Transportation Research Part C 26 (2013) 203–213

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

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

A data fusion framework for real-time risk assessment on freeways

Mohamed Ahmed *, Mohamed Abdel-Aty

Department of Civil, Environmental and Construction Engineering, University of Central Florida, Engineering II – 325 Orlando, FL 32816, United States

ARTICLE INFO

Article history:
Received 20 February 2012
Received in revised form 25 June 2012
Accepted 5 September 2012

Keywords:
Freeway safety
Automatic Vehicle Identification
Remote Traffic Microwave Sensor
Data mining
Data fusion
Stochastic Gradient Boosting

ABSTRACT

The increased deployment of non-intrusive detection systems such as Automatic Vehicle Identification (AVI) and Remote Traffic Microwave Sensors (RTMSs) provides access to real-time traffic data from multiple sources. The availability of such rich data enhances the reliability of travel time estimation and route guidance systems, however, utilization of these data is absent in the context of proactive safety management systems. This paper presents a framework for real-time risk assessment on a freeway in Colorado by fusing data from two different detection systems (AVI and RTMS), real-time weather and roadway geometry. Stochastic Gradient Boosting (SGB), a relatively recent and promising machine learning technique is used to calibrate the models. SGB's key strengths lie in its capability to fit complex nonlinear relationships, handling different types of predictors (nominal and categorical) and accommodating missing values with no need for prior transformation of the predictor variables or elimination of outliers. Boosting multiple simple trees together overcomes the drawback of single tree models of poor prediction accuracy and provides fast and superior predictive performance. The proposed framework is considered a good alternative for real-time risk assessment on freeways because of its high estimation accuracy, robustness and reliability.

1. Introduction

Accurate and reliable estimation of increased risk of crashes is critical to the success of proactive safety management strategies on freeways. In recent years, the advances in electronics have had a tremendous impact on enhancing and improving detection systems, new non-intrusive traffic detection devices are in use more these days because of their easiness of installation and maintenance in addition to their accuracy and affordable cost. Moreover, some freeways have multiple non-intrusive detection systems in place such as the Automatic Vehicle Identification (AVI) and Remote Traffic Microwave Sensor (RTMS). AVI is used mainly for toll collection and for travel time estimation purposes along freeways while RTMS are used mostly for operation and incident management. Research in the field of freeway traffic management has utilized extensively traffic data collected from inductive loop detectors in real-time crash prediction (Oh et al., 2001; Golob and Recker, 2001; Lee et al., 2002, 2003; Abdel-Aty et al., 2004; Abdel-Aty and Pande, 2005; Pande and Abdel-Aty, 2006a,b; Hourdos et al., 2006; Hossain and Muromachi, 2012). Recently, the usefulness of the collected traffic data from AVI has been investigated in real-time safety assessment (Ahmed and Abdel-Aty, 2011; Ahmed et al., 2011, in press-a,b).

Traffic data from AVIs and RTMSs as well as weather data are collected on 15-mile of mountainous Interstate-70 in Colorado to provide roadway users with important information about travel time, congestion, adverse weather conditions and lane closure due to occasional avalanche danger, maintenance on the road and/or road crashes. This information is provided as a part of an Intelligent Transportation System (ITS) and is dynamically disseminated in real time to road users via Dynamic

* Corresponding author. Tel.: +1 407 823 4552; fax: +1 407 823 3315.
E-mail address: mahmed@knights.ucf.edu (M. Ahmed).

0968-090X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2012.09.002

Message Signs (DMSs). This system utilizes AVI to estimate the segment travel time by monitoring the successive passage times of vehicles equipped with electronic tags at designated locations. Main traffic flow parameters are collected using RTMS. It is worth mentioning that the AVIs and RTMSs are providing different measures of speeds; AVIs measure space-mean-speed (SMS), which is defined by Gerlough and Huber, 1975 as "the mean of the speeds of the vehicles traveling over a given length of road and weighted according to the time spent traveling that length", whereas RTMSs measure time-mean-speed (TMS) which is the arithmetic mean of the speed of vehicles passing a point during a given time interval. Hence, TMS only reflects the traffic condition at one specific point. On the other hand, SMS is the average speed of all the vehicles occupying a given stretch of the road over some specified time period (there are several definitions of SMS depending on how it is calculated (Hall, 1996); the definition by Gerlough and Huber, 1975 is the best to describe the AVI's SMS).

Weather condition is considered one of the most important factors that can contribute to crash occurrences. In previous studies weather data are always estimated from crash reports, in this study real-time weather data are gathered by weather stations located on the roadway section.

Although in previous research efforts by the authors, it was found that classical statistical models provide interpretable models and acceptable accuracy of crash prediction using AVI and real-time weather data (Ahmed et al. 2011, in press-a), in this study we propose a framework to augment even more traffic data from multiple sources, weather and geometry data using an advanced machine learning (ML) technique. Machine learning methods are known for their superior classification and prediction performance over the classical statistical ones. In order to enhance the accuracy and increase the reliability of the real-time crash prediction, Stochastic Gradient Boosting (SGB), a recent and promising machine learning technique is attempted to uncover previously hidden patterns preceding a crash relative to non-crash conditions from the large amounts of roadway geometry, weather and AVI and RTMS traffic data.

The following sections illustrate the procedures of preparing the data, modeling technique, interpretation and evaluation, risk assessment framework and the conclusions.

2. Data description and preparation

There were five sets of data used in this study; roadway geometry data, crash data, and the corresponding AVI, RTMS and weather data. The crash data were obtained from CDOT for a 15-mile segment on I-70 for 13 months (from October 2010 to October 2011). Traffic data consists of space mean speed captured by 12 and 15 AVI detectors located on each east and west bounds, respectively along I-70. Volume, occupancy and time mean speed are collected by 15 RTMSs on each direction. AVI estimates SMS every 2-min while RTMS provides traffic flow parameters every 30-s. Weather data were recorded by three automated weather stations along the roadway section for the same time period. The roadway data were extracted from Roadway Characteristics Inventory (RCI) and Single Line Diagrams (SLDs).

