Application of Stochastic Gradient Boosting Technique to Enhance Reliability of Real-Time Risk Assessment

Use of Automatic Vehicle Identification and Remote Traffic Microwave Sensor Data

Mohamed M. Ahmed and Mohamed Abdel-Aty

This study proposes a new and promising machine learning technique to enhance the reliability of real-time risk assessment on freeways. Stochastic gradient boosting (SGB) is used to identify hazardous conditions on the basis of traffic data collected from multiple detection systems such as automatic vehicle identification (AVI), remote traffic microwave sensors (RTMS), real-time weather stations, and roadway geometry. SGB's key strengths lie in its capability to fit complex nonlinear relationships; it handles different types of predictors and accommodates missing values with no need for prior transformation of the predictor variables or elimination of outliers, as with real-time applications. Boosting multiple simple trees together overcomes the poor prediction accuracy of single-tree models and provides fast and superior predictive performance. Three models are calibrated: a full model that augments all available data and another two models to compare explicitly the prediction performance of traffic data collected from different sources (AVI and RTMS) at the same location. The results from the preliminary analysis as well as the calibrated models indicate that crash prediction by AVI is comparable to that by RTMS data. Moreover, the full model achieves superior classification accuracy by identifying about 89% of crash cases in the validation data set with only a 6.5% false positive rate. Because of its superior classification performance and its minimal required data preparation, SGB is recommended as a promising technique for real-time risk assessment.

In recent years, advances in electronics have had a tremendous impact on enhancing and improving traffic surveillance systems; new nonintrusive traffic detection devices are in use more these days because of their ease of installation and maintenance in addition to their accuracy and affordable cost. The increased deployment of nonintrusive detection systems such as automatic vehicle identification (AVI) and remote traffic microwave sensors (RTMS) provides access to real-time traffic data from multiple sources. AVI is used mainly for toll collection and for travel time estimation purposes along freeways, whereas RTMS are used mostly for operations and incident management. 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. Research in the field of incorporating safety into freeway traffic management has utilized extensively traffic data collected from inductive loop detectors in real-time crash prediction (*1–9*). Recently, the usefulness of the collected traffic data from AVI has been investigated in real-time safety assessment (*10–13*).

Traffic data from AVI and RTMS as well as weather data are collected on a 15-mi stretch of mountainous Interstate-70 in Colorado to provide roadway users with important information about travel time, congestion, adverse weather conditions, and lane closures due to the danger of occasional avalanches, maintenance on the road, road crashes, or all three. Weather 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 Ahmed et al. it was found that classical statistical models provide interpretable models and acceptable accuracy of crash prediction with AVI and real-time weather data (*11, 12*), in this study a nonparametric machine learning technique is proposed to classify hazardous conditions by using traffic data from multiple sources, weather data, and geometry data. Machine learning methods are known for their superior classification and prediction performance over classical statistical ones. In order to enhance the accuracy and increase the reliability of real-time crash prediction, stochastic gradient boosting (SGB), a recent and promising machine learning technique, is used to uncover previously hidden patterns preceding a crash relative to noncrash conditions from the large amounts of roadway geometry, weather, and AVI and RTMS traffic data.

DATA DESCRIPTION AND PREPARATION

Five sets of data were used in this study: roadway geometry, crashes, and the corresponding AVI, RTMS, and weather data. The crash data were obtained from Colorado Department of Transportation for a 15-mi segment on I-70 for 13 months (from October 2010 to October 2011). The traffic data consist of space mean speed (SMS) captured by 12 and 15 AVI detectors located eastbound and westbound, respectively, along I-70. Volume, occupancy, and time mean speed (TMS) were

Department of Civil, Environmental, and Construction Engineering, University of Central Florida, 4000 Central Florida Boulevard, Orlando, FL 32816-2450. Current affiliation for M. M. Ahmed: Department of Civil and Architectural Engineering, College of Engineering and Applied Science, University of Wyoming, Engineering Building, Room 3055, 1000 East University Avenue, Laramie, WY 82071. Corresponding author: M. M. Ahmed, mahmed@uwyo.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2386, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 26–34.
DOI: 10.3141/2386-04

collected by 15 RTMS stations in each direction. AVI estimates SMS every 2 min, and 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 the roadway characteristics inventory and single-line diagrams.

In a previous study by the authors, 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 direction, and therefore these AVI segments and their respective RTMS stations were considered in the data extraction process and modeling (*10*). The crashes were assigned to the AVI segment and to the closest RTMS station; upstream and downstream AVI segments as well as three RTMS stations in the upstream and downstream directions were identified to extract their corresponding traffic data. The upstream, crash, and downstream segments were denoted U, C, and D, respectively, and the upstream and downstream RTMS stations were named US and DS, respectively, and assigned numbers in order from the closest to the farthest ones. Also, most of the RTMS stations occur at the same location as the AVI tag readers. The arrangement of RTMS and AVI segments and their spacing are illustrated in Figure 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 a 6-min level to obtain averages, standard deviations, and logarithms of the coefficient of variation (standard deviation divided by the average of the traffic parameters) of 2-min SMS obtained from AVI and 30-s TMS, volume, and occupancy raw data obtained from RTMS. In previous studies, a 5-min aggregation level was found to reduce the noise in the data and to provide better results than other aggregation levels (*4–7*). The 6-min aggregation level was chosen to have consistent time periods between AVI and RTMS data.

