, speed, occupancy and congestion index between detectors 1 and 3 and the flow of detector 2 (Fig. 2). The results for the model with five clusters along with their associated risk ratios are presented in Table 4. It can be observed that Clusters 2 and 4 have substantially higher risk ratios as compared to other clusters in case of the d/s of on ramps. They share among them 55% of the crash samples. For the d/s of off ramp and the u/s of on ramp, Clusters 2 and 3 are found to be risky and they contain 38% and 53% of the crash samples. In case of the u/s of the off ramps, Clusters 0, 1 and 3 are markedly affiliated with greater risk and they include 65% of the crashes taking place in that region. Apart from the u/s of on ramps, the flow on the ramp does not imply to have much impact in differentiating risky and safe clusters. This may be because detectors 1 and 3 are quite closely spaced and the flow in the ramp in most cases is the difference in flow between these two detectors. Analogous to the BFS, the ramp areas also have two distinct situations when the traffic condition becomes hazardous: (i) a condition when the downstream is faster and upstream vehicles are slow moving and much more closely spaced than the downstream vehicles and (ii) a heavily congested downstream followed by a speedy upstream flow; except for the d/s of the off ramp where only the first condition is valid. Of these two situations, the former seem to pose higher threat for the d/s of on ramps (similar to the situation in the BFS) whereas the later is more prevailing for other situation. In all the cases, the occupancy and congestion index differences are substantially higher than the safer clusters, albeit their values differ substantially among the groups of the ramp areas. The results vividly ascertain that a common model to assess real-time safety or a set of common solutions to prevent crashes in real-time may not be adopted for each of the four groups of ramp areas considered in this study.

Next, classification trees have been grown using CART method for the four ramp conditions involving the crash data in the high risk clusters (marked with bold font face in Table 3). The results are shown in Fig. 7. The number of crash samples and their share based on types can be found in the root node of the corresponding groups from Fig. 7. Only rear-end crashes and sideswipe crashes are found in the risky clusters. A further insight into the data reveals

