Proactive Assessment of Accident Risk to Improve Safety on a System of Freeways

MTI Report 11-15

MINETA
TRANSPORTATION INSTITUTE
M T I

MINETA TRANSPORTATION INSTITUTE

The Norman Y. Mineta International Institute for Surface Transportation Policy Studies (MTI) was established by Congress as part of the Intermodal Surface Transportation Efficiency Act of 1991. Reauthorized in 1998, MTI was selected by the U.S. Department of Transportation through a competitive process in 2002 as a national "Center of Excellence." The Institute is funded by Congress through the United States Department of Transportation's Research and Innovative Technology Administration, the California Legislature through the Department of Transportation (Caltrans), and by private grants and donations.

The Institute receives oversight from an internationally respected Board of Trustees whose members represent all major surface transportation modes. MTI's focus on policy and management resulted from a Board assessment of the industry's unmet needs and led directly to the choice of the San José State University College of Business as the Institute's home. The Board provides policy direction, assists with needs assessment, and connects the Institute and its programs with the international transportation community.

MTI's transportation policy work is centered on three primary responsibilities:

Research

MTI works to provide policy-oriented research for all levels of government and the private sector to foster the development of optimum surface transportation systems. Research areas include: transportation security; planning and policy development; interrelationships among transportation, land use, and the environment; transportation finance; and collaborative labor-management relations. Certified Research Associates conduct the research. Certification requires an advanced degree, generally a Ph.D., a record of academic publications, and professional references. Research projects culminate in a peer-reviewed publication, available both in hardcopy and on TransWeb, the MTI website (http://transweb.sjsu.edu).

Education

The educational goal of the Institute is to provide graduate-level education to students seeking a career in the development and operation of surface transportation programs. MTI, through San José State University, offers an AACSB-accredited Master of Science in Transportation Management and a graduate Certificate in Transportation Management that serve to prepare the nation's transportation managers for the 21st century. The master's degree is the highest conferred by the California State University system. With the active assistance of the California Department of Transportation, MTI delivers its classes over a state-of-the-art videoconference network throughout the state of California and via webcasting beyond, allowing working transportation professionals to pursue an advanced degree regardless of their location. To meet the needs of employers seeking a diverse workforce, MTI's education program promotes enrollment to under-represented groups.

Information and Technology Transfer

MTI promotes the availability of completed research to professional organizations and journals and works to integrate the research findings into the graduate education program. In addition to publishing the studies, the Institute also sponsors symposia to disseminate research results to transportation professionals and encourages Research Associates to present their findings at conferences. The World in Motion, MTI's quarterly newsletter, covers innovation in the Institute's research and education programs. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

REPORT 11-15

PROACTIVE ASSESSMENT OF ACCIDENT RISK TO IMPROVE SAFETY ON A SYSTEM OF FREEWAYS

Anurag Pande, Ph.D.
Cornelius Nuworsoo, Ph.D.
Cameron Shew

May 2012

TECHNICAL REPORT DOCUMENTATION PAGE

1. **Report No.** CA-MTI-12-1006	2. **Government Acession No.**	3. **Recipient's Catalog No.**
4. **Title and Subtitle** Proactive Assessment of Accident Risk to Improve Safety on a System of Freeways		5. **Report Date** May 2012
		6. **Performing Organization Code**
7. **Authors** Anurag Pande, Ph.D., Cornelius Nuworsoo, Ph.D. and Cameron Shew		8. **Performing Organization Report** MTI Report 11-15
9. **Performing Organization Name and Address** Mineta Transportation Institute College of Business San José State University San José, CA 95192-0219		10. **Work Unit No.**
		11. **Contract or Grant No.** DTRT07-G-0054
12. **Sponsoring Agency Name and Address** California Department of Transportation Office of Research—MS42 P.O. Box 942873 Sacramento, CA 94273-0001 U.S. Department of Transportation Research & Innovative Technology Admin. 1200 New Jersey Avenue, SE Washington, DC 20590		13. **Type of Report and Period Covered** Final Report
		14. **Sponsoring Agency Code**

15. **Supplemental Notes**

16. **Abstract**

This report describes the development and evaluation of real-time crash risk-assessment models for four freeway corridors: U.S. Route 101 NB (northbound) and SB (southbound) and Interstate 880 NB and SB. Crash data for these freeway segments for the 16-month period from January 2010 through April 2011 are used to link historical crash occurrences with real-time traffic patterns observed through loop-detector data.

The crash risk-assessment models are based on a binary classification approach (crash and non-crash outcomes), with traffic parameters measured at surrounding vehicle detection station (VDS) locations as the independent variables. The analysis techniques used in this study are logistic regression and classification trees.

Prior to developing the models, some data-related issues such as data cleaning and aggregation were addressed. The modeling efforts revealed that the turbulence resulting from speed variation is significantly associated with crash risk on the U.S. 101 NB corridor. The models estimated with data from U.S. 101 NB were evaluated on the basis of their classification performance, not only on U.S. 101 NB, but also on the other three freeway segments for transferability assessment. It was found that the predictive model derived from one freeway can be readily applied to other freeways, although the classification performance decreases. The models that transfer best to other roadways were determined to be those that use the least number of VDSs–that is, those that use one upstream or downstream station rather than two or three.

The classification accuracy of the models is discussed in terms of how the models can be used for real-time crash risk assessment. The models can be applied to developing and testing variable speed limits (VSLs) and ramp-metering strategies that proactively attempt to reduce crash risk.

17. **Key Words** Real-time crash risk; Data mining; Classification tree; Proactive traffic management; Loop detector data	18. **Distribution Statement** No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161		
19. **Security Classif. (of this report)** Unclassified	20. **Security Classif. (of this page)** Unclassified	21. **No. of Pages** 86	22. **Price** $15.00

Form DOT F 1700.7 (8-72)

Library of Congress Catalog Card Number:
2012930704

051712

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This report describes the development and evaluation of real-time crash risk-assessment models for four freeway corridors: U.S. Route 101 NB (northbound) and SB (southbound) and Interstate 880 NB and SB. Crash data for these freeway segments for the 16-month period from January 2010 through April 2011 are used to link historical crash occurrences with real-time traffic patterns observed through loop-detector data.

The analysis techniques used in this study are logistic regression and classification trees, which are among the most common data mining tools. The crash risk-assessment models are based on a binary classification approach (crash and non-crash outcomes), with traffic parameters measured at surrounding vehicle detection station (VDS) locations as the independent variables. The authors developed the classification-performance assessment methodology that accounts for the rarity of crashes compared to non-crash cases in the sample rather than using a pre-specified threshold-based classification.

Prior to developing the models, some data-related issues such as data cleaning and aggregation were addressed. The modeling efforts revealed that the turbulence resulting from speed variation is significantly associated with crash risk on the U.S. 101 NB corridor. The models estimated with data from U.S. 101 NB were evaluated on the basis of their classification performance, not only on U.S. 101 NB, but also on the other three freeways for assessment of transferability. It was found that the predictive model derived from one freeway can be readily applied to other freeways, although the classification performance decreases. The models that transfer best to other roadways were determined to be those that use the least number of VDSs–that is, those that use one upstream or downstream station rather than two or three.

The classification accuracy of the models is discussed in terms of how the models can be used for real-time crash risk assessment. The models can be applied to developing and testing variable speed limits (VSLs) and ramp-metering strategies that proactively attempt to reduce crash risk.

I. INTRODUCTION

Much progress has been made in recent years in shifting from reactive (incident detection) to proactive (real-time crash risk assessment) traffic strategies as traffic safety on freeways continues to be a concern. Reliable models that can use real-time loop-detector information and distinguish normal flow conditions from crash-prone conditions are keys to implementing crash-preventive measures. This area of research has gained increased attention since vehicle detector stations (VDSs) on freeways have been able to gather real-time traffic data, and the ability to collect, archive, and analyze these data has grown.

This report presents the findings of a study sponsored by the Mineta Transportation Institute (MTI) and carried out jointly by California Polytechnic State University, San Luis Obispo and San José State University. The study team developed statistical models relating traffic-flow variables to crash likelihood and also tested the transferability of these models to other, nearby freeway corridors. A few past studies have demonstrated that statistical links between real-time traffic-flow variables (such as average speed, volume, occupancy, and their respective standard deviations) and crash likelihood can be established. However, those studies focused primarily on one particular highway corridor. The research reported here advances the current body of knowledge by exploring whether driver characteristics and behavior in close geographic proximity are similar enough to accurately apply the estimated classification models from one roadway segment to another. While safety applications using intelligent transportation systems (ITSs) need to be examined further, this study took the following steps to estimate the crash risk-estimation models and assess their transferability:

1. Assemble a database of archived loop-detector data for four study segments within the milepost range in the vicinity of the San José, CA, metropolitan area—U.S. Route 101 NB and SB (northbound and southbound) and Interstate 880 NB/SB— for the 16-month period from January 1, 2010, to April 30, 2011.

2. Assemble a database of observed crash data for the same period, including information on date, time, and location of crashes, from the Performance Measurement System (PeMS) database for the study period.

3. Create a database of "normal" conditions that provides 10 "normal" observations for each crash. The date, time, and location of these non-crashes were randomly chosen from the range of all possible combinations of date, time, and location for the 16-month period. These were times/locations in which no crashes were observed; using these data along with the crash information, the researchers set up the database for binary classification.

4. Extract loop-detector data for all crash and non-crash events, using the date, time, and milepost information from the PeMS database.

5. Perform statistical (logistic regression) and data mining (classification tree) analyses to fit the most appropriate classification model for explaining the effects of traffic-flow variables on crash risk. These variables are measured at different locations

upstream and downstream of the crash and from different time durations prior to the crash to gain an understanding of the spatiotemporal impact these variables have on crash risk.

6. Select the best models estimated from the U.S. 101 NB crash and non-crash data and use them to score the datasets (which include both crash and non-crash observations) for U.S. 101 SB and I-880 NB and SB.

7. Examine the classification performance of the models on these datasets in the context of a real-time application.

Chapter II reviews relevant past research efforts, including those aimed at real-time identification of crash-prone conditions. Chapter III presents background information on the study area and the data-preparation process. Chapter IV presents the results of the logistic-regression and data mining models and examines how well they performed on nearby freeways. Chapter V discusses the conclusions drawn from these results and other relevant issues regarding their application.

II. LITERATURE REVIEW

This chapter reviews previous studies from the literature on traffic safety, with real-time identification of crash-prone conditions on freeways, and on data mining applications in incident detection and crash analysis. The safety studies are further categorized into exploratory studies and studies establishing statistical links. All of these studies are fairly recent, indicating that the idea of using loop-detector data for traffic-safety applications is still in its early stages.

SAFETY APPLICATIONS OF ARCHIVED INTELLIGENT TRANSPORTATION SYSTEM DATA

Golob, Recker, and Alvarez (2004b) categorized traffic-safety-related studies into two groups: aggregate studies and disaggregate analysis. In the aggregate studies, units of analysis represent counts of crashes or crash rates for specific time periods (typically months or years) and locations (specific roads or networks). The traffic flow in these studies is represented by the parameters of statistical distributions of traffic, such as annual average daily traffic (AADT), for similar time and location (Zhou and Sisiopiku 1997). In the disaggregate analysis, the units of analysis are the crashes themselves, and traffic flow is represented by parameters of traffic flow at the time and location of each crash.

While determination of freeway crash patterns has been the stated focus of the traffic-safety literature, most of the efforts are aggregate studies. Disaggregate studies are relatively new and have been made possible by recent enhancements in the ability to collect, store, and analyze real-time traffic data through intelligent transportation system (ITS) applications. This section summarizes and critically reviews the disaggregate analyses in the literature.

Exploratory Studies

Hughes and Council (1999) were among the first authors to use loop-detector data to explore the relationship between freeway safety and peak-period operations. They concluded that macroscopic measures such as AADT and even hourly volume correlate poorly with real-time system performance. Most of their work relied upon the data coming from a single milepost location during peak periods, on which they tried to overlay the crash time at that location to make inferences about the changes in system performance as the time of the crash approaches. The changes in performance were also examined from "snapshots" provided by cameras installed on the freeway.

One of their most important observations was that "design inconsistency"—the non-uniform application of geometric design standards—is a key factor in crash causation. Future research should consider "traffic-flow consistency," that is, the variability in traffic parameters (such as speed, volume, and occupancy) as an important variable from a human-factor standpoint. Hughes and Council also expressed a need for determining the exact time of a crash to avoid the "cause and effect" fallacy.

Studies Establishing Statistical Links

Madanat and Liu (1995) developed an incident-likelihood prediction model using loop data as input. The focus of their research was on enhancing existing incident-detection algorithms with the likelihood of crashes and overheating vehicles. The methodology they used for analysis was binary logit. They concluded that merging sector, visibility, and rain are statistically the most significant factors for crash-likelihood prediction.

Lee, Saccomanno, and Hellinga (2002) introduced the concept of "crash precursors" and hypothesized that the likelihood of crash occurrence is significantly affected by short-term turbulence of traffic flow. They identified factors such as speed variation along the length of the roadway (i.e., the difference between the speeds upstream and downstream of the crash location) and across the three lanes at the crash location. Another important factor they identified was traffic density at the instant of the crash. Weather, road geometry, and time of day were used as external controls. With these variables, they developed a crash-prediction model using log-linear analysis. The log-linear model was chosen so that the exposure could be easily determined; this would have been difficult if a logit model had been used. To test the goodness of fit of the model, a Pearson chi-square test was performed, measuring how close the expected frequencies are to the observed frequencies for any combination of crash precursors and control factors. At the 95% confidence level, the model yielded a good fit.

In a subsequent study, Lee, Hellinga, and Saccomanno (2003) continued work along the same lines and modified their earlier model. They incorporated an algorithm to obtain a better estimate of the time of the crash and the length of the time slice (prior to the crash), that is, the duration to be examined. They concluded that variation of speed has a relatively longer-term effect on crash potential than density or the average speed difference between upstream and downstream ends of roadway sections. They also observed that the average variation-of-speed difference across adjacent lanes does not have a direct impact on crashes, and hence it was eliminated from the model.

The prediction models in both studies relied upon the log-linear models developed in the past to estimate crash frequencies on freeways, using aggregate measures of traffic-flow variables. However, they determined the crash precursors included in the model in an objective manner and did not base them on their subjective categorization. In a related study (Lee, Hellinga, and Saccomanno (2004), they proposed the application of the models and estimated real-time crash potential. The main focus of this study was on reducing the crash potential obtained from the model through different control strategies of variable speed limits (VSLs). To mimic responses of the drivers to changes in speed limits, they used the microscopic simulation tool PARAMICS. At least on the simulated data, the VSLs showed significant safety benefits in terms of estimated reduction in crash potential.

Gayah et al. (2006) similarly used PARAMICS to assess the effectiveness of various ITS strategies in mitigating crash-prone conditions on the previously studied Interstate 4 corridor in Orlando, FL. They also concluded that VSL significantly reduced the potential for crashes with high-speed conditions preceding them, but that such a benefit could be achieved only by ramp metering in the congested regime.