Three time slices of 6 min before the crash time were extracted. For example, if a crash happened on September 16, 2010 (Sunday), at 14:00, at Milepost 210.1 eastbound, the corresponding 18-min window was extracted for this crash of time intervals (13:42 to 14:00) recorded by AVI Segment 6 (mile markers start at 209.79 and end at 210.60), upstream AVI Segment 5, and downstream AVI Segment 7 as well as three RTMS stations in the upstream and three in the downstream direction. Time Slice 1 was discarded in the analysis since it would not provide enough time for successful intervention to reduce the crash risk in a proactive safety management system.

Moreover, 1-h speed profiles were generated (about 30 min before and 30 min after the crash time) to verify the reported crash time. The modeling procedure required noncrash data; a random selection from the whole remaining AVI and RTMS data sets in which 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 (three parameters times three AVI segments times two time slices) and 108 (nine parameters times six RTMS times two time slices) input variables were prepared from AVI and RTMS raw data, respectively.

Similarly, weather data for crash cases and noncrash cases 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; hence they do not follow a specific time pattern. The stations report frequent readings because the weather conditions change within a short time; if the weather conditions remain the same, the station does 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 miles) at which an object can be clearly perceived against the background sky; visibility impairment can be the result of both natural causes (e.g., fog, mist, haze, snow, rain, windblown dust) and human-induced activities (e.g., transportation, agricultural activities, and fuel combustion).

The basic parameters that define the geometric characteristics of the roadway section for each crash and noncrash case 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 SGB models were calibrated: the full model utilizing all data and another two models using only traffic data collected from AVI and RTMS. Each of these data sets was partitioned into 70% for training and 30% for validation with random sampling. In random sampling every observation in the data set has the same probability of being written to the sample; for example, of the 70% of the population that is selected for the training data set, each observation in the input data set has a 70% chance of being selected. Partitioning provides mutually exclusive data sets; two mutually exclusive data sets share no observations with each other. Partitioning is needed for machine learning models to have part of the data set for training in order to fit a preliminary model and find the best model weights. Since machine learning techniques have the capacity for overtraining, a validation

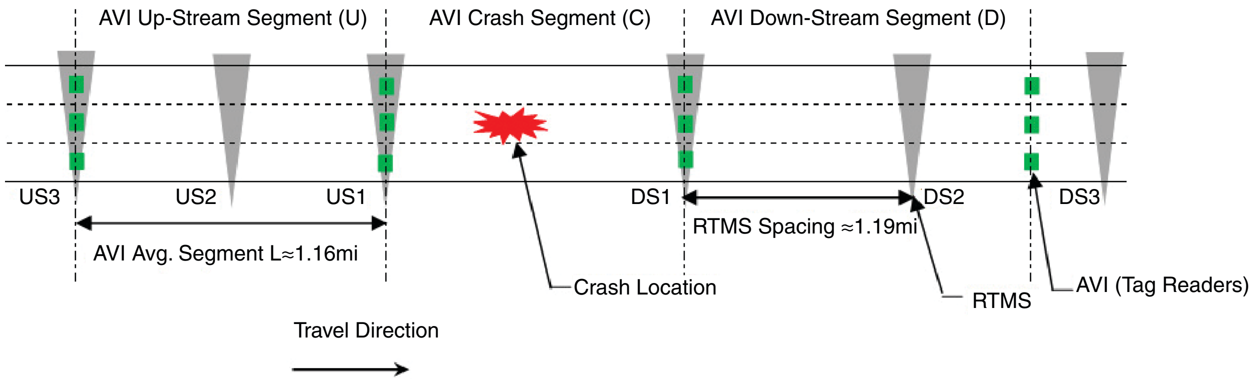


FIGURE 1 Arrangement of RTMS and AVI segments.

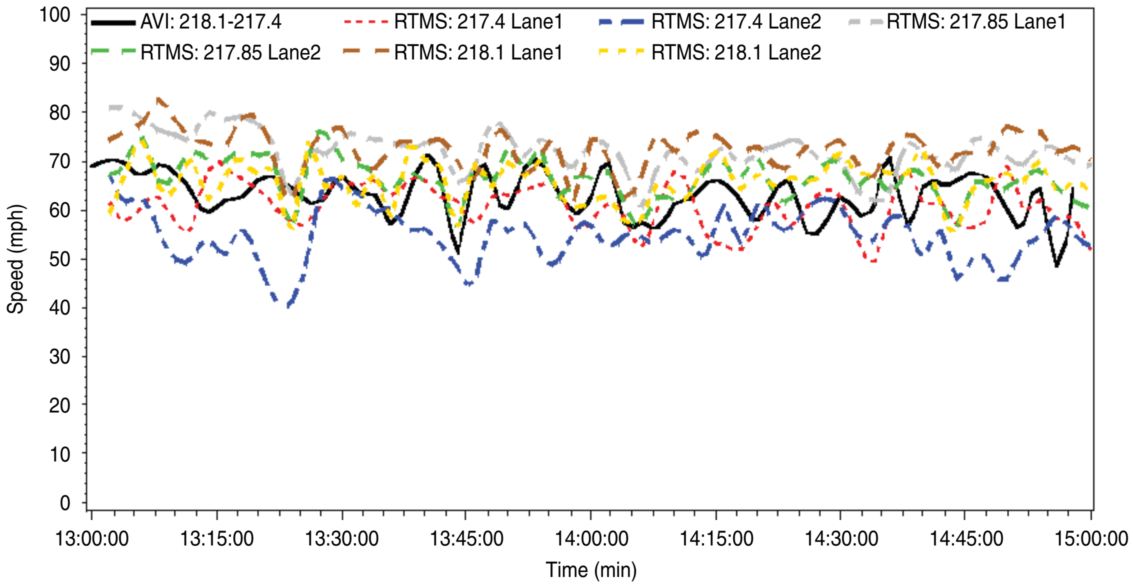
data set will be used to retreat to a simpler fit rather than to calibrate the model based only on the training data set. The validation part of the original data set is used for fine-tuning the machine learning models to assess the prediction accuracy of each model. Other data mining models (e.g., the artificial neural network and decision trees) were also estimated and compared with the SGB technique; however, they are not presented in this paper for brevity. Generally, SGB outperformed the decision tree models and performed relatively better than the artificial neural network models. 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 data set to examine the effect of short-term turbulence of traffic, geometry, and weather only. A total of 186 crashes and 744 noncrashes were finally considered in the analysis.