In a previous study by the authors (Ahmed and Abdel-Aty, 2011), it was found that crash occurrence was mostly related to the AVI crash segment, one segment in the upstream and another segment in the downstream directions and therefore these AVI segments and their respective RTMS stations were considered in the data extraction process and modeling parts. The crashes have been assigned to the AVI segment and to the closest RTMS station; upstream and downstream AVI segments as well as three RTMSs in the upstream and downstream were identified to extract their corresponding traffic data. The upstream, crash, and downstream segments were named U, C and D, respectively while the upstream and downstream RTMSs were named US and DS respectively and assigned numbers in order from the closest to the farthest ones. It is worth mentioning also that most of the RTMSs are located exactly at the same location of the AVIs' tag readers. The arrangement of RTMS and AVI segments and their spacing are illustrated in Fig. 1.

AVI and RTMS data corresponding to each crash case were extracted in the following process; the location and time of occurrence for each of the 186 crashes were identified. Traffic data were aggregated to 6-min level to obtain averages, standard deviations, and logarithm of coefficient of variations (standard deviation divided by the average of the traffic

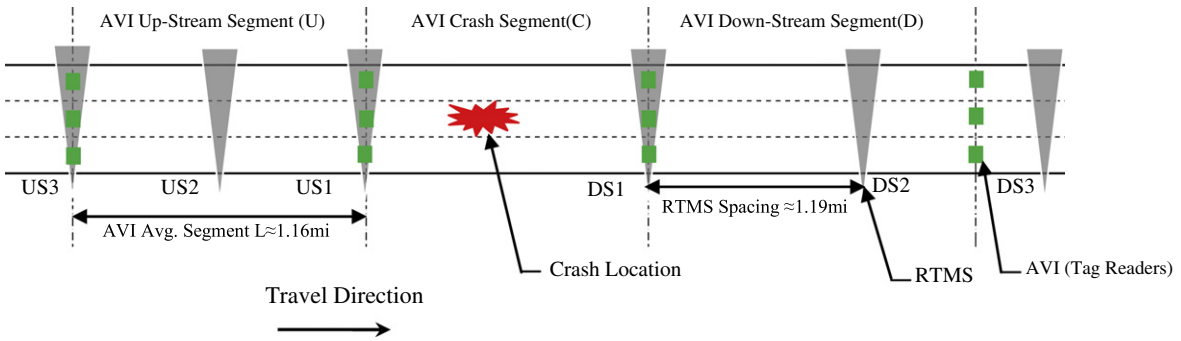

Fig. 1. Arrangement of RTMS and AVI segments.

parameters) of 2-min space mean speed obtained from AVIs and 30-s time mean speed, volume, and occupancy raw data obtained from RTMSs. The 6-min aggregation level was chosen to have consistent time periods between AVIs and RTMSs.

Three time slices of the 6-min prior to the crash time were extracted. For example if a crash happened on September 16, 2010 (Sunday) at 14:00, at the milepost of 210.1 EB. The corresponding 18-min window for this crash of time intervals (13:42–14:00) recorded by AVI segment 6 (mile marker starts at 209.79 and ends at 210.60), upstream AVI segment 5 and downstream AVI segment 7 as well as three RTMSs in the upstream and three in the downstream were extracted. Time slice 1 was discarded in the analysis since it would not provide enough time for successful intervention to reduce crash risk in a proactive safety management strategy.

Moreover, the actual crash time might not precisely be known. Golob and Recker, 2004 discarded the 2.5 min of traffic data immediately preceding each reported crash time to avoid uncertainty of the actual crash time. In general with the proliferation of mobile phones and CCTV cameras on freeways, crash time is almost usually immediately identified. One-hour speed profiles were also generated (about 30 min before and 30 min after the crash time) to verify the reported crash time. The modeling procedure required non-crash data, a random selection from the whole remaining AVI and RTMS datasets where there was no crash within 2-h before the extraction time was utilized in the study to represent the whole population of different traffic patterns, weather conditions and roadway characteristics. A total of 18 (3 parameters $\times$ 3 AVI segments $\times$ 2 time slices) and 108 (9 parameters $\times$ 6 RTMSs $\times$ 2 time slices) input variables are prepared from AVI and RTMS raw data, respectively.

Similarly, weather data for crash cases and non-crash ceases were extracted. Automated weather stations monitor the weather conditions continuously and the weather parameters are recorded according to a specific change in the reading threshold and hence they do not follow a specific time pattern. The stations report frequent readings as the weather conditions change within short time; if the weather conditions remain the same the station would not update the readings. However, these readings were aggregated over certain time periods to represent the weather conditions. For example; precipitation described by rainfall amount or snowfall liquid equivalent for 10 min, 1 h, 3 h, 6 h, 12 h and 24 h and the estimated average hourly visibility which provides an hourly measure of the clear distance in miles that drivers can see. Visibility in general can be described as the maximum distance (in mile) that an object can be clearly perceived against the background sky, visibility impairment can be the result of both natural (e.g., fog, mist, haze, snow, rain, windblown dust, etc.) and human induced activities (transportation, agricultural activities, and fuel combustion). The automated weather stations do not directly measure the visibility but rather calculate it from a measurement of light extinction which includes the scattering and absorption of light by particles and gases.

The basic parameters that define the geometrical characteristics of the roadway section for each crash and non-crash cases were considered in this study, these parameters include longitudinal grade, curve radius, deflection angle, degree of curvature, number of lanes, and width of median.

Multiple Stochastic Gradient Boosting models were calibrated for each dataset separately as well as for fused data from all sources. Each of these data were partitioned into 70% for training, 30% for validation using random sampling, in random sampling every observation in the dataset has the same probability of being written to the sample. For example, the 70% of the population that is selected for the training dataset, then each observation in the input dataset has a 70% chance of being selected. Partitioning provides mutually exclusive datasets; two mutually exclusive datasets share no observations with each other. Partitioning is needed for machine learning (ML) models to have part of the dataset for training in order to fit a preliminary model and find the best model weights using this training dataset, and since ML techniques have the capacity for overtraining, validation dataset will be used to retreat to a simpler fit than to calibrate the model based only on the training dataset. Validation part of the original dataset is used for ML models fine-tuning to assess the prediction accuracy of each model. Although crashes involving driving under the influence of alcohol or drugs and distraction related crashes were less than 3% of the total crashes, they were excluded from the crash dataset to examine the effect of short-term turbulence of traffic, geometry and weather only. A total number of 186 crashes and 744 non-crashes were finally considered in the analysis.