Continuing this trend of investigating advanced traffic management (ATM) strategies, Nezamuddin et al. (2011) used VISSIM to model VSL, peak-period shoulder-lane use, and both strategies together. Their study assessed the effects of these strategies on speed, throughput, and safety on a section of the Missouri-Pacific Expressway in Austin, TX. Speed harmonization and a reduction in the number of stops per vehicle and vehicle conflicts were achieved with VSL; however, this came at the expense of operating speed. Shoulder use increased operating speed and decreased traffic density but also increased speed variability and had many other safety considerations that must be addressed. Ramp metering was not considered in this study.

In a similar study, weather, environmental, and loop-detector data were analyzed for association with different incident types (Songchitruksa and Balke, 2006). It was found that five-minute average occupancy and coefficient of variation in speed had the strongest association with crash risk, and other factors such as visibility, time of day, and lighting condition strongly affected the type of incident that occurred.

A study by Pande, Abdel-Aty, and Hsia (2005) utilized within-stratum one-covariate logistic-regression models to determine the relative risk of crash occurrence, measured by the hazard ratio. This ratio represents the increase in the risk of crash occurrence (in log odds) resulting from changing the covariate by one unit. The study found that the log of the coefficient of variation in speed and average occupancy (expressed as a percentage) and standard deviation of volume most significantly affected the likelihood of crash occurrence. Additionally, it determined that computing these parameters at five-minute time intervals was more closely associated with crash risk than computing them at three-minute intervals. Contour plots of spatiotemporal variation of crash risk were created, and the one representing the log of the coefficient of variation in speed most clearly demonstrated increasing crash risk as the time and location of the crash were approached. The authors also proposed a methodology for identifying crash-prone conditions in real-time, for potential use in proactive traffic management.

Oh et al. (2001) showed that five-minute standard deviation of 30-second speed measurements was the best indicator of "disruptive" traffic flow leading to a crash, as opposed to "normal" traffic flow. They used the Bayesian classifier to categorize the two possible traffic-flow conditions. Since the Bayesian classifier requires a probability-distribution function for each competing class, the standard deviations of speed over crash and non-crash cases were used to fit non-parametric distribution functions, using Kernel smoothing techniques. The potential application of the model in real-time was also demonstrated.

A more detailed analysis of patterns in crash characteristics as a function of real-time traffic flow was performed by Golob and Recker (2003). The methodology used was non-linear (non-parametric) canonical correlation analysis (NLCCA) with three sets of variables. The first set comprised a seven-category segmentation variable defining lighting and weather conditions; the second set was made up of crash characteristics (collision type, location, and severity); the third set consisted of real-time traffic-flow variables. NLCCA requires reducing collinearity in the data, so a principal component analysis (PCA) was performed

to identify relatively independent measurements of traffic-flow conditions. The results of the PCA are shown in Table 1.

Table 1.   Interpretation of Principal Components and Variable Selection

Factor	Interpretation	Represented by
1	Central tendency of speed	Median volume/occupancy, interior lane
2	Central tendency of volume	Mean volume, left lane
3	Temporal variation in volume, left and interior lanes	Variation in volume, left lane
4	Temporal variation in speed, left and interior lanes	Variation in volume/occupancy, interior lane
5	Temporal variation in speed, right lane	Variation in volume/occupancy, right lane
6	Temporal variation in volume, right lane	Variation in volume, right lane

SOURCE: Golob and Recker (2003).

Golob and Recker concluded that collision type is the best-explained characteristic and is related to the median speed and left-lane and interior-lane variations in speed. Moreover, the severity of the crash tracks the inverse of the traffic volume and is influenced more by volume than by speed.

In a later study, Golob, Recker, and Alvarez (2004a) used data for more than 1,000 crashes over six major freeways in Orange County, CA, and developed a software tool called Flow Impacts on Traffic Safety (FITS) to forecast the types of crashes that are most likely to occur for the flow conditions being monitored. A case-study application of this tool on a section of State Route (SR) 55 was also demonstrated.

Golob and Recker (2004) showed that certain traffic-flow regimes are more conducive to traffic crashes than others. They found that of the eight traffic-flow regimes that exist on the six freeways in Orange County, CA, nearly 76% of all crashes occurred in the four traffic regimes that represent flow nearing or at congestion. This displays a correlation between the types of flow and crashes and indicates that understanding the patterns in real-time traffic flow might be the key to "predicting" crashes on urban freeways. It should be noted that none of these studies included non-crash-loop data as a measure of "normal" traffic conditions.

The link between traffic congestion and freeway crashes was also noted by Zhang et al. (2005) in a study that explored the relationship between crashes, weather conditions, and traffic congestion. The study showed that the relationship between the "relative risk ratio" (a measure of crash probability) resembles an inverted U-shaped curve with a peak value during moderate congestion and low points at free flow and heavy congestion.

Park and Ritchie (2004) showed that lane-changing behavior and the presence of long vehicles within a freeway section have a significant impact on section-speed variability. Section-speed variance rather than point-speed variance was used to demonstrate the traffic changes more efficiently. The traffic data for their study were not obtained from conventional single- or dual-loop detectors. Instead, a state-of-the-art vehicle-signature-based traffic-monitoring technology that provided individual vehicle trajectories as well as accurate vehicle classification was used.

Pande and Abdel-Aty (2006) further correlated lane-changing maneuvers with both sideswipe and angle crashes on the inner lanes of a freeway. Classification trees using data collected from loop detectors on the Interstate 4 corridor identified average speed upstream and downstream of the crash location and difference in occupancy of adjacent lanes as having significant association with the crash/non-crash binary variable. Satisfactory classification accuracy indicated the potential for real-time application in identifying risk for lane-change-related crashes.

Another study by Pande and Abdel-Aty (2006a) analyzed rear-end crashes occurring under two flow regimes: extended congestion and near-free-flow five to 10 minutes prior to a crash. In the first case, the coefficient of variation in speed and average occupancy distinguished crashes from randomly selected non-crash cases. In the second case (nearly free-flow conditions preceding a crash), average speed and occupancy downstream of the crash location were identified as significant factors. The authors proposed a strategy for real-time identification of crash-prone conditions, using neural-network-based classifiers.

While almost all studies have indicated a relationship between crash occurrence and speed variability, Kockelman and Ma (2004) found no evidence that speeds or their variations trigger crashes. The study was conducted for the same location as that used by Golob, Recker, and Alvarez (2004b). The sample size was limited to 55 severe crashes that occurred during January 1998, and with such a small sample, their conclusions are suspect. Similarly, Ishak and Alecsandru (2005) were unable to separate pre-incident, post-incident, and non-incident traffic regimes from one another, which indicated that conditions before a crash might not be discernible in real-time. The study was performed using part of the ITS-archived data from Interstate 4 in Orlando, FL, that was used by Pande (2003). However, data for only 116 crashes were used, which raises concerns about the validity of the findings from this research.

Various modeling methodologies have been explored by researchers, including probabilistic neural network (PNN) (Abdel-Aty and Pande, 2005), matched case-control logistic regression (Abdel-Aty et al., 2004), split models (Abdel-Aty, Uddin, and Pande, 2005), multilayer perceptron (MLP)/radial basis function (RBF) neural-network architectures (Pande, 2003), and generalized estimation equation (Abdel-Aty and Abdalla, 2004). The data for these studies were collected from a 13.2-mile central corridor of Interstate 4 in Orlando. All the studies made significant contributions toward enriching the literature. However, there remains considerable scope for improvement.

Critical Review

It is evident that exploring the loop data in traffic-safety research is still in its preliminary stages. Some of the aforementioned studies do have potential application in real-time proactive traffic management, but they have not fully analyzed the "recipe" of crashes. In addition, the statistical analysis in some cases is not really sound, from a theoretical point of view.

Lee, Hellinga, and Saccomanno (2003) have an advantage over other research groups—the availability of dual loops placed close to each other (38 loops on a 10-km stretch of freeway in Canada). Their analysis is based on a log-linear crash frequency model. As this model is not based on classification, it cannot decipher whether or not conditions are risky in real-time and is therefore unsuitable for real-time classification of the loop-data patterns.

Golob and Recker (2003) have established sound statistical links between environmental factors, traffic flow as obtained from loop data, and crash occurrence, but their findings are limited by the fact that the traffic data are obtained from single-loop detectors, and speed has to be estimated using a proportional variable (volume/occupancy). FITS has limited application due to a systematic pattern of missing values within the data used in its development.

The classification model developed by Oh et al. (2001) seems to have the most promising online application, but because of limited crash data (only 52 crashes), it remains far from being implemented in the field. The only factor used for classification is the five-minute standard deviation of speed; other significant factors such as geometry, weather, and other traffic-flow variables are not considered. It is also to be understood that for a crash-prediction model to be useful, it is necessary to classify the data much ahead of the crash occurrence time and not just five minutes prior to the crash to provide the Regional Transportation Management Center (RTMC) with some time for analysis, prediction, and dissemination of the information.

The use of limited crash and traffic data causes concerns about the findings of Ishak and Alecsandru (2005) as well. In their study, pre-incident, post-incident, and non-incident traffic-flow regimes are described by 30-second average speed and its variation depicted through spatiotemporal contour charts. Using second-order statistical analyses, Ishak and Alecsandru measured the charts for smoothness, homogeneity, and randomness. No consistent pattern for any of the statistical measures was found within the three categories of traffic regime. Therefore, it was concluded that conditions belonging to these regimes could not be differentiated based on loop data. However, only 116 crashes were used in the analysis, with speed and its variation as the only independent parameters. It is likely that more crash and non-crash data, along with different flow parameters from a range of stations located around crash locations, would have yielded better results. The findings from previous studies by Abdel-Aty et al. (differentiating pre-crash from non-crash) and Al-Deek et al. (separating post-incident from non-incident) used the loop data from the same corridor, making this postulation all the more plausible.

Investigators deem the most critical issue not addressed by past research to be the issue of transferability. Since gathering data from different sources and combining them is a significant effort, it would be worthwhile to know whether models developed from one freeway can be applied to the data from other freeways. While it may be unreasonable for models developed with data from a dense urban freeway environment to perform well on a rural freeway corridor, no studies have tested even models from the same geographical area to other freeways in close proximity. This study makes an effort in that direction.

APPLICATIONS OF DATA MINING IN TRANSPORTATION

Data mining is defined as the process of extracting valid, previously unknown, and ultimately comprehensive information from large databases (Hand, Mannila, and Smyth, 2001). Over the years, data mining has emerged as a powerful instrument offering value across a broad spectrum of information-intensive industries, including banking and logistics. The potential of various data mining techniques in the field of transportation engineering, however, remains underutilized, with the exception of neural-network applications for incident detection.

The incident-detection algorithms are the most relevant data mining applications for this research problem, since detecting an incident also involves classification of traffic-flow patterns emanating from loop detectors. The critical distinction is that while we are interested in pre-crash data, detection algorithms involve analysis of post-incident loop data. The following section reviews data mining-based incident-detection algorithms.

Incident-Detection Algorithms

Cheu and Ritchie (1995) developed three types of neural-network models to classify traffic data obtained from loop detectors: multilayer feed forward (MLF), self-organizing feature map (SOFM), and adaptive resonance theory 2 (ART2). Their objective was to use the classified output to detect lane-blocking freeway incidents.

Artificial neural network models (ANNs) were designed to classify the input data into one of the two states: an incident or an incident-free condition. ANNs were trained using post-incident loop-detector data generated from Integrated Traffic Simulator (INTRAS), a microscopic traffic-simulation model, because, according to the authors, it would have been impractical to put extensive effort into collecting real-life data. INTRAS initially generated the incident and incident-free input vectors in a ratio of 1:4. The incident input vectors were later replicated to make the number of state 1 and state 2 vectors equal in the training dataset. The input vectors used were 16-dimensional, consisting of upstream and downstream detectors' volume and occupancy at 30-second slices after the time of the incident. The performance of these networks on field evaluation data indicated that MLP neural networks produce consistently better results than the other two networks; the results were also better than those obtained by the traditional detection algorithms.

Abdulhai and Ritchie (1999) tried to identify the requirements of a successful detection framework and found that inability to address the issues of predicted probability of incident occurrence is a major shortcoming of detection algorithms. They proposed the concept

of using statistical distance and a modified probabilistic neural-network model (PNN2) in addition to a Bayesian-based traditional PNN model to detect the patterns in the loop data. They also reported that these two models were competitive with the more frequently used MLP neural networks for incident detection.

Ishak and Al-Deek (1999) conducted a study that did not use simulation data and training and testing of the neural-network models for incident detection; rather, it used real-life loop data only. In this regard, studies by Al-Deek, Ishak, and Khan (1996) and Al-Deek, Garib, and Radwan (1998) on incident detection are remarkable. The data used by Ishak and Al-Deek (1999) were collected from the same Interstate 4 corridor for which the initial crash-prediction models were developed by Pande and Abdel-Aty (2008). Input patterns of various dimensions were attempted, and the network size was changed accordingly to achieve better performance. They found that when using the MLF neural network, the incidents might be detected better with the speed patterns alone than with occupancy patterns or a combination of speed-occupancy patterns.

Data Mining Applications in Traffic Safety

Sayed and Abdelwahab (1998) compared the fuzzy K-nearest neighbor algorithm and the MLP neural network for identifying crash-prone locations. Results showed that the MLP produced slightly more accurate results and achieved higher computational efficiency than fuzzy classification.

Awad and Janson (1998) applied an MLP to model truck crashes at interchanges in the state of Washington. Results of the neural-network model were compared with those from a linear-regression model. The comparison was based on the root mean squared error (RMSE). The trained neural network showed a better fit when the training data are presented. However, the ability of the trained ANN to predict "unseen" test data was unsatisfactory.

Mussone, Ferrari, and Oneta (1999) used an MLP approach to analyze traffic crashes that occurred at intersections in Milan, Italy. Results showed that the neural-network models could extract information such as factors explaining crashes and contributing to a higher degree of danger.

Through a sequential review of the literature, we observed that the only neural-network architecture explored for traffic-safety analysis was the MLP until Abdelwahab and Abdel-Aty (2001) developed Fuzzy Adaptive Resonance Theory (Fuzzy ART) neural networks to predict driver-injury severity in traffic crashes at signalized intersections. These models were compared with the MLP architecture and it was concluded that MLP models were superior to the ordered logit model and Fuzzy ART. In a later study (Abdelwahab and Abdel-Aty, 2002), ANN models were used for traffic-safety analysis of toll plazas. Driver-injury severity (no injury, possible injury, evident injury, severe injury/fatality) and location of the crash (before the plaza, at the plaza, past the plaza) were analyzed using MLP as well as a radial basis function (RBF) neural network. Abdelwahab and Abdel-Aty (2002) reported that the nested logit model was the best model for analyzing crash location, while the RBF neural network was the best model for analysis of driver-injury severity.

Pande and Abdel-Aty (2008) explored PNNs, an implementation of the Bayesian classifier, on the Interstate 4 corridor in Orlando to identify rear-end-crash-prone conditions. These crashes were divided into those occurring under (1) congested and (2) relatively free-flow conditions preceding the crash, and decision tree-based classification determined that while their frequencies are comparable, the first condition is much rarer and can hence be described as a "crash-prone" condition. PNN-based classification models were also developed for the free-flow regime.