PRELIMINARY ANALYSIS OF AVI AND RTMS DATA

The RTMS stations provide TMS, flow, and lane occupancy and AVI provides only SMS. There are significant differences in the nature of the collected speed data from RTMS and AVI systems; the AVI SMS is defined by Gerlough and Huber 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 the RTMS TMS is the arithmetic mean of the speed of vehicles passing a point during a given time interval (*14*). Hence, TMS only reflects the traffic condition at one specific point. In contrast, 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 (*15*); the definition by Gerlough and Huber (*14*) is the best to describe the AVI SMS.] It is difficult to describe the measure of safety risk from fundamental notions of TMS and SMS without detailed analyses, and hence better understanding of these systems is essential in the safety context. Key questions therefore are, What level of accuracy could be achieved from each system? Which is more advantageous in real-time crash prediction, RTMS or AVI?

One main advantage of traffic data collected in this study is that both AVI and RTMS are at the same location, as illustrated in Figure 1. The spatial distribution of these detection devices facilitated direct comparison of the speed data. For a preliminary comparison between speed data collected from AVI and those from RTMS, speed profiles were generated for different scenarios of normal traffic conditions, crashes with property damage only, and crashes with an injury or fatality. For each case, an AVI segment and its corresponding RTMS were selected and 2 h worth of data were extracted, for example, 1 h before and 1 h after the crash. Although AVI can provide lane-by-lane information similar to RTMS, detailed specific lane speeds are only archived by RTMS.

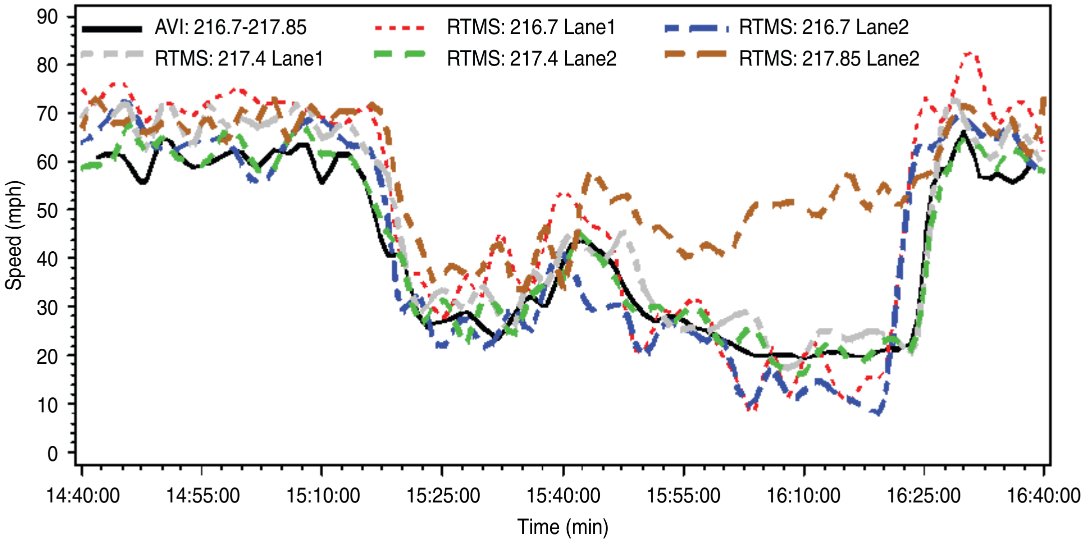
The results from various speed profiles revealed that patterns of speed from AVI and RTMS are comparable; in normal traffic conditions with no crashes, the AVI speed profile was between the profiles of upstream and downstream RTMS (Figure 2). Average speeds captured by RTMS on outer lanes are generally lower than those on the inner lanes because they are mostly for slower trucks. For crashes with property damage and an injury or fatality, the speed profiles before the crash times were very comparable except that the AVI system recorded higher speed variations than RTMS; this finding can be explained by the fact that the SMS for AVIs is aggregated temporally and spatially (the time and length needed to travel the AVI segment), and moreover the AVI SMS are aggregated across lanes. Figure 2 also shows the speed profile for a property-damage-only crash that occurred at 15:40 at Milepost 217.7 (the AVI crash segment starts at Milepost 216.7 and ends at Milepost 217.85). The crash was preceded by a drop in the average speed 25 min before and a high variability in speed 15 min before the time of the crash. The trend in all speed profiles from AVI and RTMS is similar; this finding provides enough evidence that AVI and RTMS data can substitute for each other when either one of them is not available.