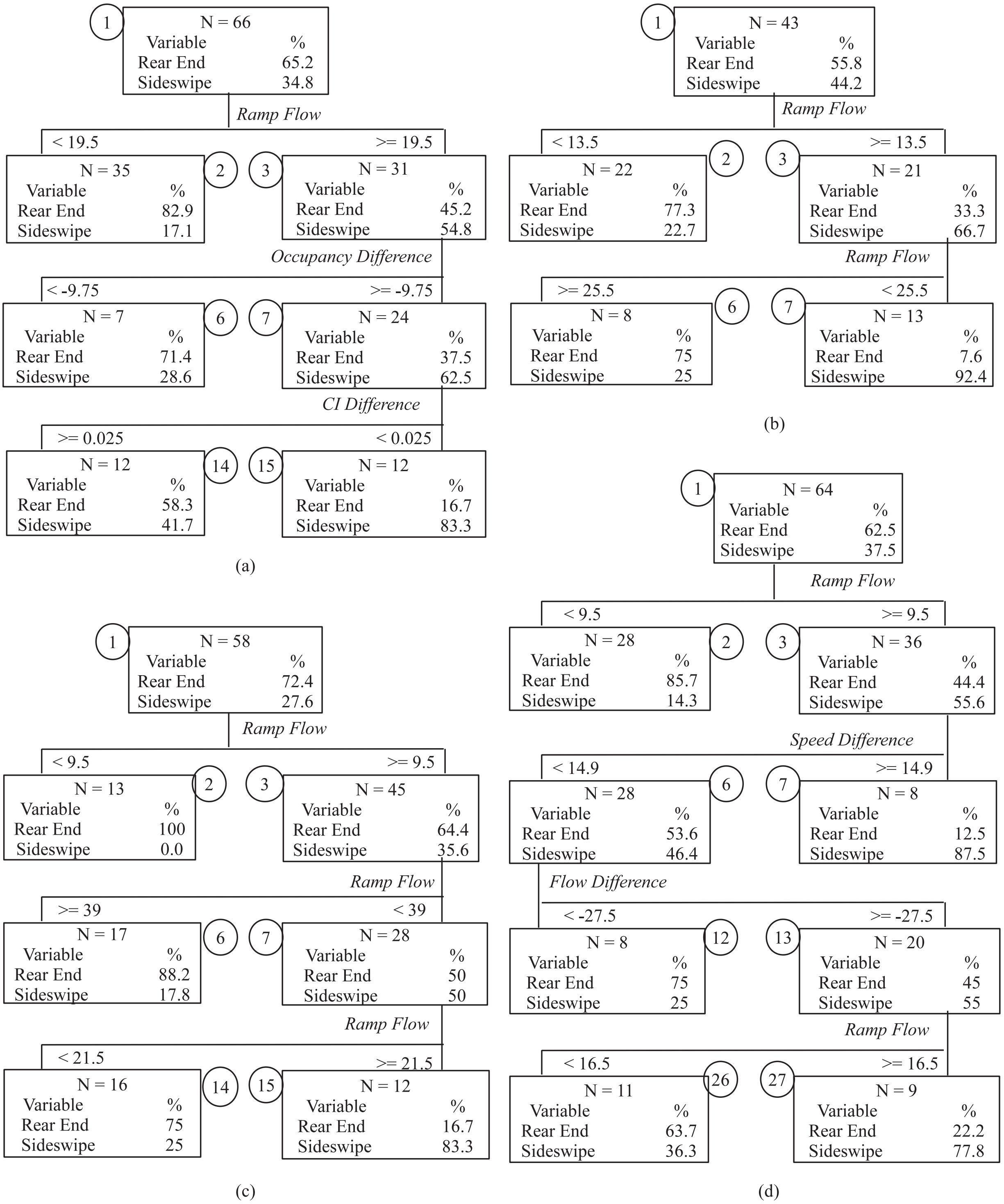

Fig. 7. Classification trees for crash cases within the high risk clusters of different locations within the ramp vicinity (a – d/s on, b – d/s off, c – u/s on and d – u/s off).

that crashes hitting the road furniture near ramps are associated with high speed. This may be the reason for the risky clusters not containing any crash sample of this type. Interestingly, although the flow in the ramp was not found to be the strongest variable to distinguish risky clusters from the relatively safer clusters, it is the most important splitting variable in determining the type of crash. Around 83% of the crashes in the d/s of on ramps are found to be rear-end when the ramp flow is less than 20 vehicles/5 min. Sideswipe crashes seem to be more or equally dominating for the same location when the ramp flow is higher than or equal to 20 vehicles/5 min and the upstream is more congested than the downstream. The crashes in the d/s of the off ramps are almost evenly divided between rear-end and sideswipe and can be independently classified by the ramp flows. If the flow is more than 25 vehicles or less than 14 vehicles/5 min then the predominant crash type is rear-end collision. Otherwise, sideswipe crashes occur more often. Rear-end crashes dominate for all the ramp flow conditions in case of the u/s of an on ramp except for situations when the ramp flow is within the range of 22–38 vehicles/5 min. This specific flow condition may create a dilemma for the drivers on a two lane urban expressway about whether to change the lane or not to avoid the merging traffic ahead. Most of the crashes in the u/s of the off ramp are rear-end during a low ramp flow condition (less than 10 vehicles/5 min). The sideswipe crashes occur with high frequency when the ramp flow is higher than 16 vehicles/5 min. This is understandable as many vehicles are changing lanes to either exit from the expressway or avoid blocking the vehicles approaching the ramp.

5. Discussion and conclusion

The study attempts to identify significant predictors and traffic patterns for crashes on urban expressways utilizing high-resolution traffic sensor data. From the ascertained knowledge, it provides insight into the underlying crash phenomena. Unlike the previous studies on conventional expressways, this research separates different sections of urban expressways into five groups – the basic freeway segments (BFS), d/s and u/s of on and off ramps. Subsequently, it formulates a hypothesis that factors defining the crash risk, patterns and phenomena are not generic and may vary among these grouped sections. To verify this, the research employs data from Shibuya 3 and Shinjuku 4, two of the busiest routes of the Tokyo Metropolitan Expressways. The locations are ideal for such a study as they are densely packed with detectors with an average inter-detector spacing as low as 250 m. Rather than solely relying on available literatures and engineering judgments, the research introduces random multinomial logit to rank the most important variables. If not the first then it is one of the first implementations of RMNL in transportation research. The study applies expectation maximization clustering, a new probability based algorithm to identify crash patterns associated with high risk. Lastly, it unveils the association of different crash types with different traffic conditions for different locations on the urban expressway by constructing classification and regression trees.