Data mining techniques other than neural networks have also appeared in recent traffic-safety literature. Vorko and Jovic (2000) used multiple-attribute entropy models to classify injuries of school-age passengers. Sohn and Shin (2001) employed neural networks and decision-tree algorithms to develop classification models for road-traffic crash severity (bodily injury or property damage) as a function of potentially correlated categorical factors. They noted that classification accuracy of the individual models from both algorithms was relatively low, and the use of data-fusion or ensemble algorithms increased the classification accuracy. Data-fusion techniques combine classification results obtained from individual classifiers and are known to improve classification accuracy when some results of relatively uncorrelated classifiers are combined. The resulting performance is usually more stable than that of a single classifier.

Pande and Abdel-Aty (2007) proposed a multiple-model framework incorporating the findings of earlier studies on rear-end and lane-change-related crashes on the Interstate 4 corridor in Orlando. The models satisfactorily identified both of these cases, as well as related single-vehicle crashes. This work elaborates on Pande's doctoral dissertation (Pande, 2005), which attempted to identify the unique precursors to each crash type and develop models that can be hybridized and applied in real-time as part of a proactive traffic-management strategy.

A study conducted by Xu et al. (2011) on a 9.2-mile stretch of the I-880 corridor in Hayward, CA, used loop-detector data gathered by researchers at the University of California, Berkeley. The researchers classified traffic into five homogeneous flow states using K-means clustering analysis to compare occupancy data for a one-crash case with four non-crash cases, all occurring at the same time and location between loop detectors. The researchers developed four logistic-regression models, indicating odds ratios four to five times higher for the "risky" scenarios of free flow upstream to a congested downstream regime and congested upstream flow to free flow downstream, and an odds ratio two times higher for flow in the transition region between uncongested and congested flow, when compared with the base case of free flow. The crash risk in the case of congested, homogeneous flow was not statistically different from that in the case of free flow. Discriminant functions developed using linear combinations of the lane-occupancy variables were able to correctly categorize the type of flow with 97.2% accuracy, and they can be deployed in real-time.

Pham, El Faouzi, and Dumont (2011) considered not only the speed and variability in speed as explanatory variables for crash risk, but also meteorological conditions (namely precipitation). Focusing on a 10-km stretch of the A1 motorway near Bern, Switzerland, between 2002 and 2007, the authors analyzed 120 rear-end and sideswipe crashes.

Data were collected for 30 minutes before each crash (in five-minute intervals), as well as for non-crash cases. PCA was used to normalize and transform traffic situations to self-organizing maps (SOMs), which partition the data points into clusters. Random Forests™ was then used to develop risk-identification models for each of eight defined flow regimes. Six of the regimes performed with acceptable accuracy (70% of crash and non-crash cases correctly identified). The two that performed poorly did not have enough data to develop a good statistical model. It was found that rain had a much stronger influence in medium-flow regimes than under either congested or free-flow conditions. For most of the traffic regimes, lane speed and lane variation in speed were the most significant factors in determining crash risk.

CONCLUSIONS FROM THE LITERATURE REVIEW

The findings of an extensive review of relevant literature demonstrate the applications, albeit limited, of ITS archived data and data mining techniques in the field of traffic safety.

The issues not addressed adequately by studies using real-time loop-detector data for predicting crashes are referred to by Golob, Recker, and Alvarez (2004b) as disaggregate studies. The most significant of these issues in the research reported here is that of transferability. Therefore, a sufficiently large database of crash and non-crash data was assembled for this study from a subset of the major freeways/expressways in the city of San José. The models developed from U.S. 101 NB data were applied to the other three corridors for which data were assembled. PeMS, managed by Caltrans, provided the archived ITS data (collected and stored on a continuous basis), as well as the incident data. Chapter III discusses these data sources and the details of the four corridors in the context of the present research problem.

III. STUDY AREA

This study examines four freeway segments: U.S. 101 NB and SB and I-880 NB and SB in the San José area of Santa Clara County, CA. These freeway corridors run through dense urban development and are among the busiest in the South Bay area. The logistic-regression and data mining models are estimated using the U.S. 101 NB data, and the models are applied on U.S. 101 SB, I-880 NB, and I-880 SB to evaluate their transferability. This chapter provides details of these segments and describes the process of data collection and preparation.

FREEWAY CORRIDORS

U.S. 101

U.S. 101 (also known as the Bayshore Freeway) is the primary north-south corridor through the city of San José. The route runs as a six-lane freeway through the suburbs of Gilroy and Morgan Hill in southern Santa Clara County. North of Morgan Hill, U.S. 101 gains a high-occupancy vehicle (HOV) lane in each direction (expanding to an eight-lane freeway) through the rural area known as Coyote. The freeway wanders in and out of the San José city limits and unincorporated land for approximately eight miles. At the junction of SR 85, U.S. 101 enters the area conventionally accepted as the boundary of the city of San José. The route continues as an eight-lane freeway through the junctions of SR-82, I-280/I-680, I-880, and SR-87, then enters the city of Santa Clara. It continues through the South Bay cities of Sunnyvale, Mountain View, and Palo Alto, finally running up the Peninsula through San Mateo County to San Francisco.

The study segment of interest for U.S. 101 NB is 17.1 miles long, starting at milepost 375.31 and ending at milepost 392.37. The study segment for U.S. 101 SB starts at milepost 392.45 and ends at milepost 375.81, for a total length of 16.6 miles. Figure 1 and Figure 2 present schematic diagrams of the VDSs along these routes. In the diagrams, VDS identification (ID) numbers are truncated to the last four digits and superimposed on the route.

Study Area

Figure 1.   U.S. 101 NB Corridor and VDS Locations

Figure 2.   U.S. 101 SB Corridor and VDS Locations

I-880

I-880 (also known as the Nimitz Freeway) is a six-lane freeway with no dedicated HOV lanes. Its officially designated beginning is north of the I-280 junction. The freeway extends south of this interchange as SR 17, running between Santa Cruz, CA, and San José, CA. I-880 runs north through the city of San José for approximately seven miles, connecting to SR 82, crossing over SR 87 (with no interchange) and connecting to U.S. 101. I-880 next enters the city of Milpitas and finally crosses the Alameda County line, running up the East Bay to Oakland. An improvement project has been under way since 2010 to reconfigure the I-280/I-880 interchange. The goal is to provide a dedicated ramp from

I-280 NB to I-880 NB; the connection is currently shared with the busy Stevens Creek Boulevard interchange, causing merging and weaving issues.

The study segment of interest for I-880 NB is 8.1 miles long, starting at milepost 0.13 and ending at milepost 8.27. The segment for I-880 SB starts at milepost 9.01 and ends at milepost 0.9, for a total length of 8.1 miles. Figure 3 and Figure 4 present schematic diagrams of the routes, along with VDS locations. The study location is shown in Figure 5.
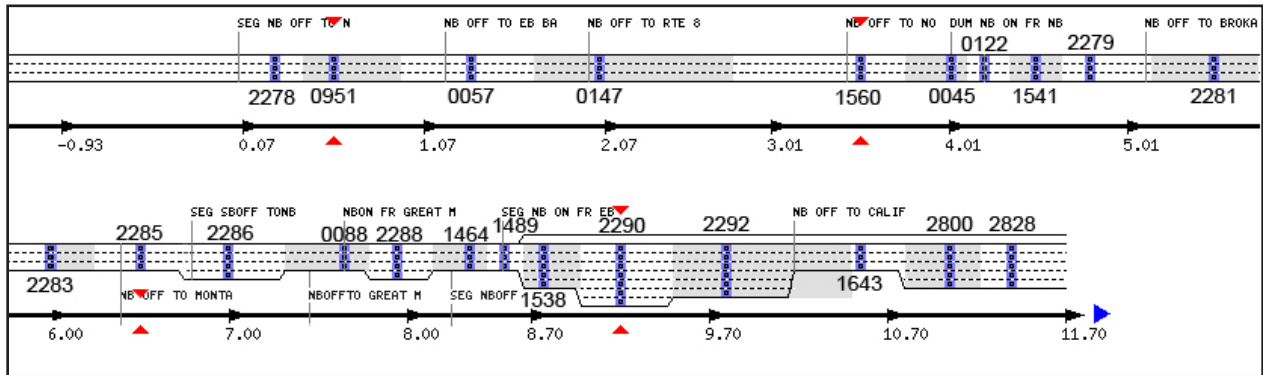
Figure 3.   I-880 NB Corridor and VDS Locations
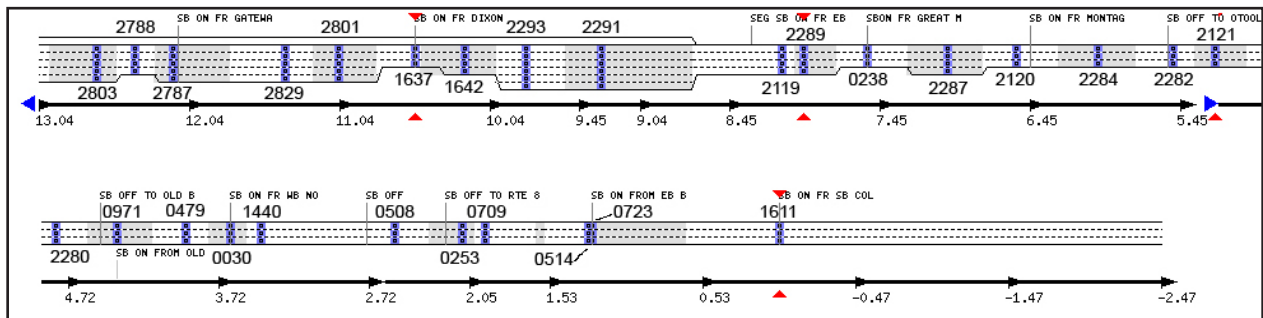
Figure 4.   I-880 SB Corridor and VDS Locations
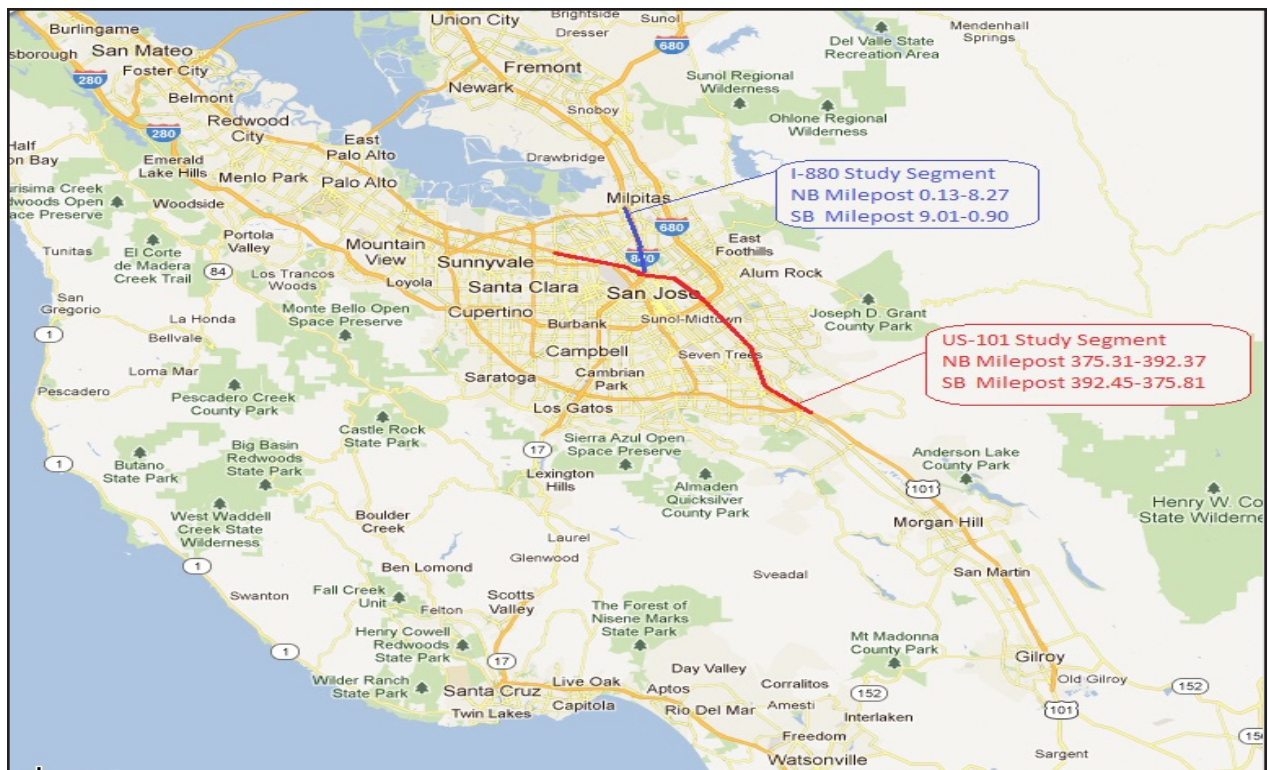
Figure 5.   Study Location