(a)

FIGURE 2 AVI and RTMS speed profiles: (*a*) normal condition with no crash occurrence.

(continued)



(b)

FIGURE 2 *(continued)* AVI and RTMS speed profiles: (*b*) crash at Milepost 217.7.

METHODOLOGY

SGB is a machine learning technique that was introduced by Friedman (*16*). This technique, which is also known as multiple additive regression trees (MART) and TreeNet, is 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; that is, it is 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 (*17*). Boosting these simple trees forms a single predictive model. The gradient boosting tree method has been proposed as a recent advancement in data mining that combines the advantages of the nonparametric 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 low 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.

One of the key features of SGB is its ability to handle a large number of mixed predictors (quantitative and qualitative) without preprocessing of rescaling or transformation; this ability allows real-time traffic and weather data to be directly fed into the SGB algorithms without time-consuming processes. The machine learning technique used in this study was chosen to deal with the curse of dimensionality, which is usually found in real-time applications; in this study more than 125 covariates were used to discover traffic and weather patterns that preceded crash cases. Conventional statistics cannot handle such a large number of predictors and may also suffer from multicollinearity. Unlike classical statistical techniques, SGB is a nonparametric machine learning technique that does not require a linear form between the target variables and the covariates. Moreover, by using classification and regression trees as the basic learner, SGB can automatically handle the missing values; this feature can still yield an accurate prediction in the case where one of the important variables is missing with no need to consider prior data imputation (*18*). 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. An 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 of single-tree models is instability and poor predictive performance, especially for larger trees; this drawback 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.

SGB is uniquely advantageous over other merging techniques because it follows a sequential forward stagewise procedure. The process of boosting is an optimization technique to minimize a loss function by adding a new simple learner (a tree) at each step that best reduces the loss function; the first tree selected by the algorithm maximally reduces the loss function. The residuals are the main focus for each following step; weighted resampling to boost the accuracy of the model is performed by giving more attention to observations that are more difficult to classify. As the model enlarges, the existing trees are left unchanged; however, a fitted value for each observation is to be reestimated 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 a correct classification receive a lower sampling weight, whereas 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–no-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 by Friedman (*16*) an approximation $F(x)$ of the function $F^*(x)$ linking x (traffic, weather, and geometry predictors) to *y* (crash–no-crash), which minimizes the expected value of a loss function $\Theta(y, F(x))$ over the joint distribution of all (y, x)-values, is needed:

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

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

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

where

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

where

R_{mi}, $i = 1, 2, \ldots, I_m$ = disjoint regions that collectively cover space of all joint values of X,
γ_{mi} = constant assigned to each such region, and
R_{mi} = 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:

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

A commonly used loss function for classification is given by

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

where

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

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:

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

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 Equation 7 will be achieved (*4*).

RESULTS AND DISCUSSION

In this study, SGB models were fitted in SAS Enterprise Miner 6.1 (*19*). The SGB model was iterated 50 times with different random samples in the validation data set to stabilize the error rate. The optimization parameters were set at SAS default values: shrinkage (learning rate) = 0.1, training proportion (different training observations are taken in each iteration) = 60, maximum branch = 2 (binary tree), and the maximum depth (number of generations) = 2.

In machine learning applications, the data may easily include hundreds of variables; a key question therefore is whether 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 data set but do not necessarily increase the performance on the actual validation data set; this problem can be easily controlled for by keeping an eye on the overfitting. 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. However, 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 a two-dimensional figure. However, boosted trees are formed of a linear combination of many trees (hundreds and in some cases thousands of trees), and therefore they forfeit this important feature. The two main components of interpretation are identifying the variables' importance and understanding their effect on the classification problem; these components are provided in all conventional regression models. Although SGB provides insight 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–noncrash cases; hence guidelines are provided for the required countermeasures to reduce the increased risk of crashes in real time. Previous research with classical and Bayesian statistics was conducted to achieve such an objective (*12*). As mentioned earlier, one of the main goals of this research is to enhance the reliability of the classification of crashes in real time, and hence interpretation is not the main focus of this study.

Unlike other black-box machine learning techniques, SGB can be summarized and interpreted. The relative importance of predictor variables can be conveniently calculated. The variables' importance is based on the number of times a variable is selected for the splitting rule, weighted by the squared improvement to the model as a result of each split, and averaged over all trees as explained by Friedman and Meulman (*20*). The role of a predictor in a tree could be as a main splitter or a surrogate. A variable can be considered highly important even if it never appears as a node splitter since it may be used in surrogate splits in the tree-growing process; 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 for the primary variable in the case of missing values. Figure 3 shows the selected variable subsets and their relative importance for each of the calibrated models. The input variables characterized by a relative importance less than 25% were discarded in the SGB models.