The preliminary exploratory analysis suggests that crash density is higher near the ramp vicinities than the long BFS. Among the crash types, rear-end crashes are most frequent followed by the sideswipe collisions and the collisions with road furniture. They also vary based on the location of the expressway. For example, 60% of the crashes on the BFS are rear-end collisions, however, the share can be as low as 31% for the u/s of on ramps. Variables related to congestion index, occupancy and speed stand out as the most important variables for the BFS. The relatively lower significance of flow as a variable underscores the importance of using high-resolution traffic flow data for detecting and effective countervailing crash prone situations as compared to classical approaches based on daily, monthly or yearly flow data. The study also recommends that crash mechanism on the BFS of a two lane urban expressway can be best observed through two detectors roughly placed 250 m upstream and downstream from the crash location. This finding holds high significance for building real-time crash prediction models. Regarding the clustering outcome for the BFS, the study identifies two specific patterns – (i) a high speed upstream traffic meeting a slow moving downstream and (ii) a rapidly progressing downstream trailed by a congested upstream. The former condition was also identified in several previous studies conducted on conventional expressways. However, the later condition has not been widely discussed in previous literatures. Moreover, it seems to be riskier and more conspicuous than the former condition. A further investigation also reveals that the later condition manifests when both the upstream and downstream traffic are congested but suddenly the downstream congestion starts alleviating. The suddenly availed possibility to increase speed to match the downstream may require drivers to engage multiple cognitive abilities and also different drivers may have different perceptions about the desired speed which may form conditions influencing crash. In order to classify crashes based on their types CART analysis has been employed and the occupancy difference emanated as the most significant variable, albeit speed and congestion index differences also came out significant. Rear-end crashes dominate in the BFS, however, sideswipe crashes are prominent when the downstream is faster than the upstream but have similar level of occupancy. This may be because when the congested condition starts dissipating, different drivers adopt different level of acceptable acceleration rate. As they are moving from slow speed to higher speed, the fear of having a severe crash may be lower in their mind. Also, they may show signs of compensating the loss in travel time by shifting to a faster lane. Severity data were not available for this study. However, it may happen that crashes of this nature are associated with lower level of severity or property damage only. However, on a two lane expressway, when there is even a slight crash, the cars must stop and wait for the intervention of law enforcement organizations to resolve the issue. Therefore, it is important to identify these crash conditions even if they are not of high severity and come up with counter measures to prevent those. Also, it will be interesting to have an experiment conducted with driving simulator to observe how the behavior of drivers varies in accelerating from a congested traffic condition. It may also happen that the same driver exhibits different behavior depending on the time he had to wait due to congestion. For devising countermeasures, the impact between a fast moving upstream and a congested downstream can be pacified by reducing the upstream speed through variable message signs (VMS) coupled with a warning sign. If the situation evolves when a congested downstream suddenly starts alleviating creating substantial difference in speed and occupancy between the upstream and the downstream then speed limits can be imposed on the downstream to prevent the rear-end crashes. However, sideswipe crashes take place in a situation where the downstream is fast but the occupancy difference is low. To surmount this, restriction in lane changing can be applied in addition to posting warning messages through VMS. As having a lane change restriction is not yet a common practice, authorities may think of introducing new signs or directly posting the instructions through VMS. For the ramp vicinities, the difference in occupancy and level of congestion play more important role than the speed difference to distinguish between safe clusters and the high risk clusters. Analogous to the BFS, level of congestion difference between the downstream and upstream seems to give rise to hazardous condition for all the ramp areas except for the d/s of the off ramp. In that case, only congested downstream and fast flowing upstream affirms a hazardous condition. Regarding the crash types, the d/s of the off ramps and the u/s of on ramps could solely be classified with the ramp flow. However, the d/s of on ramps needed more information on the occupancy and congestion index difference for classification. The information regarding speed and flow difference is useful for the crashes taking place in the u/s of the off ramps. Several crash patterns affiliated to specific crash types have also been specifically identified. For example, rear-end crashes are the prominent types of collisions for the d/s of the off ramp during high flow (more than 25 vehicles/5 min; Node 6 in Fig. 7b) and low flow (less than 14 vehicles/5 min; Node 2 in Fig. 7b) conditions. However, the share of the sideswipe crashes is substantial for a moderate ramp flow. This may be because the extreme conditions make it easy for the drivers to decide not to maneuver whereas a moderate flow puts the drivers into dilemma. The outcomes imply that ramp metering (Lee et al., 2005; Abdel-Aty et al., 2007) assisted with speed controlling measures can be effective to reduce crashes near ramp vicinities. A lane change restriction may also be imposed in the downstream where a specific volume of flow on the ramp poses substantial threat of sideswipe collisions. However, it should be carefully examined if there is any off ramp nearby as some vehicles may need to maneuver to access it.