DATA COLLECTION AND PREPARATION

Crash Data

The study considers crashes that occurred during a 16-month period from January 2010 through April 2011. This time period was chosen because new loop detectors had been installed on U.S. 101 in 2009. Crash data were downloaded from the California Highway Patrol (CHP) Incidents section of Caltrans' PeMS database. Figure 6 shows a sample of the downloaded data. Important variables for our analysis include each incident's unique ID number, time of occurrence, and milepost.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	#	District	Area	Fwy	Start	Duration	Abs Postmile	CA Postmi	Location	Description				
2	472	4	SAN JOSE	SR101-N	1/1/2010 2:45	126	379.56	29.931	NB HELLYER AV ONR TO NB US101	1183 - Traffic Collision - No Details				
3	971	4	SAN JOSE	SR101-N	1/1/2010 9:27	150	383.55	34.113	NB US101 JSO NB I680	1183 - Traffic Collision - No Details				
4	980	4	SAN JOSE	SR101-N	1/1/2010 9:35	0	382.42	32.784	NB US101 AT TULLY RD	1183 - Traffic Collision - No Details				
5	304	4	SAN JOSE	SR101-N	1/5/2010 7:13	40	387.04	37.603	NB US101 JSO N 13TH ST	1183 - Traffic Collision - No Details				
6	419	4	SAN JOSE	SR101-N	1/5/2010 8:05	4	385.22	R35.782	NB US101 JSO E JULIAN ST	1182 - Traffic Collision - Property Damage				
7	1218	4	SAN JOSE	SR101-N	1/5/2010 14:49	53	387.04	37.603	NB US101 JSO N 13TH ST	20002 - Hit and Run - No Injuries				
8	1951	4	SAN JOSE	SR101-N	1/6/2010 19:08	12	391.5	41.862	NB US101 AT MONTAGUE EXWY	1182 - Traffic Collision - Property Damage				
9	1900	4	REDWOOL	SR101-N	1/7/2010 18:33	5	375.31	R26.237	NB US101 JSO SR85	1183 - Traffic Collision - No Details				
10	295	4	HOLLISTEF	SR101-N	1/8/2010 13:38	0	376.32	R26.85	NB US101 JNO BERNAL RD	1183 - Traffic Collision - No Details				
11	1106	4	SAN JOSE	SR101-N	1/8/2010 13:42	11	376.32	R26.85	NB US101 JNO BERNAL RD	1183 - Traffic Collision - No Details				
12	355	4	GOLDEN G	SR101-N	1/9/2010 5:00	0	377.26	R28.185	NB US101 JSO BLOSSOM HILL RD	1183 - Traffic Collision - No Details				
13	362	4	SAN JOSE	SR101-N	1/9/2010 5:13	2	379.36	29.931	NB US101 JSO HELLYER AV	1183 - Traffic Collision - No Details				
14	1722	4	SAN JOSE	SR101-N	1/9/2010 18:16	24	387.6	38.164	NB US101 JSO NB I880	20002 - Hit and Run - No Injuries				
15	1950	4	REDWOOL	SR101-N	1/11/2010 18:38	0	375.31	R26.237	NB US101 JSO SR85	1183 - Traffic Collision - No Details				
16	139	4	SAN JOSE	SR101-N	1/12/2010 3:57	29	387.6	38.164	NB US101 JSO NB I880	1182 - Traffic Collision - Property Damage				
17	2026	4	SAN JOSE	SR101-N	1/12/2010 20:04	1	383.75	34.113	NB US101 TO NB I680 CON	1183 - Traffic Collision - No Details				
18	2045	4	SAN JOSE	SR101-N	1/12/2010 20:16	0	383.75	34.113	NB US101 TO NB I680 CON	1183 - Traffic Collision - No Details				
19	2059	4	SAN JOSE	SR101-N	1/12/2010 20:27	27	382.22	32.784	NB US101 JSO TULLY RD	1182 - Traffic Collision - Property Damage				
20	2062	4	SAN JOSE	SR101-N	1/12/2010 20:29	1	380.77	31.133	NB US101 AT CAPITOL EXWY	1183 - Traffic Collision - No Details				
21	2241	4	SAN JOSE	SR101-N	1/12/2010 22:38	57	382.42	32.784	NB US101 ON EB TULLY RD OFR	1179 - Traffic Collision - Ambulance Responding				
22	2256	4	SAN JOSE	SR101-N	1/12/2010 22:52	66	380.77	31.133	NB US101 ON NB CAPITOL EXWY C	1183 - Traffic Collision - No Details				
23	2260	4	SAN JOSE	SR101-N	1/12/2010 22:55	45	382.42	32.784	NB US101 AT TULLY RD	1183 - Traffic Collision - No Details				
24	2264	4	SAN JOSE	SR101-N	1/12/2010 22:58	0	382.42	32.784	NB US101 AT TULLY RD	1179 - Traffic Collision - Ambulance Responding				
25	2310	4	SAN JOSE	SR101-N	1/12/2010 23:28	0	380.77	31.133	NB US101 AT CAPITOL EXWY	1183 - Traffic Collision - No Details				
26	2319	4	SAN JOSE	SR101-N	1/12/2010 23:37	0	382.42	32.784	NB US101 ON EB TULLY RD OFR	1182 - Traffic Collision - Property Damage				
27	464	4	SAN JOSE	SR101-N	1/13/2010 8:19	13	392.17	42.533	NB US101 AT GREAT AMERICA PKV	1182 - Traffic Collision - Property Damage				
28	651	4	SAN JOSE	SR101-N	1/13/2010 9:33	67	390.43	40.591	NB US101 JNO DE LA CRUZ BLVD	1183 - Traffic Collision - No Details				
29	1622	4	GOLDEN G	SR101-N	1/13/2010 16:50	0	392.17	42.533	GREAT AMERICA PKWY AT NB US1	1183 - Traffic Collision - No Details				
30	2239	4	SAN JOSE	SR101-N	1/14/2010 19:30	0	380.77	31.133	NB US101 AT CAPITOL EXWY	20002 - Hit and Run - No Injuries				
31	648	4	SAN JOSE	SR101-N	1/15/2010 9:35	12	380.77	31.133	NB CAPITOL EXWY ONR TO NB US:	1183 - Traffic Collision - No Details				
32	796	4	GOLDEN G	SR101-N	1/16/2010 11:12	1	377.26	R28.185	NB US101 JSO BLOSSOM HILL RD	1179 - Traffic Collision - Ambulance Responding				
33	1237	4	SAN JOSE	SR101-N	1/16/2010 14:36	0	390.43	40.591	NB US101 JNO DE LA CRUZ BLVD	1183 - Traffic Collision - No Details				
34	385	4	SAN JOSE	SR101-N	1/17/2010 4:52	28	377.26	R28.185	NB US101 JSO BLOSSOM HILL RD	1183 - Traffic Collision - No Details				
35	819	4	SAN JOSE	SR101-N	1/17/2010 12:19	34	383.86	34.224	NB STORY RD ONR TO NB US101	1179 - Traffic Collision - Ambulance Responding				
36	835	4	SAN JOSE	SR101-N	1/17/2010 12:36	59	383.55	34.113	NB US101 JSO NB I680	1183 - Traffic Collision - No Details				
37	846	4	SAN JOSE	SR101-N	1/17/2010 12:41	0	382.42	32.784	NB US101 AT TULLY RD	1183 - Traffic Collision - No Details				
38	876	4	SAN JOSE	SR101-N	1/17/2010 12:55	9	385.62	R35.782	NB US101 JNO MCKEE RD	1179 - Traffic Collision - Ambulance Responding				
39	889	4	SAN JOSE	SR101-N	1/17/2010 13:03	0	383.66	34.224	NB US101 JSO STORY RD	1183 - Traffic Collision - No Details				

Figure 6. Crash Data from PeMS

SOURCE: PeMS database.

The predictive models were developed from the crash data from U.S. 101 NB. There were 2,176 crashes during the study period, the type, number, and percentage of which are shown in Table 2.

Table 2. Crash Details for U.S. 101 NB

Crash Type	Number	Percentage of Total Crashes
1181 - Traffic collision, minor injuries	38	1.7
1182 - Traffic collision, property damage	754	34.7
1179 - Traffic collision, ambulance responding	257	11.8
1144 - Possible fatality	2	0.1
20002 - Hit and run, no injuries	182	8.4
20001 - Hit and run, injuries or fatalities	5	0.2
1183 - Traffic collision, no details	938	43.1
Total	**2,176**	**100.0**

Traffic Information

Once the crash data were obtained, a list of all VDS locations on the study segments was compiled along with their respective mileposts. A sample list is shown in Figure 7. The variables of interest for this study include the VDS number and the milepost.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Fwy	Dist	County	City	CA PM	Abs PM	VDS	ID	Lane	Type	HOV	LDS
3	SR101-S	4	Santa Clara		16.4	365.71	402336	412033	1	Mainline	No	403810
4					16.4	365.71	402336	412034	2	Mainline	No	403810
5					16.4	365.71	402336	412035	3	Mainline	No	403810
6					16.4	365.71	402336	412036	4	Mainline	No	403810
7				Morgan Hill	16.7	366.01	401579	409582	1	Mainline	No	403156
8					16.7	366.01	401579	409583	2	Mainline	No	403156
9					16.7	366.01	401579	409584	3	Mainline	No	403156
10				SB ON FR EB COCHRAN RD (R17.576)								
11				SB ON FR WB COCHRAN RD (R17.857)								
12	SR101-S	4	Santa Clara	Morgan Hill	17.89	367.2	401668	409824	1	Mainline	No	403286
13					17.89	367.2	401668	409825	2	Mainline	No	403286
14					17.89	367.2	401668	409826	3	Mainline	No	403286
15				SB OFF TO COCHRAN RD (R18.145)								
16	SR101-S	4	Santa Clara		18.8	368.11	402338	412041	1	Mainline	No	403812
17					18.8	368.11	402338	412042	2	Mainline	No	403812
18					18.8	368.11	402338	412043	3	Mainline	No	403812
19					18.8	368.11	402338	412044	4	Mainline	No	403812
20				San Jose	19.5	368.81	402340	412049	1	Mainline	No	403814
21					19.5	368.81	402340	412050	2	Mainline	No	403814
22					19.5	368.81	402340	412051	3	Mainline	No	403814
23					19.5	368.81	402340	412052	4	Mainline	No	403814
24					20.3	369.61	402342	412057	1	Mainline	No	403816
25					20.3	369.61	402342	412058	2	Mainline	No	403816
26					20.3	369.61	402342	412059	3	Mainline	No	403816
27					20.3	369.61	402342	412060	4	Mainline	No	403816
28				SB ON FR SCHELLER AVE (R21.05)								
29				SB OFF TO SCHELLER AVE (R21.51)								
30	SR101-S	4	Santa Clara	San Jose	21.8	371.11	402344	412065	1	Mainline	No	403818
31					21.8	371.11	402344	412066	2	Mainline	No	403818
32					21.8	371.11	402344	412067	3	Mainline	No	403818
33					21.8	371.11	402344	412068	4	Mainline	No	403818
34					22.29	371.6	401661	409802	1	Mainline	No	403279
35					22.29	371.6	401661	409803	2	Mainline	No	403279
36					22.29	371.6	401661	409804	3	Mainline	No	403279
37					22.29	371.6	401661	409805	4	Mainline	No	403279
38					23.29	372.6	400223	402377	1	Mainline	No	402861
39					23.29	372.6	400223	402378	2	Mainline	No	402861
40					23.29	372.6	400223	402379	3	Mainline	No	402861
41					23.29	372.6	400223	402380	4	Mainline	No	402861

Figure 7.   VDS Locations, by Milepost

Traffic data from these VDS locations were downloaded from the Data Clearinghouse section of PeMS for the entirety of Caltrans District 4 (Bay Area). The following variables for each VDS were included: time and date, milepost and average speed, volume, and lane-occupancy information measured every 30 seconds, by corresponding VDS. Among these variables, only volume and lane occupancy are measured, and the 30-second average speed is calculated (in the database) using these measurements. Figure 8 shows a sample of the downloaded raw loop-detector data.

We next matched the traffic data to the corresponding crash events. The crash time and locations were known from the database (see sample in Figure 6), and each crash was merged with corresponding traffic data from six VDS locations—the three VDSs nearest to the location of the crash in the upstream direction and the nearest three in the downstream direction. The spatial arrangement of locations is shown later in this chapter, in Figure 11.

VDSs were typically spaced 0.5 to 0.8 mile apart. The time horizon for each event was the period up to 20 minutes before the crash and five minutes after the crash. The period of up to five minutes after the crash was used only to verify the incident's occurrence (and is typically relevant only for incident detection); it will therefore not be discussed further in this report.

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10	VAR11	VAR12	VAR13
1	01JAN10:00:00:18	400223	1	0.0078	78	1	0.0089	86	1	0.0078	78	0	0
2	01JAN10:00:00:18	400237	2	0.01	86	2	0.0117	78	3	0.0206	71	3	0.045
3	01JAN10:00:00:18	401178	0	0	149	2	0.01	149	4	0.01	149	2	0.01
4	01JAN10:00:00:18	401464	1	0.0044	86	2	0.0139	65	3	0.0106	49		
5	01JAN10:00:00:18	401489	1	0.0067	81	1	0.005	96	0	0	0		
6	01JAN10:00:00:18	401538	0	0	0	0	0	0	0	0	0	2	0.0133
7	01JAN10:00:00:18	401642	2	0.0122	78	2	0.0139	71	3	0.0189	71	2	0.0111
8	01JAN10:00:00:18	401658	1	0.0078	71	1	0.005	86	3	0.0139	65	0	0
9	01JAN10:00:00:18	401660	1	0.0094	78	1	0.0078	78	0	0	0		
10	01JAN10:00:00:18	401661	2	0.0172	78	1	0.0078	78	0	0	0	1	0.0117
11	01JAN10:00:00:18	401662	0	0	0	0	0	0	1	0.0094	78	1	0.0106
12	01JAN10:00:00:18	401663	0	0	0	0	0	0	2	0.0167	78	0	0
13	01JAN10:00:00:18	401668	1	0.0061	86	4	0.0283	78	1	0.0078	86		
14	01JAN10:00:00:18	401669	0	0	0	1	0.0078	71	2	0.0122	78	0	0
15	01JAN10:00:00:18	401670	0	0	0	0	0	0	1	0.0056	97	0	0
16	01JAN10:00:00:18	401906	0	0	0	0	0	0	1	0.0072	78	1	0.0078
17	01JAN10:00:00:18	401907	0	0	0	4	0.0311	65	0	0	0		
18	01JAN10:00:00:18	401908	3	0.0189	71	1	0.0056	65	3	0.0194	84	2	0.0133
19	01JAN10:00:00:18	401909	0	0	0	0	0	0	0	0	0		
20	01JAN10:00:00:18	401995	0	0	0	1	0.0044	86	1	0.0044	86	0	0
21	01JAN10:00:00:18	402119	0	0	0	0	0	0	0	0	0	0	0
22	01JAN10:00:00:18	402120	0	0	0	2	0.015	96	1	0.0128	109		
23	01JAN10:00:00:18	402121	2	0.0128	78	4	0.0194	71	0	0	0		
24	01JAN10:00:00:21	400001	3	0.0222	78	0	0	0	1	0.0067	71	1	0.0078
25	01JAN10:00:00:21	400014	0	0	0	1	0.0061	71	1	0.0067	9	1	0.0072
26	01JAN10:00:00:21	400030	0	0	0	0	0	0	1	0.0089	65		
27	01JAN10:00:00:21	400045	0	0	0	5	0.0367	60	0	0	0		
28	01JAN10:00:00:21	400057	0	0	0	0	0	0	0	0	0		
29	01JAN10:00:00:21	400069	0	0	0	2	0.0161	67	3	0.0128	71	1	0.0056
30	01JAN10:00:00:21	400088	0	0	0	0	0	0	0	0	0		
31	01JAN10:00:00:21	400109	0	0	0	1	0.0083	65	3	0.0233	65	3	0.0122
32	01JAN10:00:00:21	400138	4	0.1539	14	0	0	0	3	0.0211	65	6	0.0233
33	01JAN10:00:00:21	400147	0	0	0	3	0.0161	86	3	0.0161	65		
34	01JAN10:00:00:21	400160	0	0	0	1	0.0039	71	2	0.0122	65	2	0.0089
35	01JAN10:00:00:21	400172	0	0	0	2	0.0144	71	4	0.0283	71	0	0
36	01JAN10:00:00:21	400181	0	0	0	1	0.0094	78	0	0	0		
37	01JAN10:00:00:21	400195	0	0	0	2	0.0117	93	0	0	0	2	0.0144
38	01JAN10:00:00:21	400206	0	0	0	5	0.0206	60	6	0.0456	65	4	0.0317
39	01JAN10:00:00:21	400227	0	0	0	0	0	0	1	0.01	65	1	0.0061
40	01JAN10:00:00:21	400238	0	0	0	2	0.0133	3	2	0.0128	86		
41	01JAN10:00:00:21	400253	1	0.0056	71	5	0.0322	78	0	0	0		
42	01JAN10:00:00:21	400325	0	0	0	0	0	0	0	0	0	3	0
43	01JAN10:00:00:21	400394	1	0.0067	71	0	0	0	0	0	0	0	0
44	01JAN10:00:00:21	400400	2	0.0128	60	2	0.0117	60	0	0	0	1	0.0094
45	01JAN10:00:00:21	400440	1	0.0067	78	1	0.0067	71	1	0.0067	78	1	0.0061
46	01JAN10:00:00:21	400474	0	0	0	0	0	0	0	0	0	0	0
47	01JAN10:00:00:21	400508	1	0.0056	71	2	0.0106	71	0	0	0		
48	01JAN10:00:00:21	400514	0	0	0	1	0.0072	65	2	0.015	60		

Figure 8.   Raw Data from VDSs

SOURCE: PeMS database.