3. Methodology

3.1. Stochastic Gradient Boosting

The Stochastic Gradient Boosting (SGB) is a machine learning technique that was introduced by (Friedman, 2001). This technique which is also known under other names such as Multiple Additive Regression Trees (MARTs), and TreeNet is technically suitable to be used for all data mining problems including regression, logistic regression, multinomial classification and survival models. The general idea of boosting is to create a series of simple learners known as "weak" or basic learners, i.e., a classifier that has a slightly lower error rate than random guessing. Most of the boosting algorithms use binary trees with only two terminal nodes as the basic learner (Hastie et al., 2001). Boosting these simple trees forms a single predictive model. The gradient boosting trees method has been proposed as a recent advancement in data mining that combines the advantages of the non-parametric tree-based methods and the strengths of boosting algorithms. It showed outstanding prediction performance in different fields including; real-time credit card fraud detection and terrorism culpability. The fraud detection application has some similarity to real-time crash prediction; with thousands of credit, debit and online

transactions taking place every minute; the probability of a fraud transaction is very small and the variables' space is relatively high, the mechanism that is deployed to monitor all transactions in real-time may be adopted in traffic safety applications.

Some of the key features of Stochastic Gradient Boosting are its ability of handling large number of mixed predictors (quantitative and qualitative) without preprocessing of rescaling or transformation which allows real-time traffic and weather data to be directly fed into the SGB algorithms without any time consuming processes. Moreover, by using CART as the basic learner, SGB can automatically handle the missing values which can still yield an accurate prediction in case of missing one of the important variables with no need to consider prior data imputation (Breiman et al, 1983). SGB has the capability of resisting the outliers in predictors and it can perform well with partially inaccurate data, therefore any erroneous traffic data can be handled easily without cleaning. Additional advantage of tree-based models is the robustness of variable selection; tree models have the capability of excluding irrelevant input variables. The main disadvantage, however, of single tree models is instability and poor predictive performance especially for larger trees which can be mitigated by other techniques that can improve model accuracy such as boosting, bagging, stacking, model averaging and ensemble which merges results from multiple models. Stochastic Gradient Boosting is uniquely advantageous over other merging techniques because it follows sequential forward stagewise procedure. The process of boosting is an optimization technique to minimize a loss function by adding a new simple learner (tree) at each step that best reduces the loss function, first tree is selected by the algorithm that maximally reduces the loss function. The residuals are the main focus for each following step by performing weighted resampling to boost the accuracy of the model by giving more attention to observations that are more difficult to classify. As the model enlarges, the existing trees are left unchanged; however, fitted value for each observation is to be re-estimated at each new added tree. The sampling weight is adjusted at the end of each iteration for each observation with respect to the accuracy of the model result. Observations with correct classification receive a lower sampling weight while incorrectly classified observations receive a higher weight. In the next iteration, a sample with more misclassified observations would be drawn.

SGB was used for classification in which, traffic, weather, and geometry variables are used as independent variables x to identify the binary crash $y \in \{-1, 1\}$, by using a "training" sample $\{y_i, x_i\}_1^N$ of known (y,x) values. The goal of estimating the function that maps the traffic, weather and geometry features to crashes is to be used for prediction of the increased risk for future observations, where only x is known. As explained in Friedman (2001) we need to obtain an approximation $F(x)$ of the function $F^*(x)$ linking x (traffic, weather and geometry predictors) to y (crash/no-crash), that minimize the expected value of a loss function $\Theta(y,F(x))$ over the joint distribution of all (y,x) values

$$F^*(x) = \arg\min_{F(x)} E_{y,x}\Theta(y, F(x)) \tag{1}$$

As mentioned earlier, the boosting idea is to build an additive model on a set of basic functions (weak classifier). In case of using a single tree as the individual classifier, the boosted tree model will be a sum of many simple trees;

$$f_T(x) = \sum_{m=1}^{M} T_m(x; \gamma_m, R_m) \tag{2}$$

where

$$T_m(x; \gamma_m, R_m) = \sum_{i}^{I_m} \gamma_{mi} I(x \in R_{mi}) \tag{3}$$

where R_{mi}, $i = 1, 2, \ldots, I_m$ are disjoint regions that collectively cover the space of all joint values of X. γ_{mi} is a constant that is assigned to each such region. R_{mi} is the ith terminal node in tree m with fitted value of γ_{mi}. Ideally, γ_{mi} and R_{mi} are fitted by minimizing a loss function;

$$\min_{\{\gamma_m R_m\}_1^M} \sum_{j-1}^{N} \Theta\left(y_j, \sum_{m=1}^{M} T_m(x; \gamma_m, R_m)\right) \tag{4}$$

Commonly used loss function for classification is given by;

$$\Theta(y, \widehat{F}) = 2\log(1 + exp(-2y\widehat{F})) \tag{5}$$

where

$$F(x) = \frac{1}{2}\log\left[\frac{\Pr(Y = 1|x)}{\Pr(Y = -1|x)}\right] \tag{6}$$

The solution can be approximated by iteratively adding a single tree at each step without adjusting the parameters of the existing trees as mentioned earlier. Therefore, by adding tree $k + 1$, the following equation can be minimized

$$\sum_{j=1}^{N} \Theta\left(y_j, \sum_{m=1}^{K} T_m(x_j; \gamma_m, R_m) + T_{k+1}(x_j; \gamma_{k+1}, R_{k+1})\right) \tag{7}$$

as a function of γ_{k+1} and R_{k+1}, holding $\gamma_1, \ldots, \gamma_k$ and $R_1, \ldots, R_k$ fixed. After M iterations (7) will achieve (4).

4. Results and discussion

4.1. Model estimation, interpretation and diagnostics

This section explains how the calibration, interpretation and evaluation processes were performed.

In this study, Stochastic Gradient Boosting models were fitted in SAS Enterprise Miner 6.1. The SGB was iterated 50 times with different random samples in the validation dataset to stabilize the error rate. The optimization parameters were set at SAS default values; shrinkage (learn rate) = 0.1, train proportion (different training observations are taken in each iteration) = 60, maximum branch = 2 (binary tree), and the maximum depth (number of generation) = 2.

In machine learning applications, the data may include easily hundreds of variables; a key question therefore whether or not all these variables actually lead to true information gain? The answer is obviously, no, since there are a lot of redundant variables that may increase the performance of the learning dataset but they do not necessarily increase the performance on the actual validation dataset which can be easily controlled for by keeping an eye on the over-fitting. Many data mining techniques such as neural networks, near-neighbor, kernel methods, and support vector machines perform worse when extra irrelevant predictors are added, and therefore a variable selection technique should always precede the modeling. On the other hand tree-based models are highly resistant to the inclusion of irrelevant variables; tree-based models perform automatic variable subset selection.