SGB models were estimated for three different data sets; Model 1 was calibrated by using all available data collected from AVI, RTMS, and weather stations as well as geometric characteristics for crash–noncrash cases. To examine and compare the prediction accuracy that can be achieved by using data collected from AVI and RTMS, another two models were calibrated: Model 2, based on RTMS data, and Model 3, based on AVI data.

It may be observed from Model 1 results that the most important variables are traffic data collected from RTMS such as average

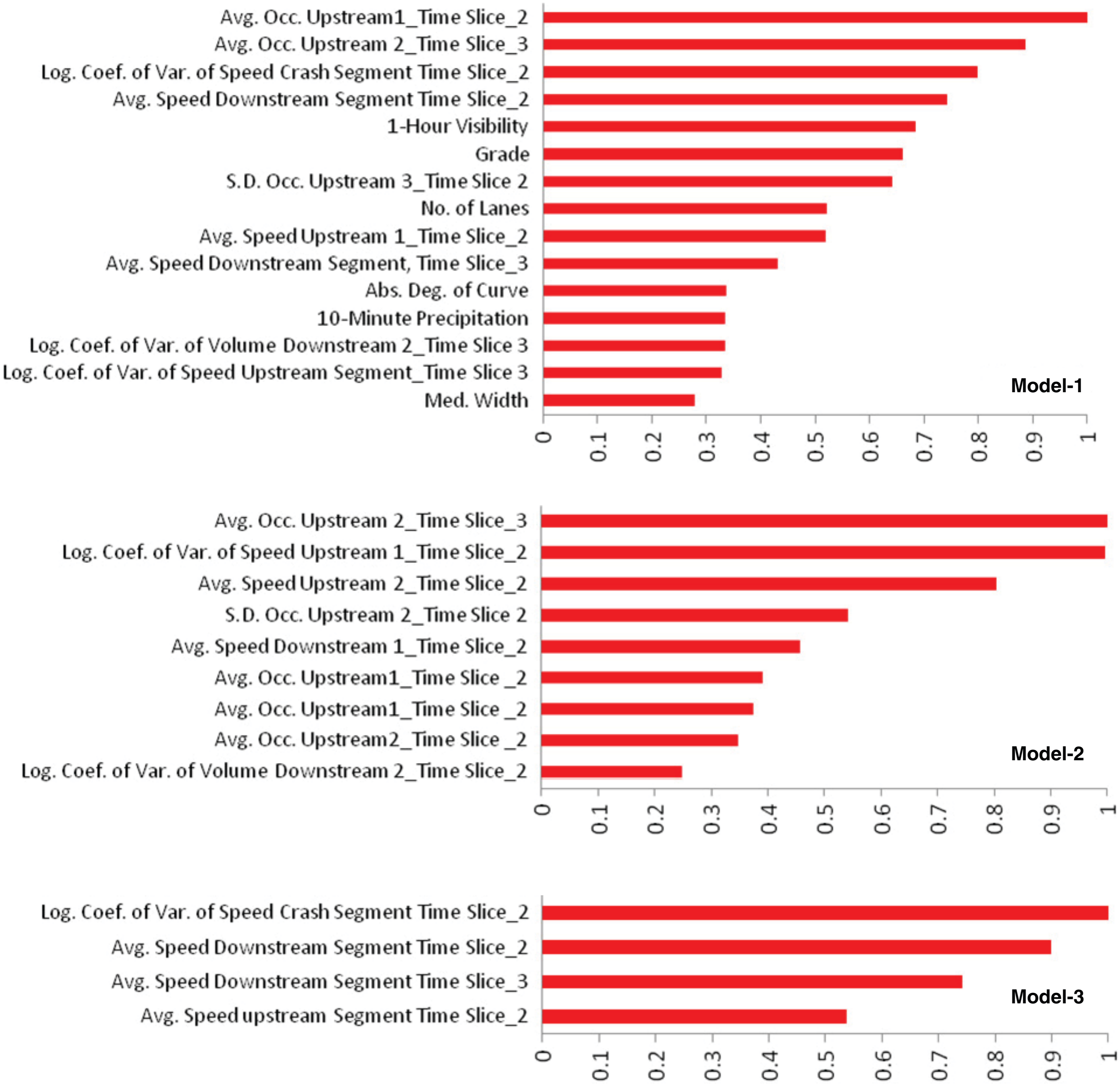


FIGURE 3 Importance of variable subsets (avg. = average; occ. = occupancy; log. coef. = logarithm of coefficient; var. = variation; S.D. = standard deviation; abs. deg. = absolute degree; med. = median).

occupancies from US-2 and US-3 sensors during Time Slices 2 and 3, respectively (Time Slice 2 was 6 to 12 min before the crash and Time Slice 3 was 12 to 18 min before the crash), followed by the logarithm of the coefficient of variation of speed from the AVI crash segment at Time Slice 2 and average speed from the AVI downstream segment at Time Slice 2; other RTMS and AVI variables were selected but had less relative importance.