Non-Crash Events

Since the modeling approach adopted here was binary classification, we collected traffic data for both crash and non-crash cases. The traffic data for the non-crash cases would be representative of normal conditions on the freeways, whereas the traffic data for the crash cases represent crash-prone conditions. To represent normal freeway traffic conditions, we generated a sample of random traffic conditions. As the crashes occurred both on- and off-peak, non-crashes for the same conditions were generated to sample overall traffic conditions. To generate the random non-crash sample, the total study period was divided into one-minute periods from which a random sample of times could be selected as the time of the non-crash event. Similarly, milepost locations for non-crash cases could also be drawn from any milepost from the beginning to the end of the corresponding corridor. All possible combinations of date-time and mileposts were used to derive a sample of non-crash cases. To adequately represent normal conditions, for every crash event in the analysis, ten non-crash events were generated. An earlier study testing different ratios of crash to non-crash events found that the number of non-crashes included had no effect on the classification accuracy of the model (Pande, Abdel-Aty, and Hsia, 2005). A snapshot of the process of generating the random non-crash sample is shown in Figure 9. Excel's RANDBETWEEN function is used in the process.

	A2	▼	⊙	*fx*	=RANDBETWEEN(37531,39237)								
◢	A	B	C	D	E	F	G	H	I	J	K	L	M
1											Date	Time	Milepost
2	38470.00	384.7	323	11/20/2010	1/1/2010	994	0.690278	0:00	16:34		11/20/2010	16:34	384.7
3	39016.00	390.16	271	9/29/2010	1/1/2010	313	0.217361	0:00	5:13		9/29/2010	5:13	390.16
4	38664.00	386.64	230	8/19/2010	1/1/2010	1020	0.708333	0:00	17:00		8/19/2010	17:00	386.64
5	38818.00	388.18	114	4/25/2010	1/1/2010	625	0.434028	0:00	10:25		4/25/2010	10:25	388.18
6	38858.00	388.58	225	8/14/2010	1/1/2010	25	0.017361	0:00	0:25		8/14/2010	0:25	388.58
7	39229.00	392.29	212	8/1/2010	1/1/2010	1271	0.882639	0:00	21:11		8/1/2010	21:11	392.29
8	38755.00	387.55	23	1/24/2010	1/1/2010	671	0.465972	0:00	11:11		1/24/2010	11:11	387.55
9	38421.00	384.21	47	2/17/2010	1/1/2010	1087	0.754861	0:00	18:07		2/17/2010	18:07	384.21
10	38584.00	385.84	481	4/27/2011	1/1/2010	299	0.207639	0:00	4:59		4/27/2011	4:59	385.84
11	38024.00	380.24	296	10/24/2010	1/1/2010	253	0.175694	0:00	4:13		10/24/2010	4:13	380.24
12	38015.00	380.15	81	3/23/2010	1/1/2010	1249	0.867361	0:00	20:49		3/23/2010	20:49	380.15
13	39009.00	390.09	109	4/20/2010	1/1/2010	747	0.51875	0:00	12:27		4/20/2010	12:27	390.09
14	38375.00	383.75	198	7/18/2010	1/1/2010	937	0.650694	0:00	15:37		7/18/2010	15:37	383.75
15	38730.00	387.3	419	2/24/2011	1/1/2010	1006	0.698611	0:00	16:46		2/24/2011	16:46	387.3
16	37541.00	375.41	472	4/18/2011	1/1/2010	1383	0.960417	0:00	23:03		4/18/2011	23:03	375.41
17	38552.00	385.52	456	4/2/2011	1/1/2010	12	0.008333	0:00	0:12		4/2/2011	0:12	385.52
18	38143.00	381.43	196	7/16/2010	1/1/2010	318	0.220833	0:00	5:18		7/16/2010	5:18	381.43
19	37941.00	379.41	229	8/18/2010	1/1/2010	1335	0.927083	0:00	22:15		8/18/2010	22:15	379.41
20	38311.00	383.11	60	3/2/2010	1/1/2010	1096	0.761111	0:00	18:16		3/2/2010	18:16	383.11
21	37756.00	377.56	306	11/3/2010	1/1/2010	442	0.306944	0:00	7:22		11/3/2010	7:22	377.56
22	38390.00	383.9	450	3/27/2011	1/1/2010	615	0.427083	0:00	10:15		3/27/2011	10:15	383.9
23	37715.00	377.15	182	7/2/2010	1/1/2010	887	0.615972	0:00	14:47		7/2/2010	14:47	377.15
24	38555.00	385.55	94	4/5/2010	1/1/2010	42	0.029167	0:00	0:42		4/5/2010	0:42	385.55
25	39043.00	390.43	300	10/28/2010	1/1/2010	369	0.25625	0:00	6:09		10/28/2010	6:09	390.43
26	37972.00	379.72	197	7/17/2010	1/1/2010	320	0.222222	0:00	5:20		7/17/2010	5:20	379.72
27	38286.00	382.86	397	2/2/2011	1/1/2010	484	0.336111	0:00	8:04		2/2/2011	8:04	382.86
28	38468.00	384.68	23	1/24/2010	1/1/2010	1116	0.775	0:00	18:36		1/24/2010	18:36	384.68
29	38856.00	388.56	453	3/30/2011	1/1/2010	383	0.265972	0:00	6:23		3/30/2011	6:23	388.56
30	38859.00	388.59	128	5/9/2010	1/1/2010	13	0.009028	0:00	0:13		5/9/2010	0:13	388.59
31	39064.00	390.64	410	2/15/2011	1/1/2010	535	0.371528	0:00	8:55		2/15/2011	8:55	390.64
32	38104.00	381.04	206	7/26/2010	1/1/2010	852	0.591667	0:00	14:12		7/26/2010	14:12	381.04

Figure 9.   Random Generation of Non-Crash Events

The nearest three VDS both upstream and downstream of the event-location milepost were also identified for all of the non-crash events. The time horizon (from 20 minutes before the crash up to five minutes after the crash) was also the same as that for the crash events. Figure 10 shows a sample spreadsheet of this identification process. The station-arrangement convention for any crash is depicted in Figure 11.

	A	B	D	H	I	J	K	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV
1				VDS ID:	401578	401586	401660	US3 PM	US2 PM	US1 PM	DS1 PM	DS2 PM	DS3 PM	US3 ID	US2 ID	US1 ID	DS1 ID	DS2 ID	DS3 ID
2	Date	Time	Abs Postmi	VDS PM:	365.57	365.97	367.01												
3	2/8/2011	8:26	378.27		-12.7	-12.3	-11.26	376.49	376.87	377.77	378.93	379.68	380.63	400014	402350	400474	402353	400858	402354
4	1/27/2010	13:12	383.23		-17.66	-17.26	-16.22	381.73	382.33	382.9	383.63	384.12	384.23	402356	402358	400195	402359	401908	402361
5	2/1/2011	10:15	388.61		-23.04	-22.64	-21.6	387.48	387.9	388.37	388.92	389.23	389.74	400965	400001	402364	400863	400911	400760
6	3/31/2010	5:57	387.66		-22.09	-21.69	-20.65	385.85	386.24	387.48	387.9	388.37	388.92	400668	402362	400965	400001	402364	400863
7	4/7/2011	14:40	378.16		-12.59	-12.19	-11.15	376.49	376.87	377.77	378.93	379.68	380.63	400014	402350	400474	402353	400858	402354
8	3/23/2010	8:37	378.75		-13.18	-12.78	-11.74	376.49	376.87	377.77	378.93	379.68	380.63	400014	402350	400474	402353	400858	402354
9	1/22/2010	18:43	386.73		-21.16	-20.76	-19.72	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
10	4/4/2011	7:19	387.92		-22.35	-21.95	-20.91	386.24	387.48	387.9	388.37	388.92	389.23	402362	400965	400001	402364	400863	400911
11	2/13/2010	3:30	378.76		-13.19	-12.79	-11.75	376.49	376.87	377.77	378.93	379.68	380.63	400014	402350	400474	402353	400858	402354
12	12/26/2010	8:33	388.31		-22.74	-22.34	-21.3	386.24	387.48	387.9	388.37	388.92	389.23	402362	400965	400001	402364	400863	400911
13	2/26/2011	22:57	387.83		-22.26	-21.86	-20.82	385.85	386.24	387.48	387.9	388.37	388.92	400668	402362	400965	400001	402364	400863
14	1/24/2010	3:03	379.47		-13.9	-13.5	-12.46	376.87	377.77	378.93	379.68	380.63	381.03	402350	400474	402353	400858	402354	402355
15	1/12/2010	16:59	379.52		-13.95	-13.55	-12.51	376.87	377.77	378.93	379.68	380.63	381.03	402350	400474	402353	400858	402354	402355
16	7/11/2010	13:25	388.96		-23.39	-22.99	-21.95	387.9	388.37	388.92	389.23	389.74	390.29	400001	402364	400863	400911	400760	400964
17	4/3/2010	6:00	391.47		-25.9	-25.5	-24.46	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
18	3/26/2010	6:01	391.7		-26.13	-25.73	-24.69	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
19	3/6/2011	20:24	381.92		-16.35	-15.95	-14.91	381.45	381.46	381.73	382.33	382.9	383.63	400325	400591	402356	402358	400195	402359
20	10/22/2010	14:29	390.55		-24.98	-24.58	-23.54	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
21	1/16/2011	9:48	381.95		-16.38	-15.98	-14.94	381.45	381.46	381.73	382.33	382.9	383.63	400325	400591	402356	402358	400195	402359
22	9/16/2010	10:33	387.17		-21.6	-21.2	-20.16	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
23	1/8/2010	13:11	384.48		-18.91	-18.51	-17.47	383.63	384.12	384.23	384.74	384.83	385.47	402359	401908	402361	400400	401890	400996
24	11/28/2010	14:32	389.34		-23.77	-23.37	-22.33	388.37	388.92	389.23	389.74	390.29	390.45	402364	400863	400911	400760	400964	400227
25	2/1/2011	11:00	385.29		-19.72	-19.32	-18.28	384.23	384.74	384.83	385.47	385.85	386.24	402361	400400	401890	400996	400668	402362
26	2/25/2010	22:35	380.49		-14.92	-14.52	-13.48	377.77	378.93	379.68	380.63	381.03	381.45	400474	402353	400858	402354	402355	400325
27	8/20/2010	15:02	389.41		-23.84	-23.44	-22.4	388.37	388.92	389.23	389.74	390.29	390.45	402364	400863	400911	400760	400964	400227
28	3/13/2010	19:03	386.95		-21.38	-20.98	-19.94	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
29	10/19/2010	17:38	376.01		-10.44	-10.04	-9	373.17	373.65	375.72	376.29	376.49	376.87	402345	401658	402347	400181	400014	402350
30	4/17/2011	13:27	388.81		-23.24	-22.84	-21.8	387.48	387.9	388.37	388.92	389.23	389.74	400965	400001	402364	400863	400911	400760
31	3/29/2010	10:53	386.68		-21.11	-20.71	-19.67	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
32	7/16/2010	7:29	388.31		-22.74	-22.34	-21.3	386.24	387.48	387.9	388.37	388.92	389.23	402362	400965	400001	402364	400863	400911
33	8/19/2010	0:10	391.19		-25.62	-25.22	-24.18	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
34	11/29/2010	1:38	385.57		-20	-19.6	-18.56	384.74	384.83	385.47	385.85	386.24	387.48	400400	401890	400996	400668	402362	400965
35	2/26/2010	1:14	391.7		-26.13	-25.73	-24.69	389.74	390.29	390.45	392.32	392.34	392.42	400760	400964	400227	401994	400246	401995
36	12/2/2010	22:23	379.25		-13.68	-13.28	-12.24	376.87	377.77	378.93	379.68	380.63	381.03	402350	400474	402353	400858	402354	402355
37	2/25/2011	1:27	385.37		-19.8	-19.4	-18.36	384.23	384.74	384.83	385.47	385.85	386.24	402361	400400	401890	400996	400668	402362
38	3/23/2011	18:43	386.98		-21.41	-21.01	-19.97	385.47	385.85	386.24	387.48	387.9	388.37	400996	400668	402362	400965	400001	402364
39	9/7/2010	4:46	385.15		-19.58	-19.18	-18.14	384.23	384.74	384.83	385.47	385.85	386.24	402361	400400	401890	400996	400668	402362

Figure 10. Identification of Nearest Three Upstream and Downstream VDSs
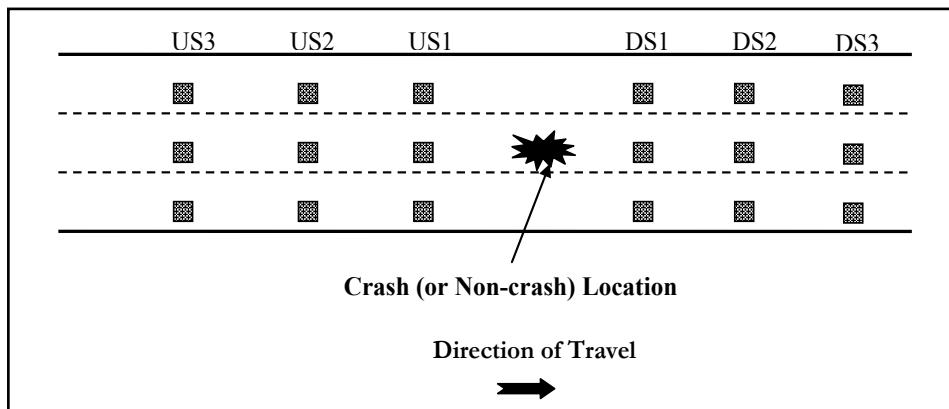

Figure 11. Arrangement of the Loop-Detector Stations

The upstream station IDs, in order of increasing distance from the crash site, are US1, US2, and US3; downstream stations, in order of increasing distance from the crash site, are DS1, DS2, and DS3 (yellow highlighted cells in Figure 10). In addition to the IDs, the spreadsheet also shows the mileposts that were identified for these VDS locations (highlighted in orange in Figure 10).

Data Aggregation

As noted by Pande and Abdel-Aty (2006a), there is significant noise in the raw 30-second loop-detector data, making them unsuitable for modeling purposes. Hence, for each of the six VDS locations identified for crash and non-crash events, individual variables were

averaged across all lanes and aggregated into five-minute intervals. The intervals are: 0–5 minutes after the event (time slice 0), 0–5 minutes before the event (time slice 1), 5–10 minutes before the event (time slice 2), 10–15 minutes before the event (time slice 3), and 15–20 minutes before the event (time slice 4). For these time slices, standard deviations of the variables were also calculated, since past studies documented in the literature noted that variation in traffic parameters was critically associated with freeway crash potential.