As an overall note it is important to mention that the high risk clusters in all five groups of the road sections have substantially high differences in their congestion indexes. This indicates that either the downstream or the upstream traffic conditions were at least partially congested. Thus, it is easier to explain the crash mechanism under low speed operation. This is also logical to believe that many high speed crashes may be associated with unsafe driving rather than traffic condition which is hazardous and thus hard to explain with traffic flow variables. They will also require education and enforcement related interventions rather than engineering solutions.

Regarding the limitations, the research uses two very busy Japanese expressways as the study area. It is hard to envisage to what extent the results can be transferable to other parts of the world. It will be interesting to see the outcome of a study conducted on an urban expressway in another country following the same methodology. Also, we fund 30 crash samples falling into Cluster 4 for the u/s of the on ramps (Table 4) having a risk ratio approximately half of the average risk. It can also happen that these crashes have a mechanism which is closer to that of the BFS rather than the ramp areas and the crashes may not have been influenced by the existence of the ramp. We recommend building of separate crash prediction models for the proposed five locations and evaluate each crash case, specially those in the u/s of the on ramps, to further confirm their crash mechanism. Moreover, although the overall sample size seemed large at the beginning, it did not remain that large after they were further separated for the BFS and ramp vicinities. Specially, in case of the d/s of the off ramps, only 43 crash samples were used to perform CART analysis. Hence, further study is recommended for this specific ramp vicinity before using the outcome of the analysis for designing countermeasures. In this study, geometric variables were not introduced. Environmental factors were not considered either due to lack of availability of data. It was assumed that the speed-flow-occupancy related variables can be considered as surrogate measures of weather as the impact of weather is translated into these variables to a great extent. Likewise, this research treats all the crashes equally as the associated severity data were not provided. Some crash patterns, such as, predominance of rear-end crashes for the d/s of the off ramp during high flow (more than 25 vehicles/5 min; Node 6 in Fig. 7b) and low flow (less than 14 vehicles/5 min; Node 2 in Fig. 7b) conditions require further investigation, probably with the assistance of a driving simulator, to shed greater light into the crash mechanism. Finally, the manuscript confines itself in identifying different crash patterns. Rather than introducing new countermeasures, it makes intelligent use of the existing ones and associates those with the appropriate crash patterns. Further study is recommended to actually investigate the proposed countermeasures.

In conclusion, it can be expressed that the study is one of the first in investigating crash phenomena using high-resolution traffic data on urban expressways outside North America and Europe. It has also been conducted involving cutting edge methods such as RMNL, CART and EM clustering. The paper demonstrates that crash phenomena and its types vary between the BFS and the ramp areas and even within the ramp areas based on their types and locations. This ascertains the needs to introduce separate approaches to predict crash prone situations as well as to prevent crashes on BFS and ramp vicinities. The findings are expected to be valuable to the expressway authorities in designing location and situation based proactive road safety management systems.

Acknowledgement

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

References

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

Abdel-Aty, M., Abdalla, F., 2004. Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes using generalized extreme equations for correlated data. Transportation Research Record 1897, 106–115.

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 1908, 51–58.