One of the main advantages of tree-based models is their simple interpretability. A single tree model can be graphically illustrated by two-dimensional figure that is easily interpreted. On the other hand, boosted trees are formed of linear combination of many trees (hundreds and in some cases thousands of trees), and therefore forfeit this important feature. The main two components of interpretation are identifying the variables importance and understanding their effect on the classification problem which are provided in all conventional regression models. Although, SGB provides insights on which variables are affecting crash occurrence and their relative importance, conventional statistics might be compulsory to provide information about the contributing effects of these predictors on the classification of crash/non-crash cases, and hence provides guidelines for the required countermeasures to reduce the increased risk of crashes in real-time. Previous research utilizing classical and Bayesian statistics were conducted to achieve such objective (Ahmed et al. in press-a,b). As mentioned earlier that one of the main goals of this research is to enhance the reliability of the classification of crashes in real-time, hence interpretation is not the main focus of this study.

Unlike other black-box machine learning techniques, SGB can be summarized and interpreted. Relative importance of predictor variables can be conveniently calculated. The variable importance is based on the number of times a variable is selected for the splitting rule and weighted by the squared improvement to the model as a result of each split, and averaged over all trees as explained in Friedman and Meulman (2003). The role of a predictor in a tree could be a main splitter or a surrogate. It is worth mentioning that a variable can be considered highly important even if it never appears as a node splitter since it may be used in a surrogate splits in the tree growing process and hence the contribution a variable can make in classification is not determined only by primary splits. For example, consider pairs of variables that contain similar information, such as speed variation from AVI and RTMS. Although only one of these variables can be used for main splits because it performs better than the other, the other variable could be the best surrogate to substitute the primary variable in case of the missing values. Fig. 2 provides the selected variable subsets and their relative importance for each of the calibrated models. The input variables characterized by a relative importance smaller than 25% have been discarded in the SGB models.

Stochastic Gradient Boosting models were estimated for four different datasets; Model-1 was calibrated using all available data collected from AVI, RTMS and weather stations as well as geometrical characteristics for crash/non-crash cases. In order to examine the prediction accuracy that can be achieved depending only on one dataset at a time and to account for any interruption of the data flow from any source, another three models were calibrated; Model-2 based on RTMS data, Model-3 based on AVI data, and Model-4 based on real-time weather data. It is worth mentioning that geometrical characteristics are always available and hence they were considered in all models.

It may be observed from Model-1 results that the most important variables are traffic data collected from RTMS such as average occupancies from US2 and US3 sensors during time slice two and three, respectively (time slice 2: 6–12 min before the crash and time slice 3: 12–18 min before the crash), followed by logarithm of the coefficient of variation of speed from AVI crash segment at time slice 2 and average speed from AVI downstream segment at time slice 2, other RTMS and AVI variables were selected but with less relative importance. It is clear that the measure of variation of speed might be more noticeable from AVI data rather than RTMS data, as mentioned earlier that space-mean-speed SMS collected from AVI provides information on a stretch of the road (AVI segment) while time-mean-speed TMS collected from RTMS reflects the traffic condition at only one specific point (RTMS station). Weather related variables are relatively important; 1-h visibility is shown at the top of the list just after some traffic variables. The 10-min precipitation variable was also selected among the important variables. Other site-related variables came out to be important including longitudinal grade, number of lanes, absolute degree of curvature, and width of median. Models 2–4 yield similar results with marginal difference in the order and value of the relative importance.

Comparison between the models' performance is subjective and depends on different criteria; misclassification rate and the area under the Receiver Operating Characteristics (ROC) were used as the main performance criteria in this analysis. The area under the ROC curve shows how well the model is at discriminating between the crash and non-crash cases in the target