As time slice 0 occurs after the crash, it is relevant only to incident detection and will not be further analyzed or discussed here. The five-minute intervals preceding a crash were selected on the basis of previous research by Pande. Generally, the closer the analysis interval to the crash time, the more accurate the model prediction will be. However, there must also be sufficient time for a traffic-management center to identify crash-prone conditions and deploy countermeasures; it is therefore likely that only the time-slice 2, 3, and 4 models will be relevant for proactive crash management.

The nomenclature for these average and standard deviations is of the form $XYZ\alpha\_b$. X takes the value A or S for average and standard deviation, respectively. Y takes the value S, V, or O for speed, volume, or lane-occupancy, respectively. $Z\alpha$ takes the value of U1, U2, U3 or D1, D2, D3, depending on the station to which a traffic parameter belongs (the nearest upstream/downstream station relative to the crash location being U1/D1, and subsequent detectors being U2/D2 and U3/D3, respectively). β takes the values 1, 2, 3, or 4, referring to the aforementioned four time slices. Hence, ASD1_2 and AVU1_2 represent average speed on station DS1 over time slice 2 and average volume on station US1 over time slice 2, respectively. All these averages and standard deviations were calculated for both crash and non-crash cases.

The data described above are used in Chapter IV to estimate and test statistical (binary logistic regression) and data mining (classification tree) models for classifying crash-prone and normal conditions on the freeways.

IV. MODELING TOOLS, ANALYSIS, AND RESULTS

As noted earlier, this study applies two different modeling tools—logistic regression and classification trees—to identify crash-prone conditions. These tools are applied to data from U.S. 101 NB, and the models estimated from the data are then applied to the U.S. 101 SB and I-880 NB/SB segments. This chapter provides the details of the statistical and data mining methods and then describes the analysis and results.

LOGISTIC REGRESSION

In a logistic regression, setting the function of dependent variables yielding a linear function of the independent variables would be the logit transformation:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

(1)

where $\pi(x) = E(Y|x)$ is the conditional mean of Y (the dependent variable representing crash occurrence; $Y = 1$ in this case) given independent variable x when the logistic distribution is used. Under the assumption that the logit is linear in continuous covariate x, the equation for the logit would be $g(x)$. Once the model (i.e., the coefficient βs) is esti-mated for the binary target variable, it can be used to score any dataset that contains the required input variable to the model (i.e., x). The output of the model is in the form of a posterior probability of crash occurrence, lying between 0 and 1. Note that the same formulation may be extended to multiple independent variables, as is the case in this research. In case of multiple independent variables, a standard stepwise variable-selection method will be used to finalize the set of variables that are significantly asso-ciated with the crash occurrence. The details of logistic regression and the stepwise variable-selection procedure can be found in standard texts on logistic-regression and binary-data modeling (e.g., Collett, 1991, and Hosner and Lemeshow, 1989).

CLASSIFICATION TREES

A classification tree represents segmentation of data created by applying a series of simple rules, each of which assigns an observation to a group based on the value of an input. One rule is applied after another, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node. The final, or terminal, nodes are called leaves. For each leaf, a decision is made and applied to all observations in that leaf. Decision trees are one of the most widely utilized tools in data mining applications and may be used for classification of categorical variables, as well as for continuous targets. (The latter application, of course, is not relevant here.) The advantage of classification trees over other modeling tools such as neural networks is that they produce a model that may represent interpretable English rules or logic statements. The other advantage of trees is that no assumptions are necessary about the data and the model form. In the next subsection, theoretical details of the classification trees are described. Since we invariably deal with binary target variables (Y=1 for crash and Y=0 for non-crash) in this study, the details of the methodology are provided in the context of a binary target. Neural networks and decision-tree algorithms have been successfully used to develop

classification models for crash severity as a function of potentially correlated categorical factors (Sohn and Shin, 2001) and, more recently, to demonstrate significant correlation between speed differentials upstream/downstream and crash risk (Pande and Abdel-Aty, 2006).

The basic action in classification-tree construction is to split each (non-terminal) node such that the descendant nodes are "purer" than the parent node. In a completely "pure" node, all of the observations belong to the same class. To achieve this, a set of candidate split rules is created that consists of all possible splits for all variables included in the analysis. For example, a dataset with 200 observations and five input variables would have up to 200 ´ 5 = 1000 splits available at the root node. These splits are then evaluated based on a criterion to choose among various available splits at every non-terminal node (including the root node). The Gini index is used as the measure (i.e., purity functions) to rank candidate splits for a binary target variable. This measure was proposed by Breiman et al. (1984).

One of the criteria is applied recursively to the descendants, which become the parents to successive splits, and so on. The splitting process is continued until the criterion of minimum reduction in impurity (i.e., reduction in the Gini Index) and/or minimum size of a node is satisfied. To stop the splitting process, one may choose the classification accuracy over the validation dataset (i.e., the dataset not used for estimating the splits) as the criterion. The classification accuracy may be assessed after every split, and the process may be terminated if it declines after a particular split. The output from the classification-tree model is also the posterior probability of an observation being a crash (a number lying between 0 and 1).

These tools are selected because they can provide not only a measure for crash vs. non-crash classification, but also because the variables included in the model can be explained. Neural networks were also considered as a tool but were not used because of their "black box" nature. In other words, the effect of individual independent variables on the output is not transparent.

METHOD FOR ANALYSIS OF CLASSIFICATION PERFORMANCE

Assumptions

Some critical issues needed to be addressed before proceeding with the modeling exercise. Crashes, however frequent on the corridors under consideration, are still rare events. Sampling their actual proportion in the dataset would mean that the sample would consist almost exclusively of non-crash cases (crash cases would be less than 0.001%). It is reasonable to assume that the crash-prone conditions, which would justify issuing warnings, are more frequent than the crashes themselves. For any model intended to be applied in real-time, the ideal sample composition for modeling would have the proportion of the two competing events the same as that in reality. However, there is currently no way to estimate the proportion of crash-prone conditions on the freeway. Also, since the number of warnings beyond a certain point would constitute an "unreasonable" number of false alarms, the decision from the models cannot be positive (i.e., cannot predict a crash) around 50% of the time. Hence, a sample with equal numbers of crash and non-crash

cases would not be ideal. Therefore, we deemed 10% to be an appropriate ratio of crash vs. non-crash cases and included ten non-crash cases for each crash in the sample.

Because of the imbalance in the proportions of crashes and non-crash cases, model performance evaluation becomes complicated. The output of the models (for any observation) is the posterior probability of the crash, i.e., a number between 0 and 1. The closer it is to unity, the more likely, according to the model, the observation is to be a crash. The overall classification accuracy based on a pre-selected threshold is usually an appropriate measure for judging the performance of the model. However, with only 9.1% of the crashes in the sample (one crash for ten non-crash cases) used for modeling, 90.9% overall classification accuracy could be achieved by a model that merely classifies every data point as non-crash. Such a model would, of course, be useless for identifying crash-prone conditions. Also, since the classification performance of the models would vary based on the cutoff set on the output from the models (i.e., the posterior probability), even the classification accuracy for each individual class (at a certain cutoff) would not be appropriate for comparing the performance of competing models. It would reflect only the performance of the model at a predetermined threshold on output posterior probability. This is especially true here, since we have two different classes of models, and their outputs are calibrated differently. The same threshold can potentially produce varying results for the different classes of modeling techniques. Therefore, a well-calibrated measure of performance evaluation was needed; that measure is presented below.

The Performance Evaluation Measure

We evaluate the performance of the estimated models by applying them to a dataset consisting of the input variables. The output of these models (for any observation) is the posterior probability of a crash. The closer the posterior probability is to unity, the more likely, according to the model, it is for that observation to be a crash. We therefore sorted the output dataset by the estimated posterior probability. In the sorted group, the top 10% of observations would be those most likely to be a crash, according to the model. The performance of a model may be measured by determining the proportion of crashes captured within various deciles of posterior probability. (A decile is defined as any of nine points that divide a distribution of ranked scores into equal intervals, each containing one-tenth of the scores.) Since these models are intended to identify an event as rare as a crash, the proportion of crashes captured within the first few deciles must be critically examined. We decided that the best model among a set of competing models would be the one capturing the highest percentage of crashes within the first three deciles (i.e., the 30th percentile). As mentioned earlier, because of the imbalance in the proportions of crash and non-crash cases in the sample, this dataset would not be a good measure for model performance evaluation.

LOGISTIC-REGRESSION ANALYSIS

Overview

The multivariate logistic-regression modeling was estimated using U.S. 101 NB data. The statistical analysis software (SAS) package (SAS Institute, 2001) was used to fit the

regression models. The target variable for the logistic-regression models was Y, with a value of 0 for non-crash cases and 1 for crashes. The independent variables of interest were average speed, standard deviation of speed, average volume, standard deviation of volume, average lane occupancy, and standard deviation of lane occupancy calculated over each VDS location and time slice. The traffic parameters (speed, volume, and lane-occupancy) were not all included simultaneously in any model, and speed-based models were created separately from the volume and occupancy-based models, as the study VDSs were all based on single-loop detectors. This implies that speed was calculated from the volume and occupancy data and was not independently measured. Including those data in the same regression model would have led to an unacceptable level of correlation in independent variables. A stepwise selection process was used to identify the most significant variables, and the model coefficients were estimated for these significant variables.

A total of 30 logistic-regression models were estimated, using traffic information from four time slices (ranging from 0 to 20 minutes before the crash in five-minute intervals) and three sets of VDS locations. For each VDS and time-slice combination, there were two models: one based on Caltrans' derived speed information and one based on independently measured volume and lane-occupancy information. The crash-risk estimation models are identified as PredX_Y_Z, where:

> X is the number of VDS stations upstream and downstream of the crash (or non-crash) location (1, 2, or 3) contributing traffic information to the model

> Y is the time slice number (1, 2, 3, or 4), as described in Chapter III

> Z is whether the model uses speed information (s) or volume and lane-occupancy information (v)

For example, Pred1_4_s would indicate that the model is developed from a dataset of speed observations from the one nearest VDS both upstream and downstream of the crash, over the period of 15–20 minutes before the crash occurred. This model utilizes traffic data from two VDS locations

As another example, Pred3_4_s would indicate that the model used the dataset of speed observations from the nearest three loop detectors both upstream and downstream of the crash, over the period of 15–20 minutes before the crash occurred. This model utilizes traffic data from a total of six VDS locations, two of which are the same as those in Pred1_4_s.

The 30 models were applied to the dataset used to estimate the models containing observations (both crash and non-crash events for U.S. 101 NB), with the posterior probability of the observation being a crash estimated for each observation. The models were then compared in terms of the cumulative proportion of crashes correctly identified within 30% of the observations they predicted were most likely to be a crash (the criterion selected based on the discussion in the previous section). The percentage of crashes identified by each model can also be examined in the context of the "performance" of a random baseline "model" that represents the percentage of crashes identified in the

sample if the observations are randomly assigned as crash and non-crash. Of course, in any set of 30% observations, such a "model" could correctly identify 30% of the crashes in the dataset. Any model can be assessed for its classification based on the difference between crashes it identifies within the first three deciles and 30%, which is the percentage of crashes that can be identified by the random, baseline "model."

Using this criterion, we selected the best model from subsets of one, two, and three upstream/downstream VDS models. Traffic parameters from time slice 1, being too close to the time of the crash used in a model, would leave absolutely no time available to process, analyze, and disseminate the information that could be used to avoid crashes. Hence, models using variables measured only during time slice 2, 3, or 4 are given further consideration.

The single loops analyzed in this study collect raw volume and occupancy data and use a predetermined effective vehicle length (g-factor) to calculate average speed; dual loops, in contrast, can measure speeds directly. Acknowledging that this g-factor will vary by lane, time of day, and loop sensitivity, PeMS calculates it for each loop for every five-minute period during an average week to improve the accuracy of the speed estimates. The smoothed g-factor is then applied to the real-time VDS data to obtain speeds. These speeds are then smoothed with an exponential filter, which is weighted based on traffic flow to produce reasonable estimates of speed (lower flow conditions require more smoothing).

In general, the volume and occupancy (v) models had a much higher classification accuracy at the 30th percentile than the speed (s) models. This is understandable, as the speeds derived by the PeMS algorithm are inherently less reflective of field conditions than the actual VDS data. Additionally, only the volume and occupancy data are reported live by Caltrans districts (in a variety of methods, including extensible markup language (XML) feed over TPC, SQLnet, and raw controller packets over RPC [remote procedure call]); speeds must be post-processed from these transmitted data. For reasons of model reliability and applicability in a real-time framework, only the volume and occupancy models will be considered further. The following section describes the U.S. 101 NB models for all available crash and non-crash data.

Comparison of Crash Models

The best models for all crashes, using the 30th-percentile selection criterion, are summarized in Table 3.

Table 3.   Best Models for All Crashes

Model Name[a]	Time Slice	Cumulative % of Crashes Captured Within First Three Deciles (30th Percentile)
Pred1_4_V	**4**	**53.463**
Pred1_3_V	3	52.276
Pred1_2_V	2	50.069
Pred2_4_V	4	56.546
Pred2_3_V	3	56.711
Pred2_2_V	**2**	**57.524**
Pred3_4_V	**4**	**61.749**
Pred3_3_V	3	61.264
Pred3_2_V	2	60.000

[a]*The best models are indicated in boldface.*

The best one-VDS model used volume and occupancy data from the fourth time slice, Pred1_4_v. The best two-VDS model used volume and occupancy data from the second time slice, Pred2_2_v. The best three-VDS model used volume and occupancy data from the fourth time slice, Pred3_4_v. Classification accuracy increases when data from more VDS locations are used. The model from three VDSs upstream and downstream is able to identify more than 61% of the crashes, a 31% improvement over the random baseline "model."

Model Details

The coefficients of the best one-VDS, two-VDS, and three-VDS logistic-regression models are shown in Tables 4, 5, and 6, respectively. Only the variables used in models based on the stepwise selection procedure are included. In addition to the model parameters, the tables provide the corresponding p-values for the model coefficients. A p-value less than 0.05 indicates that the variable is significant at the 95% confidence level. A positive (negative) coefficient means that as the value of the corresponding variable increases, the crash-risk measure increases (decreases).

Table 4.   Model Coefficients for the Best One-VDS Model

Parameter[a]	Estimate	Pr > ChiSq (p-value)
AVDS1_4	**0.1**	**<0.0001**
AVUS1_4	**0.08**	**<0.0001**
AODS1_4	**1.72**	**<0.0001**
AOUS1_4	**0.87**	**0.0058**
SVDS1_4	0.05	0.1355
SVUS1_4	**–0.1**	**0.0035**
SODS1_4	–0.57	0.2157

Syntax:

    Column 1:   A = average;   S = standard deviation
    Column 2:   O = occupancy;   V = volume   S = speed
    Columns 3 and 4:   DS = downstream;   US = upstream
    For example,   AODS = average occupancy downstream.
    [a]Bold text denotes statistical significance at the 95% confidence level.