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

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

Abdel-Aty, M., Dhindsa, A., Gayah, V., 2007. Considering various ALINEA ramp metering strategies for crash risk mitigation on freeways under congested regime. Journal Transportation Research Part C 15 (2), 113–134.

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

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

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

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

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accident Analysis and Prevention 39 (4), 657–670.

Cedar, A., Livneh, M., 1982. Relationship between road accidents and hourly traffic flow – I. Accident Analysis and Prevention 14 (1), 19–34.

Cedar, A., 1982. Relationship between road accidents and hourly traffic flow – II. Accident Analysis and Prevention 14 (1), 35–44.

Chen, H., Liu, P., Lu, J.J., Behzadi, B., 2009. Evaluating the safety impacts of the number and arrangement of lanes on freeway exit ramps. Accident Analysis and Prevention 41 (3), 543–551.

Chen, H., Zhou, H., Zhao, J., Hsu, P., 2010. Safety performance evaluation of left-side off-ramps at freeway diverge areas. Accident Analysis and Prevention 43 (3), 605–612.

Christoforou, Z., Cohen, S., Karlaftis, M.G., 2011. Identifying crash type propensity using real-time traffic data on freeways. Journal of Safety Research 42, 43–50.

Christoforou, Z., Cohen, S., Karlaftis, M.G., 2012. Integrating real-time traffic data in road safety analysis. Procedia – Social and Behavioral Science 48, 2454–2463.

Dalgaard, P., 2008. Introductory Statistics with R. Springer, NY.

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

Frantzeskakis, J.M., Iordanis, D.I., 1987. Volume-to-capacity ratio and traffic accidents on interurban four-lane highways in Greece. Transportation Research Record 1112, 29–38.

Fridstrom, L., Ifver, J., Ingebrigtsen, S., Kulmala, S.R., Thomsen, L.K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accident Analysis and Prevention 27 (1), 1–20.

Golob, T.F., Recker, W.W., Alvarez, V.M., 2003. A tool to evaluate the safety effects of changes in freeway traffic flow. In: Proceedings of the 82nd Annual Meeting of Transportation Research Board, Washington, DC.

Harb, R., Yan, X., Radwan, E., Su, X., 2009. Exploring precrash maneuvers using classification trees and random forests. Accident Analysis and Prevention 41 (1), 98–107.

Hossain, M., Muromachi, Y., 2010a. Evaluating location of placement and spacing of detectors for real-time crash prediction on urban expressways. In: Proceedings of the 89th Annual Meeting of Transportation Research Board, Washington, DC.

Hossain, M., Muromachi, Y., 2010b. Development of a real-time crash prediction model for urban expressway. Journal of EASTS 8, 2091–2107.

Jang, J.A., Choi, K., Cho, H., 2012. A fixed sensor-based intersection collision warning system in vulnerable line-of-sight and/or traffic-violation-prone environment. Intelligent Transportation System 13 (4), 1880–1890.

Lee, C., Hellinga, B., Ozbay, K., 2005. Quantifying effects of ramp metering on freeway safety. Accident Analysis and Prevention 38 (2), 279–288.

McCartt, A.T., Northrup, V.S., Retting, R.A., 2002. Types and characteristics of ramp-related motor vehicle crashes on urban interstate roadways in Northern Virginia. Journal of Safety Research 35 (1), 107–114.

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

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

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

Khan, S., Shanmugam, R., Hoeschen, B., 1999. Injury, fatal, and property damage accident models for highway corridors. Transportation Research Record 1665, 84–92.

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

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

Lee, C., Abdel-Aty, M., Hsia, L., 2006. Potential real-time indicators of sideswipe crashes on freeways. Transportation Research Record 1953, 41–49.

Miaou, S.-P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis and Prevention 37 (4), 699–720.

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

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 1953, 31–40.

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

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

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

Soman, K.P., Diwakar, S., Ajay, V., 2006. Insight into Data Mining: Theory and Practice. Prentice-Hall, India.

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

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, USA.

Xu, C., Wang, W., Liu, P., 2012. A genetic programming model for real-time crash prediction on freeways. IEEE Transactions on Intelligent Transportation Systems PP (99), 1–13.

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