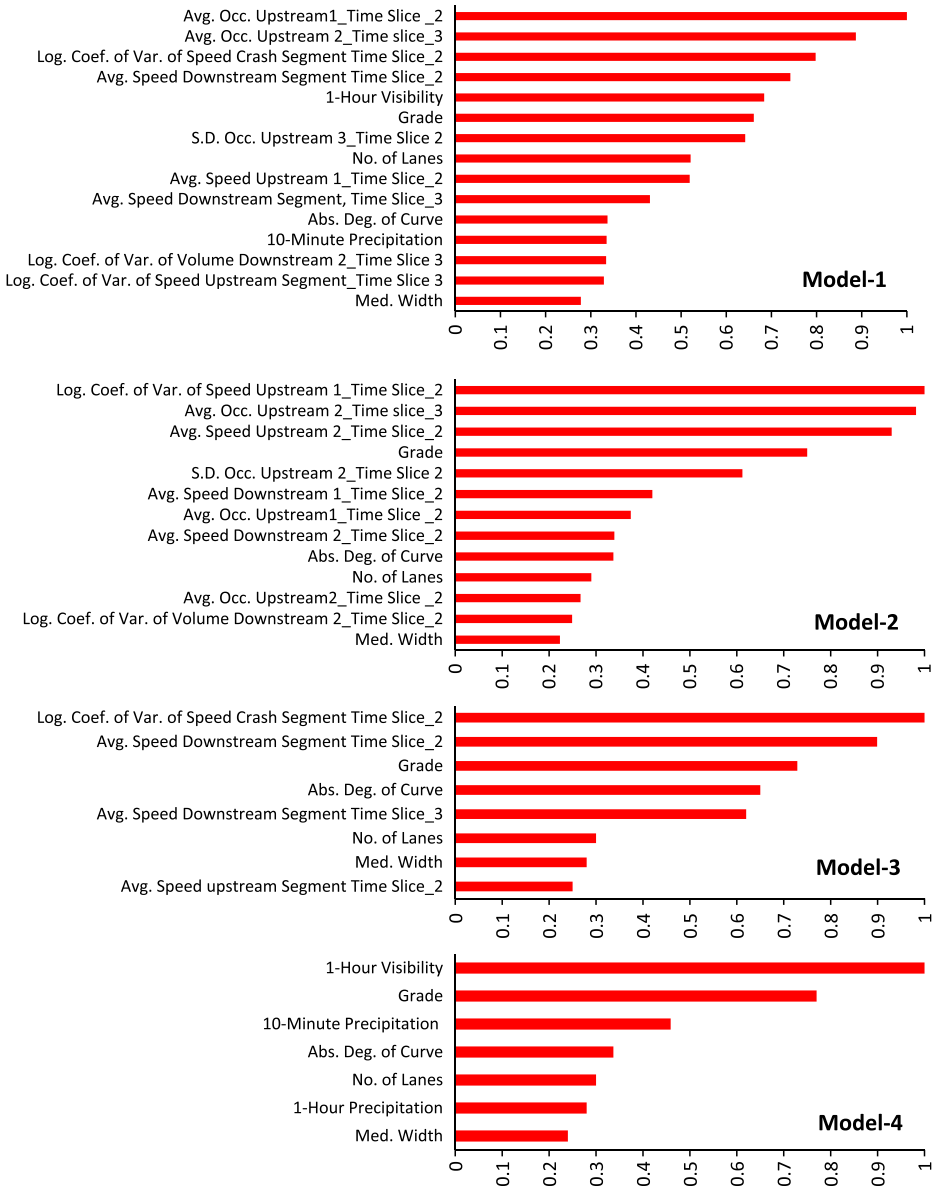

Fig. 2. Variable importance.

variable. This is similar to the misclassification rate, but the ROC curve plots sensitivity vs. 1 – specificity values for many cutoff points. Sensitivity (known also as true positive rate) is the ability to predict a crash case correctly and specificity (known as true negative rate) is the ability to predict a non-crash case correctly. The area under the curve seems to be large for the best selected model in red color (Model-1) as shown in Fig. 3 for the validation dataset. The exact areas under the ROC curves for all models validation datasets are listed in Table 1.

Generally, Model-1 is consistently superior in term of classification accuracy and area under the ROC curve. Model-2 is ranked second after the full model (Model-1), while Model-3 is relatively ranked lower than Model-1 and Model-2 but still providing satisfactory performance. Model-4 is ranked the lowest on these measures. Area under the ROC curves as shown in Fig. 3 and listed in Table 1 was found to be 0.946 for Model-1 validation dataset, 0.839 and 0.774 for Model-2 and Model-3, respectively while Model-4 achieved ROC of 0.731 all for the validation dataset.

Unlike previous studies that only reported accuracy and misclassification rate at one cutoff value, in this study the accuracy and misclassification rates are graphically illustrated for many cutoff values as shown in Figs. 4–7. In terms of accuracy and misclassification rate, also Model-1 outperformed all other individual models in all classification measures. Sensitivity analysis is important for the implementation of the proposed system in real-life application; while the overall classification rate can provide some insight of the model performance, sensitivity which is defined as the proportion of crashes (event

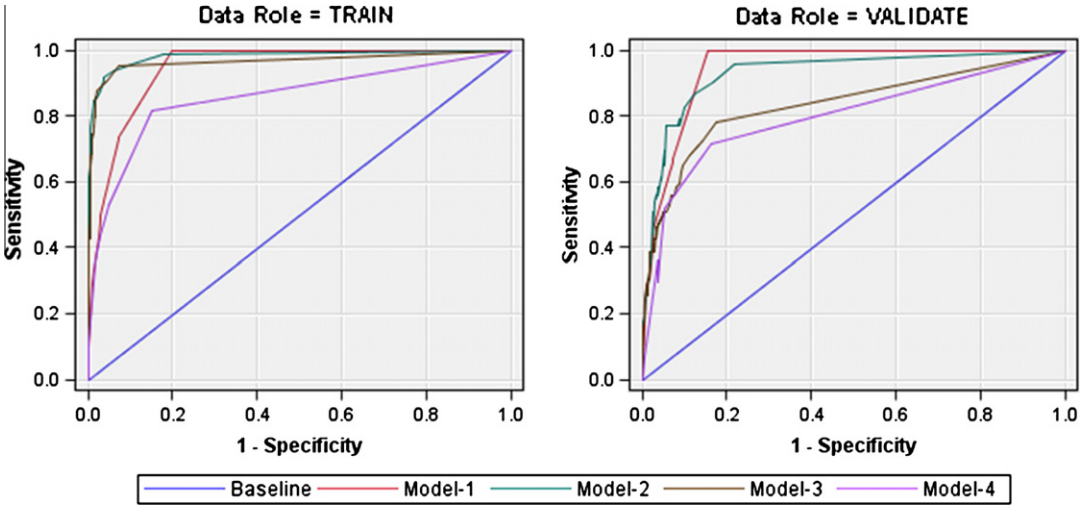

Fig. 3. Receiver operating characteristics chart.

Table 1
Validation: classification rates and ROC index.

Model	Model description	Overall classification rate (%)	True positive rate (%)	False positive rate (%)	True negative rate (%)	Valid: ROC index
Model-1	All data	92.157	88.889	6.481	93.519	0.946
Model-2	RTMS + geometry	89.838	81.538	6.601	93.399	0.839
Model-3	AVI + geometry	88.452	77.692	6.931	93.069	0.774
Model-4	Weather + geometry	85.208	61.057	6.567	93.443	0.731

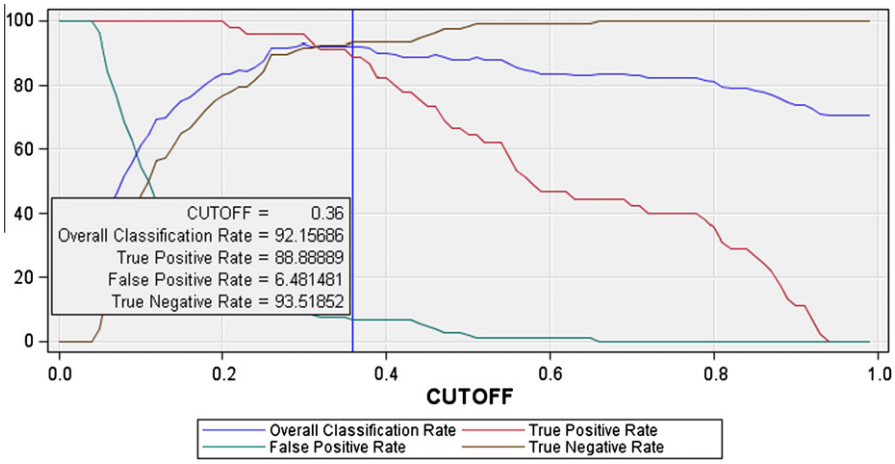

Fig. 4. Model-1 classification rates.

cases) that are correctly identified as crashes is usually the most important measure of accuracy. Other measure that may affect drivers' compliance to the management system and should be kept as minimum as possible is the proportion that is incorrectly classified as crashes (false positive rate).