It is clear that the variation of speed might be more noticeable from AVI data than RTMS data; as mentioned earlier, SMS collected from AVI provides information on a stretch of the road (the AVI segment), whereas TMS collected from RTMS reflects the traffic condition at only one specific point (the RTMS station). Weather-related variables are relatively important; 1-h visibility is at the top of the list in Figure 3 just after some traffic variables. The 10-min precipitation variable was also selected as an important variable. Other site-related variables were revealed to be important including longitudinal grade, number of lanes, absolute degree of curvature, and width of median. Models 2 and 3 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; the misclassification rate and the area under the receiver operating characteristics (ROC) curve were used as the main performance criteria in this analysis. The area under the ROC curve shows how well the model is discriminating between the crash and noncrash cases in the target variable. This variable is similar to the misclassification rate, but the ROC curve plots sensitivity versus 1 – specificity values for many cutoff points. Sensitivity (known also as the true positive rate) is the ability to predict a crash case correctly and specificity (known as the true negative rate) is the ability to predict a noncrash case correctly. The area under the curve seems to be large for the best selected model (Model 1), as shown in red in Figure 4 for the validation data set. The exact areas under the ROC curves for all model validation data sets are given in Table 1.

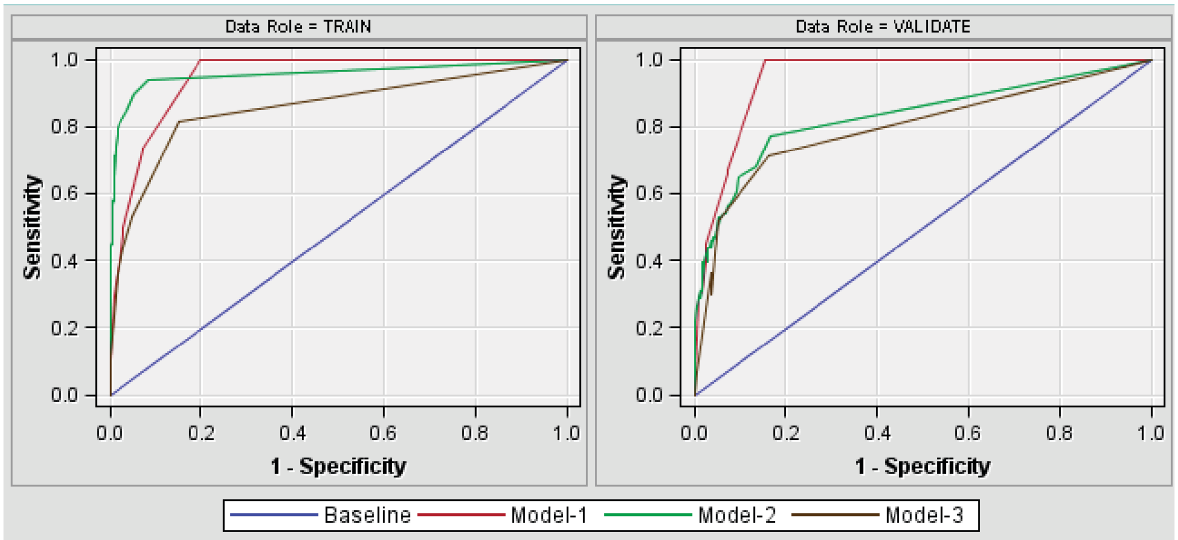


FIGURE 4 ROC curves.

Generally, Model 1 is consistently superior in terms of classification accuracy and area under the ROC curve. The RTMS model (Model 2) is ranked second after the full model (Model 1) and is followed by the AVI model (Model 3). The area under the ROC curve as shown in Figure 4 and Table 1 was found to be 0.946 for the Model 1 validation data set and 0.762 and 0.721 for Model 2 and Model 3, respectively.

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 Figure 5. In terms of accuracy and misclassification rate, Model 1 also outperformed all other individual models in all classification measures. Sensitivity analysis is important for implementation of the proposed system in a real-life application, and the overall classification rate can provide some insight into the model performance; sensitivity, which is defined as the proportion of crashes (event cases) that are correctly identified as crashes, is usually the most important measure of accuracy. Another measure that may affect drivers' compliance with the management system and should be kept as minimal as possible is the proportion that is incorrectly classified as crashes (false positive rate).

As mentioned earlier, a sensitivity analysis was conducted for the practical reason of implementing the models in a real-time proactive safety management system in which the sensitivity (capability of predicting events = 1) or prediction of a high probability of risk and reduction of false positive rates (false alarms) are considered the main focus for issuing warnings to motorists or managing speeds by using variable speed limits. Sensitivity and false positive rates were used to choose the cutoff value. As shown in Figure 5, different false positive rates can be obtained by changing the cutoff value. In order to fairly compare across the three calibrated models, cutoff values were chosen that achieve the highest possible sensitivity while preserving false positive rates at low values (less than 7.5%), specificity (the proportion of correctly identified noncrashes), and overall classification. As shown in Figure 5 and summarized in Table 1 for the chosen cutoff values, Model 1 identified about 89% of crashes correctly whereas only about 6.5% of noncrash cases were incorrectly identified as crashes. Model 1 also achieved the highest overall accuracy, about 92%. Models 2 and 3 ranked relatively lower than Model 1 in terms of overall accuracy; Model 2 performed slightly better than Model 3 with respect to the true positive rate and area under the ROC curve, as mentioned earlier. The results show that AVI data can provide comparable classification accuracy to the model using RTMS data. The calibrated Model 2 using only traffic surveillance data collected from RTMS achieved a classification accuracy of more than 73% of crashes with only 7.1% false alarms, whereas the model using only AVI data achieved more than 70% accuracy in classifying crash cases with less than 6.4% false alarms.