Table 5.   Model Coefficients for the Best Two-VDS Model

Parameter[a]	Estimate	Pr > ChiSq (p-value)
AVDS1_2	**0.05**	**0.0138**
AVUS1_2	**0.04**	**0.027**
AODS1_2	**0.91**	**0.0171**
AOUS1_2	**1.5**	**0.0443**
SVDS1_2	0.07	0.0997
SOUS1_2	–0.93	0.2134
AVDS2_2	**0.05**	**0.0442**
AVUS2_2	**0.08**	**0.0013**
AODS2_2	**1.45**	**0.0158**
AOUS2_2	–1.49	0.139
SVDS2_2	**–0.22**	**<0.0001**
SVUS2_2	–0.08	0.081
SODS2_2	**–1.85**	**0.0061**
SOUS2_2	**2.87**	**0.0024**

[a]Bold text denotes statistical significance at the 95% confidence level.

Table 6.   Model Coefficients for the Best Three-VDS Model

Parameter[a]	Estimate	Pr > ChiSq (p-value)
AVUS1_4	**0.13**	**<0.0001**
SVDS1_4	**0.08**	**0.0407**
SVUS1_4	**-0.18**	**0.0017**
AVUS2_4	**0.06**	**0.0318**
AODS2_4	**−1.24**	**0.0389**
SVDS2_4	**−0.09**	**0.0454**
SVUS2_4	−0.1	0.068
SOUS2_4	**2.7**	**<0.0001**
AVDS3_4	**−0.11**	**<0.0001**
AVUS3_4	**0.11**	**<0.0001**
AODS3_4	**1.87**	**0.0112**
AOUS3_4	**3.04**	**0.0003**
SVDS3_4	**0.16**	**0.0023**
SODS3_4	**−1.82**	**0.0169**
SOUS3_4	−1.32	0.1416

[a]*Bold text denotes statistical significance at the 95% confidence level.*

The model coefficients show that the standard deviation of occupancy downstream of a freeway location is negatively associated with crash risk, i.e., if the standard deviation of lane occupancy decreases, crash risk increases. Also, variables representing average occupancy downstream (AODS*_*) have a positive coefficient in all models, indicating that increased lane occupancy (i.e., congestion) downstream of a site increases crash likelihood. Because the specific crash type is not known, it is not possible to associate these coefficients with the relevant crash mechanism. However, the coefficients might be more readily associated with conditions prone to rear-end crashes, which are the most common crash type on urban freeways.

Model Application for Assessing Transferability

A major focus of our research project was evaluation of transferability, that is, the potential to apply the predictive model developed on one freeway segment to other similar facilities nearby. As discussed in Chapter II, previous studies have either failed to address the issue or have tried to apply the model on dissimilar facilities and failed to attain good classification accuracy.

To assess transferability, we used the coefficients of regression models shown in Tables 4, 5, and 6 to score the combined crash and non-crash data for the other three corridors— U.S. 101 SB, I-880 NB, and I-880 SB. For each observation in these datasets, a posterior probability output was obtained. We then examined the proportion of crashes correctly identified within the 30% of observations having the highest posterior probability. The

cumulative percentages of identified crashes for each model on each of the three corridors are shown in Figure 12. The model that identifies a higher proportion of crashes within the 30th percentile is considered a better model.
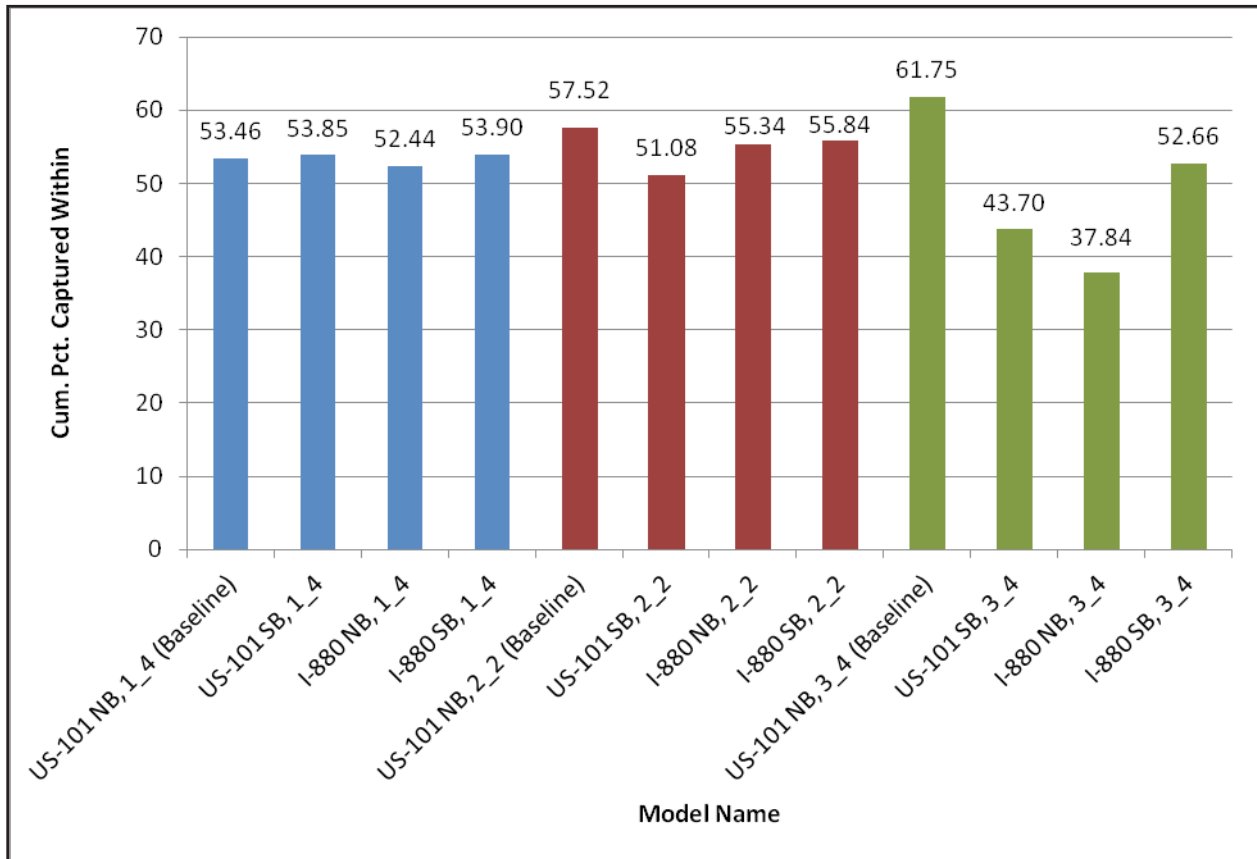

Figure 12. Transferability of the Models to Other Freeways, All Crashes

The same data for each model are presented in Tables 7, 8, and 9, along with the information on upstream/downstream stations and time slices used by each model.

Table 7. Classification Accuracy of the Best One-VDS Model Applied to Other Freeways, All Crashes

Best one-VDS model	pred1_4
VDS US/DS	1
Time slice	4
Minutes before crash	15–20
Selection criterion	30%
Percent captured within segment:	
U.S. 101 NB (estimation baseline)	53.463
U.S. 101 SB	53.846
I-880 NB	52.439
I-880 SB	53.898

Table 8.   Classification Accuracy of the Best Two-VDS Model Applied to Other Freeways, All Crashes

Best 2 VDS model	pred2_2
VDS US/DS	2
Time slice	2
Minues before crash	5–10
Selection criterion	30%
Percent captured within segment:	
U.S. 101 NB (estimation baseline)	57.524
U.S. 101 SB	51.083
I-880 NB	55.340
I-880 SB	55.844

Table 9.   Classification Accuracy of the Best Three-VDS Model Applied to Other Freeways, All Crashes

Best 2 VDS model	pred3_4
VDS US/DS	3
Time slice	4
Minues before crash	15-20
Selection criterion	30%
Percent captured within segment:	
U.S. 101 NB (estimation baseline)	61.749
U.S. 101 SB	43.700
I-880 NB	37.838
I-880 SB	52.660

The one-VDS and two-VDS models work almost as well on nearby freeways as on the roadway for which they were developed. The three-VDS model is a much less accurate predictor of crashes on the nearby roadway segments. In other words, the one- and two-VDS models are easily transferable, while the three-VDS model does not appear to be so.

The three-VDS model appears to be overfitting; we believe that traffic conditions approximately 1.5 miles from a crash location do not have a real relationship with crash risk 15–20 minutes later. The overfitting is happening on the training data; the model does not perform very well when tested with an unseen dataset.

Daytime-Only Models

We also estimated daytime-only models, assuming that late-night crashes are more likely to occur as a result of driver error or driving conditions (e.g., driving under the influence) than because of measurable traffic conditions. The modeling process and model comparison were identical to the above analysis, except that the regression models were estimated using data only for crashes and non-crash cases between the weekday hours of 5:00 am and 10:00 pm. The model results are summarized in Table 10.

Table 10. Best Models for Daytime-Only Crashes

Model Name[a]	Time Slice	Cumulative % of Crashes Captured Within the 30th Percentile (U.S. 101 NB)
Pred1_4_V	**4**	**52.362**
Pred1_3_V	3	51.866
Pred1_2_V	2	50.000
Pred2_4_V	**4**	**57.724**
Pred2_3_V	3	56.873
Pred2_2_V	2	55.285
Pred3_4_V	**4**	**60.324**
Pred3_2_V	3	59.438
Pred3_3_V	2	59.438

[a] *The best models are indicated in boldface.*

Again, the volume-occupancy models performed better than the models based on calculated speed information in almost every case. Therefore, we dropped the speed models from the analysis. The best one-, two-, and three-VDS models all used volume and occupancy data from the fourth time slice.

The daytime-only results are compared with the all-crash results in Table 12.

Table 11. Best Three Models for All Crashes and Daytime-Only Crashes

All Crashes			Daytime-Only Crashes		
Model Name	Time Slice	Cumulative % Captured	Model Name	Time Slice	Cumulative % Captured
One-VDS					
pred1_4_v	4	53.463	pred1_4_v	4	52.362
Two-VDS					
pred2_2_v	2	57.524	Pred2_4_v	4	57.724
Three-VDS					
pred3_4_v	4	61.749	pred3_4_v	4	60.324

There is no appreciable difference between the performance of the all-crash models and that of the daytime-only crash models. Hence, it is not advantageous to estimate the model for daytime crashes only. Therefore, the transferability analysis for daytime-only crashes is not discussed here.

CLASSIFICATION-TREE ANALYSIS

Overview

Classification-tree models are frequently used data mining tools. However, they tend to overfit the data, which affects their future performance on unseen datasets. Therefore, in data mining analysis, models are usually estimated with a "training dataset" of 70%

of the available observations and then validated with the remaining 30%. Validating with the unseen dataset helps identify more-robust models in terms of performance on new datasets. The choice of using a separate validation dataset was informed by the literature, which has shown the overfitting and instability problems of classification trees to be much worse than those in the regression models (e.g., Perlich, Provost, and Simonoff, 2003).

We estimated 30 different classification-tree models, and those using data from time slice 1 were then excluded, for the reasons discussed above for logistic-regression models. The speed classification-tree models were generally better than the volume-occupancy models. The classification-tree models were compared using the same metric we used for the logistic-regression models, the proportion of validation-dataset crashes identified within the top 30th percentile.

Selection of the Best Model

In addition to different one-VDS, two-VDS, and three-VDS models from different time slices, we estimated a classification-tree model using only time of crash (and non-crash) as input. This model was estimated to ensure that the models provide real differentiation between crash-prone and normal traffic conditions. If the models using traffic data are providing valuable information about crash risk, they should perform much better than the model with only time-of-crash/non-crash information. These models do, in fact, perform much better, as shown in Table 12. It is clear that while the time-of-crash model performs better than the random baseline "model" (i.e., identifies more than 30% of the crashes), its performance is significantly worse than that of the models using traffic information. The model performance in Table 12 is based on the 30% validation dataset. The results in Table 12 and Figure 13 identify the two-VDS, time-slice 3 model as the most accurate classifier on the validation dataset.

Table 12. Classification Accuracy of Classification-Tree Models

Model Name	US/DS VDS Locations	Time Slice	Cumulative % of Crashes Captured Within the 30th Percentile (validation dataset)
Pred1_4_s	1	4	56.662
Pred2_3_s	2	3	58.647
Pred3_3_s	3	3	56.309
Time-of-crash model	-	-	43.771

Model Details

Classification-tree models are a series of "if-then" rules created to classify the observations. Sample code is provided in the Appendix. The variables analyzed through classification trees for crash vs. non-crash classification can be ranked by combinations of the number of times they appear in various rules and the number of observations they contribute. The variables in the best classification-tree model (Pred2_3_s) were ranked as follows:

1. SSDS2_3

2. ASDS2_3

3. ASUS1_3

4. SSUS1_3

5. ASUS2_3

6. SSDS1_3

7. ASDS1_3

The standard deviation and averages of speed at the second downstream VDS are the two most significant variables. These results are consistent with those of past studies, which have found that the turbulence in speed downstream of a location is significantly associated with crash risk on urban freeways. The standard deviation of speed at the second upstream VDS (SSUS2_3) was the only variable found to be not associated with crash likelihood.

Transferability Analysis

The best classification-tree model (Pred2_3_s) was applied to complete sets of data from U.S. 101 SB and I-880 NB/SB. It was also applied on the complete set of U.S. 101 NB data, since the results shown in Table 12 are based on applying the tree model on the validation dataset (i.e., 30% of the observations from U.S. 101 NB). The classification accuracy in Table 13 (61.897%) is higher for U.S. 101 NB than for the validation dataset (58.657%, shown in Table 12), since the complete set also includes the 70% training data. Applying the model to the U.S. 101 NB dataset allows us to compare the classification-tree-model performance with that of the logistic-regression model.