As mentioned earlier that sensitivity analysis was conducted for the practical reason of implementing the models in real-time proactive safety management system where the sensitivity (capability of predicting events = 1) or predicting high probability of risk and reducing the false positive rates (false alarms) are considered the main focus for issuing warnings to motorists or managing speeds using Variable Speed Limits (VSLs). Sensitivity and false positive rates were used to choose the cut-off value. As shown in Figs. 4–7 that different false positive rates can be obtained by changing the cutoff value. In order to fairly compare across the four calibrated models, cutoff values have been chosen that achieve the highest possible

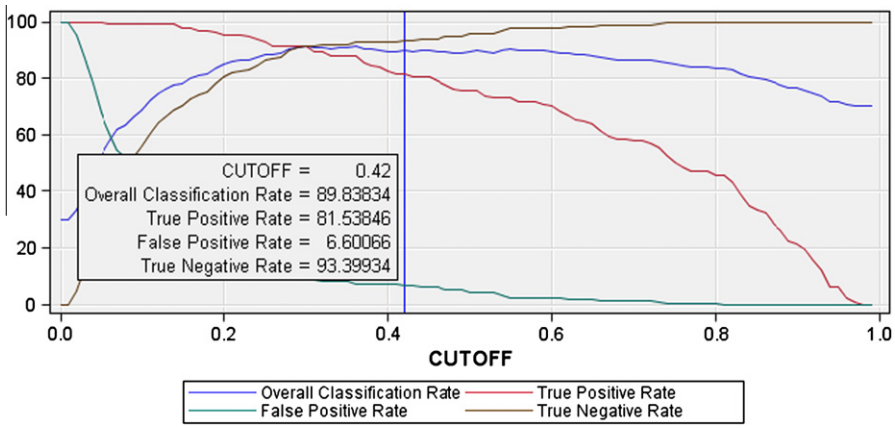

Fig. 5. Model-2 classification rates.

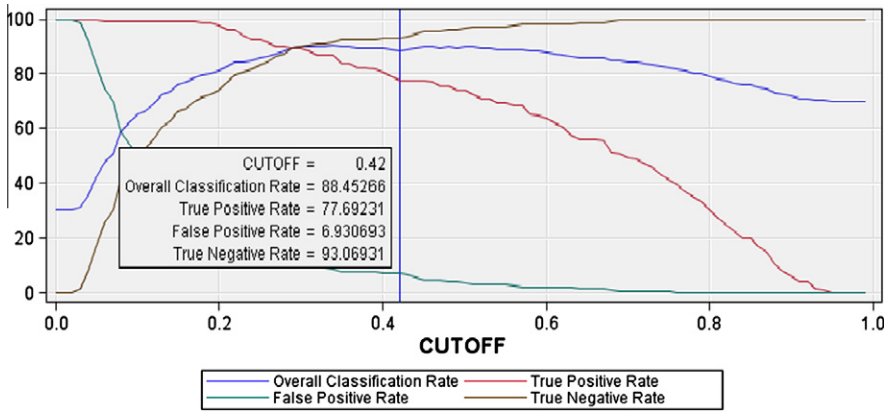

Fig. 6. Model-3 classification rates.

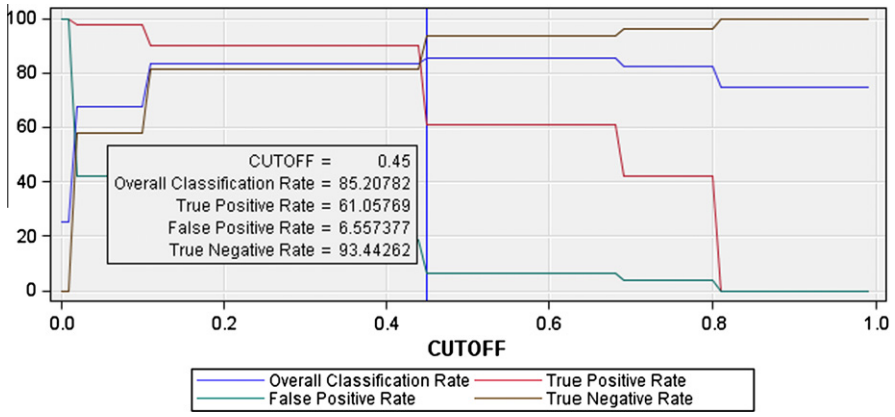

Fig. 7. Model-4 classification rates.

sensitivity while preserving false positive rates at low values under 7%, specificity (the proportion of correctly identified non-crashes) and overall classification. As illustrated in Figs. 4–7 and summarized in Table 1 for the chosen cutoff values, Model-1 identified about 89% of crashes correctly while only about 6.5% of non-crash cases were incorrectly identified as crashes; Model-1 also achieved the highest overall accuracy of about 92%. Model-2 and Model-3 ranked the second in term of overall accuracy with Model-2 performed slightly better than Model-3 to the respect of true positive rate and area under ROC curve as mentioned earlier. Model-4 achieved the lowest overall accuracy and true positive rate in the same range of false positive rate defined above.

Although Model-4 (weather and geometry model) performed not as good as the other three models, 61% accuracy of predicting high probability of crash risk with very low false positive rate is considered reasonable. The inclusion of weather information is essential in risk assessment; drivers need to have localized real-time information especially during adverse weather, including pavement conditions, visibility level, lane closure, snow, heavy rain and fog. The weather information would be more relevant if provided at the segment level rather than regional level. According to the Federal Highway Administration (Goodwin, 2002), weather contributed to over 22% of the total crashes in 2001. This means that adverse weather can easily increase the likelihood of crash occurrence. Several studies, in fact, concluded that crashes increase during rainfall by 100% or more (Brodsky and Hakkert, 1988; NTSB, 1998), while others found more moderate (but still statistically significant) increases (Andreescu and Frost, 1998; Andrey and Olley, 1990). Model-4 may provide an adequate measure of risk in scenarios where weather information is the only real-time data available and may help toward more weather responsive traffic management.