CONCLUSION

A relatively recent approach based on machine learning to identify increased risk on mountainous freeways in real time was presented. The SGB technique was utilized to analyze 186 crashes that occurred on a 15-mi mountainous freeway section of I-70 in Colorado. The analyses were set up as a binary classification problem in which

TABLE 1 Validation: Classification Rates and ROC Index

Model	Model Description	Overall Classification Rate (%)	True Positive Rate (%)	False Positive Rate (%)	True Negative Rate (%)	Validation: ROC Index
1	All data	92.157	88.889	6.481	93.519	0.946
2	RTMS	87.879	73.333	7.154	92.845	0.762
3	AVI	87.653	70.192	6.393	93.607	0.721

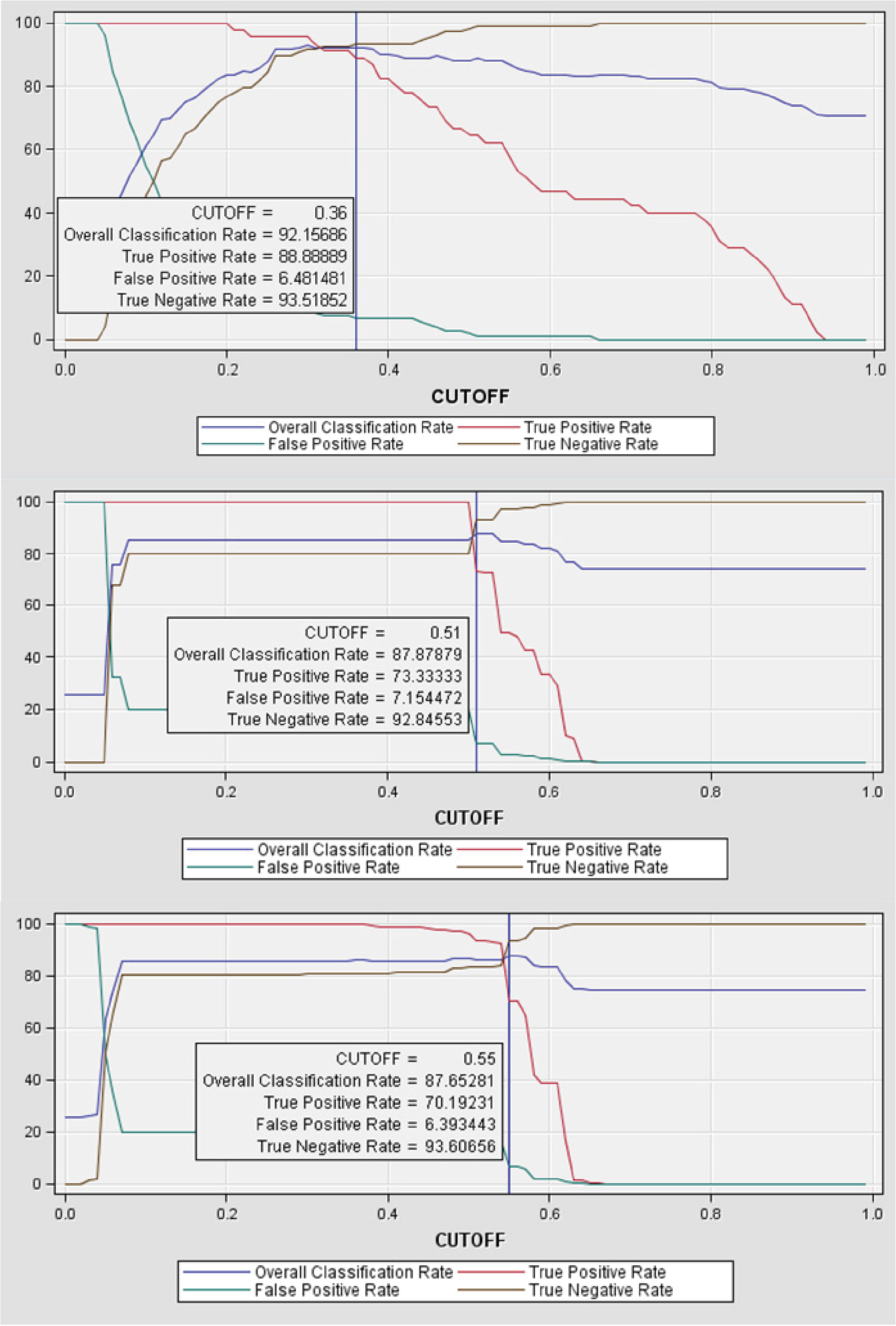


FIGURE 5 Classification rates.

traffic, geometry, and weather variables are used as independent variables to identify crashes in real time. The availability of data from two different surveillance systems, AVI and RTMS, and real-time weather and geometric characteristics on the same roadway section facilitated the collection of the most comprehensive data sets created for a real-time crash prediction study.