4.2. Risk assessment framework

The collected data on the study roadway section is one of the greatest assets that should be utilized appropriately to maximize the benefit for the roadway authority as well as for the road users. Buried within this vast amount of data is useful information that could make significant difference in how these roads are managed and operated. Fig. 8 illustrates a framework to assess the increased real-time risk depending on the availability of on-line data. The idea behind the proposed framework is based on the fact that although the traffic detection and meteorological stations became advanced enough to overcome hardware failures and malfunctions, the flow of the data might be interrupted in real-time at some point due to uncertain software and/or hardware failures. Therefore, a reliable and robust framework should be in place at all times. Moreover, another issue that was discussed but not explicitly addressed in previous studies is how different the prediction accuracy of traffic data that are collected from different sources at the same location in identifying real-time black spots on freeways.

There are four main models calibrated in the proposed framework; Model-1 based on all available data collected from AVI, RTMS, weather stations and roadway geometry, Model-2 based on geometry and RTMS data, Model-3 based on geometry and AVI data, and Model-4 based on geometry and real-time weather data. As shown in the flowchart in Fig. 8, in case of the availability of all traffic and weather data at the same time, these data would be fused together to provide the most comprehensive data and then Model-1 can be calibrated. If a hazardous traffic condition is detected, this section of the roadway would be flagged, otherwise, the section would be operated under normal condition. The other three models are calibrated for each data separately in addition to geometry data to examine how each model performs and to substitute the full model

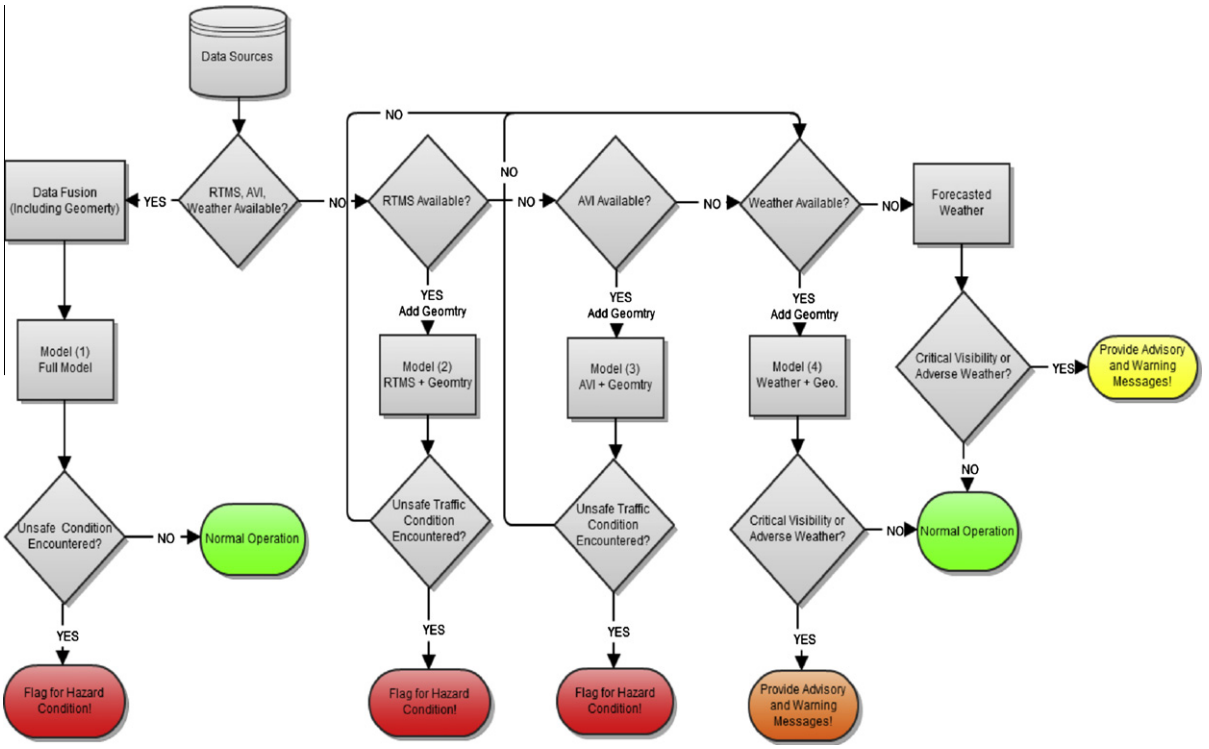

Fig. 8. Framework of the real-time risk assessment.

in case of absence of other data as mentioned earlier. Based on Model-2, a roadway section can be flagged if unsafe traffic condition was encountered otherwise Model-4 needs to be checked. If a critical visibility or adverse weather condition is predicted by Model-4 then an advisory/warning messages have to be issued to inform drivers about the situation. Although certain weather conditions might have a relatively insignificant impact on traffic (effect of light rain, snow and visibility on traffic capacities and average operating speeds), drivers may still need some advisory messages to help them in selecting the safe operating speed. In case that the real-time weather is not available, advisory messages can be issued depending on the forecasted weather. The same logic can be followed by Model-2 using data collected from AVI and geometry.

5. Conclusion

The recent advances in data collection technologies for traffic and weather on freeway sections provide valuable asset that should be utilized to increase safety and mobility and in order to maximize the benefit for highway authorities as well as for road users. These valuable data can be utilized to provide a framework for real-time risk assessment on freeways and expressways.

A relatively recent machine learning technique known under different names such as Stochastic Gradient Boosting (SGB), Multiple Additive Regression Trees (MARTs), and TreeNet was used to analyze 186 crashes occurred on 15-mile mountainous freeway section (I-70) in Colorado. The analyses were set up as a binary classification problem in which traffic, geometry, and weather variables are used as independent variable to identify crashes in real-time. The proposed learning machine methodology seems to provide all advantages that are needed in a real-time risk assessment framework. The Stochastic Gradient Boosting inherited all key strengths from tree-based models of their ability of selecting relevant predictors, fitting appropriate functions, accommodating missing values without the need for any prior transformation of predictor variables or elimination of outliers while overcoming the unstable prediction accuracy of single tree models. Boosting is considered unique among other popular aggregation methods; while ensemble, bootstrap or bagging, bagged trees and random forest can improve single tree models performance. Bagged trees and random forest can reduce variance more than single trees, however, unlike boosting; they cannot achieve any bias reduction (Prasad et al., 2006).