The proposed learning machine methodology seems to provide all the advantages that are needed in a real-time risk assessment framework. The SGB technique inherited all the key strengths from tree-based models: their ability to select relevant predictors, to fit appropriate functions, to accommodate missing values without the need for prior transformation of predictor variables, or to eliminate outliers while overcoming the unstable prediction accuracy of single-tree models. Boosting is considered unique among other popular aggregation methods, whereas ensemble, bootstrap or bagging, bagged trees, and random forest can improve single-tree model performance. Bagged trees and random forest can reduce variance more than single trees; however, unlike boosting they cannot achieve any bias reduction (*21*). The proposed methodology has a considerable advantage over classical statistical approaches. In particular, it has provided outstanding performance.

Another issue that was explicitly addressed here is how different the prediction accuracy is in identifying black spots on freeway sections in real time from traffic data collected from different traffic surveillance systems at the same location; the results showed that the accuracy of crash prediction from AVI is comparably equivalent to that from RTMS data. The calibrated model using only traffic surveillance data collected from RTMS achieved a classification accuracy of more than 73% of crashes, and the model using only AVI data achieved an accuracy of about 70% of crash cases with less than 7.5% false positive rate for both models. Moreover, the accuracy of the full model that augments all available data 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 full model (Model 1) identified about 89% of crash cases in the validation data set with only 6.5% false positives. The SGB technique provided far superior classification accuracy over conventional and Bayesian approaches. The accuracy of the prediction models using real-time traffic data in the literature was found to range between 44.73% and 75% (*1–13).*

Depending on online data availability, the results from these different models can be extended to develop a real-time risk assessment framework and maximize the benefit of such rich data to enhance the reliability of crash prediction on freeways. This framework would allow more proactive management strategies to help mitigate conditions of hazardous traffic and adverse weather.

ACKNOWLEDGMENT

The authors thank the Colorado Department of Transportation for providing the data that were used in this study and for funding this research.

REFERENCES

1. Oh, C., J. Oh, S. Ritchie, and M. Chang. Real-Time Estimation of Freeway Accident Likelihood. Presented at 80th Annual Meeting of the Transportation Research Board, 2001.
2. Golob, T., and W. Recker. *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, 2001.
3. Lee, C., F. Saccomanno, and B. Hellinga. Analysis of Crash Precursors on Instrumented Freeways. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1784,* Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 1–8.
4. Abdel-Aty, M., N. Uddin, A. Pande, M.F. Abdalla, and L. Hsia. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1897,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 88–95.
5. Abdel-Aty, M., and A. Pande. Identifying Crash Propensity Using Specific Traffic Speed Conditions. *Journal of Safety Research,* Vol. 36, 2005, pp. 97–108.
6. Abdel-Aty, M., and A. Pande. Comprehensive Analysis of the Relationship Between Real-Time Traffic Surveillance Data and Rear-End Crashes on Freeways. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1953,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 31–40.
7. Pande, A., and M. Abdel-Aty. Assessment of Freeway Traffic Parameters Leading to Lane-Change Related Collisions. *Accident Analysis and Prevention,* Vol. 38, 2006, pp. 936–948.
8. Hourdos, J. N., V. Garg, P. G. Michalopoulos, and G. A. Davis. Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1968,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 83–91.
9. Hossain, M., and Y. Muromachi. A Bayesian Network Based Framework for Real-Time Crash Prediction on the Basic Freeway Segments of Urban Expressways. *Accident Analysis* and Prevention, Vol. 45, 2012, pp. 373–381.
10. Ahmed, M., and M. Abdel-Aty. The Viability of Using Automatic Vehicle Identification Data for Real-Time Safety Risk Assessment. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 13, No. 2, 2011, pp. 459–468.
11. Ahmed, M., R. Yu, and M. Abdel-Aty. Safety Applications of Automatic Vehicle Identification and Real-Time Weather Data on Freeways. Presented at 18th World Congress on Intelligent Transport Systems, Orlando, Fla., 2011.
12. Ahmed, M.M., M. Abdel-Aty, and R. Yu. Assessment of Interaction of Crash Occurrence, Mountainous Freeway Geometry, Real-Time Weather, and Traffic Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2280*, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 51–59.
13. Ahmed, M. M., M. Abdel-Aty, and R. Yu. Bayesian Updating Approach for Real-Time Safety Evaluation with Automatic Vehicle Identification Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2280*, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 60–67.
14. Gerlough, D. L., and M. J. Huber. *Special Report 165: Traffic Flow Theory: A Monograph.* TRB, National Research Council, Washington, D.C., 1975.
15. Hall, L. F. Traffic Stream Characteristics. In *Traffic Flow Theory* (N. H. Gartner, C. J. Messer, and A. K. Rathi, eds.), FHWA, U.S. Department of Transportation, 1996, Chap. 2.
16. Friedman, H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics,* Vol. 29, No. 5, 2001, pp. 1189–1232.
17. Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag, Berlin, 2001.
18. Breiman, L., J. H. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Monterey, Calif., 1983.
19. *Getting Started with SAS Enterprise Miner TM 6.1.* SAS Institute Inc., Cary, N.C., 2009.
20. Friedman, J. H., and J. J. Meulman. Multiple Additive Regression Trees with Application in Epidemiology. *Statistics in Medicine,* Vol. 22, 2003, pp. 1365–1381.
21. Prasad, A. M., L. R. Iverson, and A. Liaw. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems,* Vol. 9, 2006, pp. 181–199.

All opinions and results are those of the authors.

The Safety Data, Analysis, and Evaluation Committee peer-reviewed this paper.