The proposed methodology has brought considerable advantage over classical statistical approaches. In particular, it has provided outstanding performance. On the other hand, machine learning techniques are being argued against for being black boxes; there are no *P* values to indicate the relative significance of model coefficients and there is no simple model with fewer variables. The proposed methods of interpretation (variable importance) and evaluation (ROC and classification) can be regarded as functional equivalence to many conventional regression techniques, thus addressing the criticisms against machine learning techniques.

Another issue that has been explicitly addressed in this study is how different the prediction accuracy of traffic data that are collected from different sources at the same location in identifying black spots on freeway sections in real-time; the results showed that crash prediction from AVI is comparably equivalent to RTMS data. Moreover, the accuracy of the main model that is augmenting information from multiple traffic detectors (AVI and RTMS), weather, and geometry performed the best in terms of classification rate and area under the ROC curve. The overall model (Model-1) identified about 89% of crash cases in the validation dataset with only 6.5% false positive.

Although the AVI system can provide measures about percentage of lane change per segment by comparing the unique tag ID for each individual vehicle at the beginning and end of the segment as well as providing space mean speed for each lane to estimate the variation in speed across lanes, the AVI algorithm and the archiving system in its current form do not report these information and hence an update of the AVI system might be needed.

This paper proposed a framework for real-time risk assessment using data from multiple sources that can achieve reliable and robust prediction performance under different scenarios of data availability. The results depict that traffic management authorities as well as road users can benefit from the wealth of collected data from multiple sources not only to alleviate traffic congestion but also to mitigate an increase in safety risk. The secondary but vital element would be the traffic control techniques (proactive intervention systems) that will be used to achieve the safer operation conditions. Route diversion, ramp metering, Variable Speed Limit (VSL), and Dynamic Message Signs (DMSs) can be used as intervention strategies. Among those strategies, VSL systems are proven to reduce recurrent congestion and speed variation, and maintain higher operating speeds on freeways. Integrating VSL and dynamic safety messages based on the estimated risk level within existing Advanced Traveler Information Systems (ATISs) would be a cost-effective added value to these systems. An informative and relevant message at the right time is the key to gain drivers' trust and compliance to the system which in return will improve the reliability of the system. Micro-simulation could be used to evaluate different scenarios of route diversion, ramp metering, and VSL. In order to come up with the most appropriate dynamic message(s), based on the findings from the statistical models, tailored sets of messages have to be tested at different traffic and weather conditions.

Acknowledgements

The authors wish to thank the Colorado Department of Transportation (CDOT) for providing the data that were used in this study, and for funding this research. The authors thank Rongjie Yu for his help in the data preparation. All opinions and results are solely those of the authors.

References

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

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

Ahmed, M., Abdel-Aty, M., Yu, R., in press-a. Crash Occurrence, Mountainous Freeway Geometry, Real-Time Weather, and Traffic Data from Automatic Vehicle Identification System: Assessment of Interaction and AVI traffic data. Transportation Research Record.

Ahmed, M., Abdel-Aty, M., Yu, R., in press-b. A Bayesian updating approach for real-time safety evaluation using AVI data. Transportation Research Record.

Ahmed, M., Abdel-Aty, M., 2011. The viability of using automatic vehicle identification data for real-time safety risk assessment. IEEE Transactions on Intelligent Transportation Systems 13 (2), 459–468.

Ahmed, M., Yu, R., Abdel-Aty, M., 2011. Safety applications of automatic vehicle identification and real-time weather data on freeways. In: The 18th World Congress on Intelligent Transport Systems, Orlando.

Andreescu, M., Frost, B.D., 1998. Weather and traffic accidents in Montreal, Canada. Climate Research 9, 225–230.

Andrey, J., Olley, R., 1990. Relationships between weather and road safety, past and future directions. Climatological Bulletin 24, 123–137.

Breiman, L., Friedman, J.H., Olshen, R., Stone, C., 1983. Classification and Regression Trees. Wadsworth.

Brodsky, H., Hakkert, A.S., 1988. Risk of a road accident in rainy weather. Accident Analysis and Prevention 20, 161–176.

Friedman, H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29 (5), 1189–1232.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Statistics in Medicine 22, 1365–1381.

Gerlough, D.L., Huber, M.J., 1975. Traffic Flow Theory: a Monograph. Special Report 165, Transportation Research Board. National Research Council, Washington, DC.

Golob, T., Recker, W., 2001. Relationships among urban freeway accidents, traffic flow, weather and lighting Conditions. California PATH. Working Paper UCB-ITS-PWP-2001-19, Institute of Transportation Studies. University of California, Berkeley.

Golob, T., Recker, W., 2004. A method for relating type of crash to traffic flow characteristics on urban freeways. Transportation Research – Part A, Policy and Practice 38, 53–80.

Goodwin, C.L., 2002. Weather Related Crashes on U.S. Highways. Federal Highway Administration. Falls Church, VA: Mitretek Systems, Inc. <http://ops.fhwa.dot.gov/Weather/best_practices/CrashAnalysis2001.pdf> (accessed December 2010).

Hall, L.F., 1996. Traffic stream characteristics. In: Gartner, N.H., Messer, C.J., Rathi, A.K. (Eds.), Traffic Flow Theory. US Federal Highway Administration.

Hastie, T., Tibshirani, R., Friedmann, J., 2001. The Elements of Statistical Learning. Springer-Verlag.

Hourdos, J., Garg, V., Michalopoulos, P., Davis, G., 2006. Real-time detection of crash-prone conditions at freeway high-crash locations. Transportation Research Record 1968, 83–91.

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

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, 67–78.

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

National Traffic Safety Board (NTSB), 1998. Fatal Highway Accidents on Wet Pavement – The Magnitude Location and Characteristics, NTIS, Springfield, VA. HTSB-HSS-80-1.

Oh, C., Oh, J., Ritchie, S., Chang, M., 2001. Real Time Estimation of Freeway Accident Likelihood.

Pande. A., Abdel-Aty, M., 2006a. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Compendium of Papers CD-ROM, Transportation Research Board 85th Annual Meeting, Washington, DC.

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

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.

SAS Institute Inc., 2009. Getting Started with SAS ® Enterprise Miner TM 6.1. SAS Institute Inc., Cary, NC.