Table 13. Classification Accuracy of the Best U.S. 101 NB Classification-Tree Model on Other Freeways

Segment	Proportion of Crashes Identified Within the 30th Percentile (classification-tree model)	Proportion of Crashes Identified Within the 30 Percentile (logistic-regression model)
U.S. 101 NB (estimation baseline)	61.897	61.749
U.S. 101 SB	46.505	43.700
I-880 NB	40.674	37.838
I-880 SB	50.368	52.660
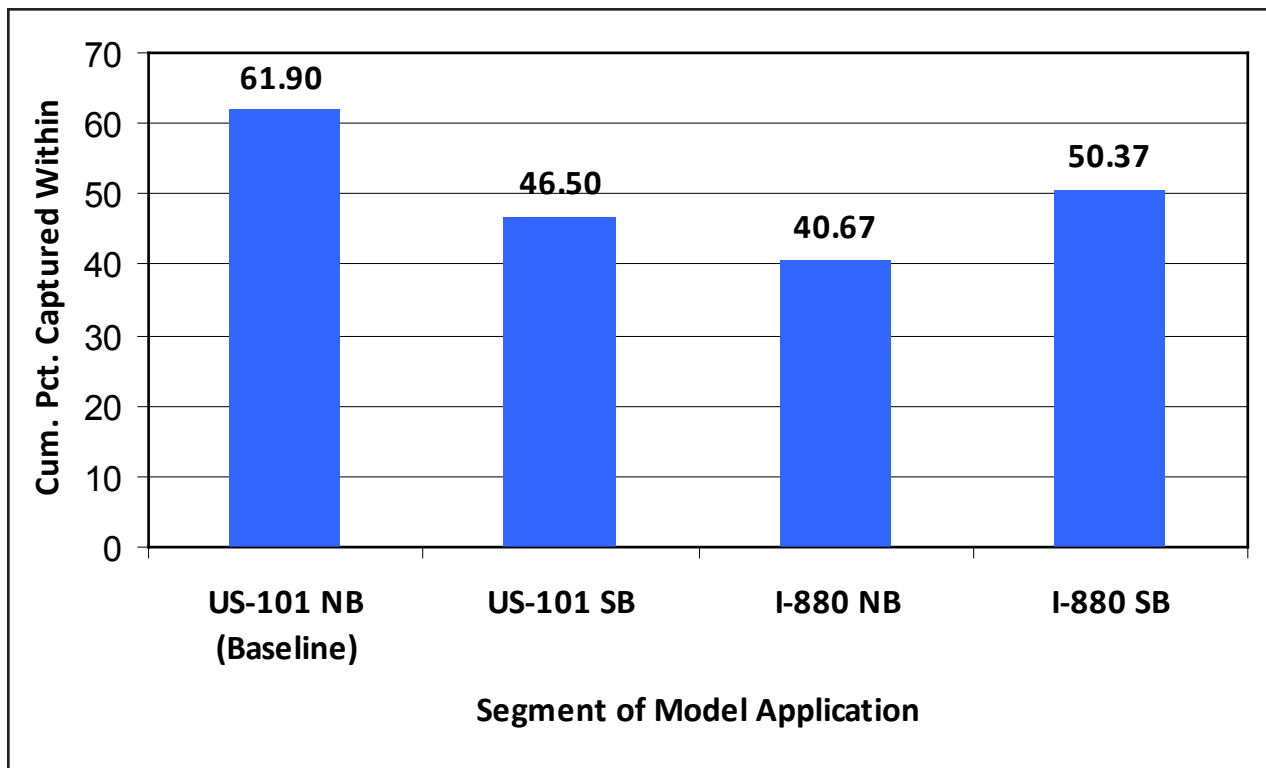
Figure 13. Analysis of Transferability of Models to Other Freeways, All Crashes

The best classification-tree model performs slightly worse on the other freeways, as was the case with best logistic-regression model. I-880 SB is the corridor to which the model estimated from U.S. 101 NB data seems to be most readily transferable.

V. REAL-TIME APPLICATION FRAMEWORK

PROCEDURE

The models developed here may be applied in real-time, as they are capable of classifying the traffic patterns measured at VDSs into posterior probability. A step-by-step procedure is shown graphically in Figure 14 and described below.

We first try to obtain data from three VDSs upstream and downstream of the location of interest, because the three-VDS models are the best estimators of crash risk. If all the VDS are in good condition after a data check, the five-minute averages and standard deviations of traffic variables are calculated for each location. Estimated model coefficients (for logistic-regression models) or "if–then" rules (for classification-tree models) can be applied to obtain the measure of crash risk at the middle of the section.

If data from all six VDSs are not available because of intermittent loop failures, a check for data availability is applied for the two-VDS model (for which a total of four VDSs is needed). Using the same procedure as that for the three-VDS application, traffic parameters are calculated for input into a model. As shown in Chapter IV, models developed using two VDSs on nearby freeways are transferable to other roadways. If models developed specifically for the segment of interest (which perform best) are not available, a two-VDS model from a nearby roadway can be used to estimate crash risk.

If there are only enough good data to run a one-VDS model, the procedure used for the two-VDS model is applied. Traffic parameters are calculated from the VDS data and input into the calibrated one-VDS model for the segment (if available). If a model has not yet been specifically developed for the location, a one-VDS model from another freeway can be applied to produce a reasonably accurate assessment of crash risk. The procedure differs for the nearby freeways in that the check for three-VDS models is not applied, and we have examined data for only two-VDS models for the nearby freeways because the three-VDS models lack transferability.
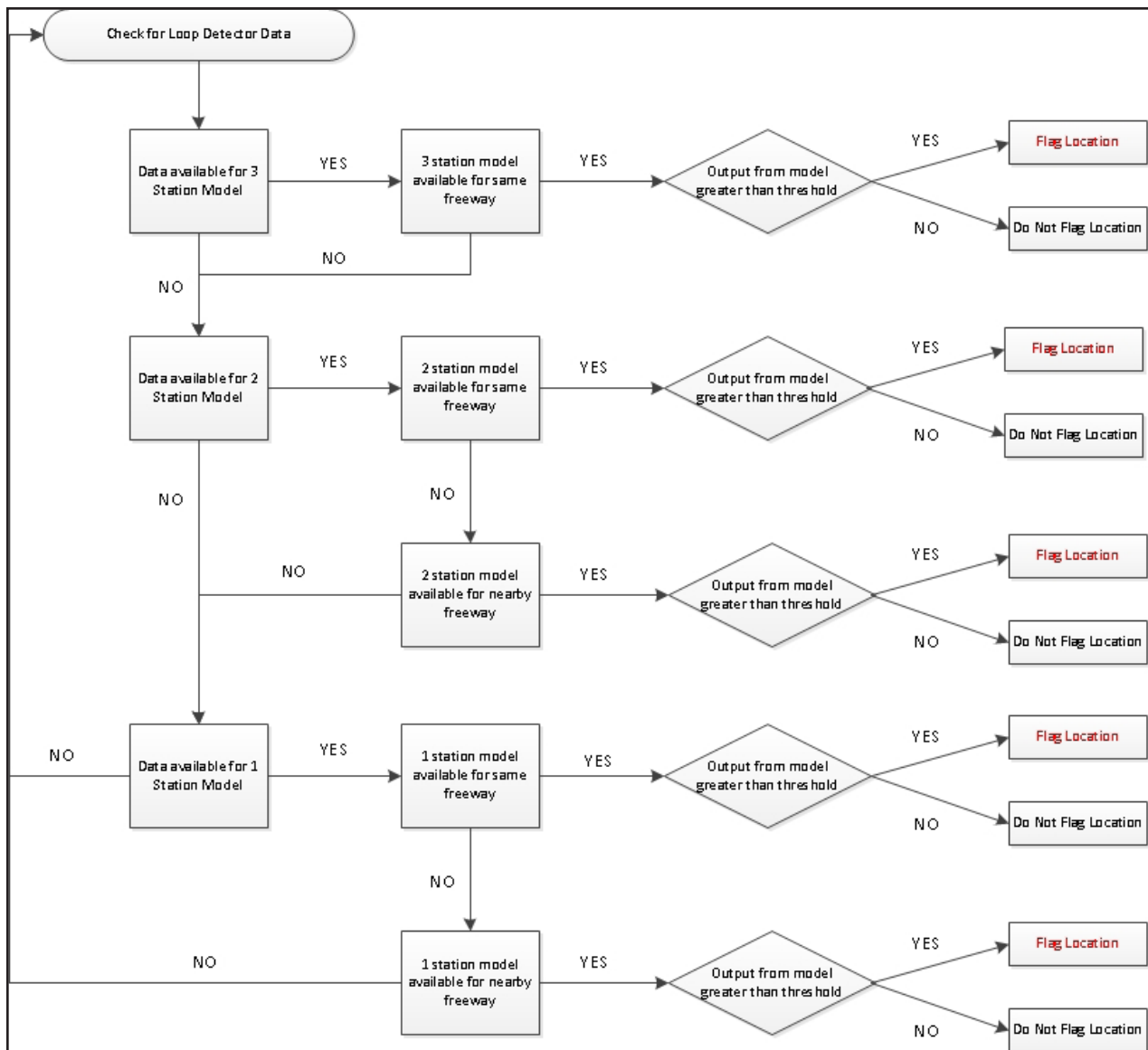
Figure 14. Real-Time Application Procedure

If the output posterior probability for a segment of freeway is consistently high, the traffic-management authorities can keep their crash-mitigation squad on alert to minimize the impacts of crash occurrence. Also, if the models trigger the warning more often on some freeway segments than on others, those segments may be closely watched through the freeway cameras. This will help identify problems associated with those locations. Our findings could also be used in the formulation of VSL or ramp-metering strategies to reduce the likelihood of crashes. These strategies can be tested using microscopic traffic-simulation models.

REAL-TIME APPLICATION ISSUES

One-VDS vs. Two-VDS vs. Three-VDS Models

Even though the one-VDS models may not always achieve as high classification accuracy as the three-VDS model for the same corridor, they have more tolerant data requirements, since the three-VDS models require that data be available from six simultaneous VDS locations. If even one of the VDSs is malfunctioning, the three-VDS model cannot be applied. A one-VDS model, on the other hand, requires data from only two VDS locations.

False Alarms

The formulation of the problem, along with the solution approach adopted here, is similar to incident detection. In fact, we estimated some models that used the data 0–5 minutes after a crash. However, the objective of this analysis is to identify crash-prone conditions—i.e., the conditions in which drivers are more likely to make errors resulting in crashes—rather than to pinpoint the occurrence of a crash. Conditions prior to crashes are not as readily identifiable (possibly due to significant human-factor involvement) as those following crashes. Crashes being such rare events, it is not possible to avoid false alarms. Also, the relative threshold (based on percentiles or posterior probability output) discussed here allows decisionmakers to get an indication of the scale of false alarms being issued based on any given model application. An absolute threshold would provide no such indication.

Adopting even the approach used here for assessing classification-tree models would result in a significant number of false alarms throughout the day. The number could be reduced somewhat by using a higher threshold (e.g., a 20-percentile value for the posterior probability), but it would still be significant. If traffic parameters from time slice 1 were used as inputs instead of parameters from time slice 2, slight improvement could be expected. However, because time slice 1 is too close to the time of a crash, there would be absolutely no time available to process, analyze, and disseminate the information that might be used to avoid crashes. Hence, we believe that the warning of crash-prone conditions could provide not event prediction but a heightened measure of crash risk.

Finally, false alarms are not as detrimental in the present application as they are in case of incident-detection algorithms. In fact, the ultimate goal of this research should be to achieve a false alarm every time a crash warning is issued. With some form of proactive countermeasure or warnings for motorists, potential crashes in crash-prone conditions might be avoided. The justification or inevitability of false alarms does not mean that an unlimited number of warnings could be issued, especially if information based on model output is being transferred to drivers on the freeway. Being judicious about the number of warnings would help to ensure that drivers would not perceive the number of warnings as excessive and would not become immune to them. The whole notion of warnings and drivers' reaction to them is beyond the scope of the present work and requires further investigation.

VI. CONCLUSIONS

The research reported here was undertaken to develop models for linking ITS-archived data with historical crashes on instrumented corridors in the San José metropolitan area and to assess their transferability to other corridors. We assembled a detailed database of all crashes that occurred on four major corridors in the area over a period of 16 months and linked it to the archived loop-detector data from the surrounding VDS locations. The analysis of the models' classification results showed that their continuous output (i.e., posterior probability) can in fact be related to real-time crash risk.

TRANSFERABILITY ANALYSIS

While crash risk-assessment models have been developed for freeways in the United States (I-4 in Orlando, FL), Canada, and the Netherlands, this research advances the knowledge regarding transferability of the models. Specifically, it critically examined the performance of models estimated with data from the U.S. 101 NB corridor on nearby corridors (U.S. 101 SB, I-880 NB, and I-880 SB) and found that the model from one corridor can be applied to other corridors, although its classification performance is not as good as it is on the original corridor.

Logistic-regression models that use data from a smaller section surrounding a crash location (one VDS) transfer better to nearby corridors than three-VDS models. One possible reason is that including traffic data from a larger segment leads to crash risk being influenced by variability in geometric factors. Over a smaller segment, the geometrical factors do not vary as widely, enabling better model performance on corridors that may have different geometric design.

As logistic-regression models include more and more VDS locations, their classification accuracy increases for the freeway segment from which they were estimated. However, they perform worse when applied to nearby freeways than models that use data from fewer VDSs. Our modeling with weekday-only data did not change the classification results in any significant way, hence the proposed models use data for all crashes. Classification-tree models have classification accuracy comparable to that of logistic-regression models. The U.S. 101 NB classification-tree model was a more accurate predictor of I-880 SB crashes than of crashes on the other two roadways examined, but it was not nearly as accurate on I-880 SB as it was for the U.S. 101 NB crashes (as was also the case for the logistic-regression models).

FUTURE WORK

This research used random generation of both times and locations to generate non-crash events. To reduce variability in modeling, consideration should be given to using actual crash locations and then randomizing times. Also, using a lasso (instead of stepwise) selection procedure for logistic regression has been suggested to reduce bias in the coefficient estimates.

In this study, all detectable incidents were treated alike in the modeling procedure. Future work might attempt to analyze incidents in terms of intensity, measured, for example, by the number of lanes closed, incident duration, and resulting effects on traffic flow.

APPENDIX A: SAMPLE CODE

BUILD MODELS FROM 101 NB CRASH AND NON-CRASH DATA

```
proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_0      AVDS1_0SVUS1_0          SVDS1_0 AOUS1_0
        AODS1_0 SOUS1_0        SODS1_0

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_0_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_0_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_1      AVDS1_1SVUS1_1          SVDS1_1 AOUS1_1
        AODS1_1 SOUS1_1        SODS1_1

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_1_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_1_vo;

by descending IP_1;
```

```
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_2      AVDS1_2SVUS1_2      SVDS1_2 AOUS1_2
        AODS1_2 SOUS1_2      SODS1_2

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_2_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_2_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_3      AVDS1_3SVUS1_3      SVDS1_3 AOUS1_3
        AODS1_3 SOUS1_3      SODS1_3

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_3_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_3_vo;

by descending IP_1;

run;
```

```
proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_4      AVDS1_4SVUS1_4      SVDS1_4 AOUS1_4
      AODS1_4 SOUS1_4      SODS1_4

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_4_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_4_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_0      AVDS1_0SVUS1_0      SVDS1_0 AOUS1_0
      AODS1_0 SOUS1_0      SODS1_0 AVUS2_0      AVDS2_0 SVUS2_0
      SVDS2_0 AOUS2_0      AODS2_0 SOUS2_0      SODS2_0

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_0_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_0_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
```

```
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_1      AVDS1_1SVUS1_1        SVDS1_1 AOUS1_1
      AODS1_1 SOUS1_1           SODS1_1 AVUS2_1       AVDS2_1 SVUS2_1
      SVDS2_1 AOUS2_1           AODS2_1 SOUS2_1       SODS2_1

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_1_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_1_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_2      AVDS1_2SVUS1_2        SVDS1_2 AOUS1_2
      AODS1_2 SOUS1_2           SODS1_2 AVUS2_2       AVDS2_2 SVUS2_2
      SVDS2_2 AOUS2_2           AODS2_2 SOUS2_2       SODS2_2

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_2_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_2_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
```

```
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_3      AVDS1_3SVUS1_3        SVDS1_3 AOUS1_3
        AODS1_3 SOUS1_3         SODS1_3 AVUS2_3       AVDS2_3 SVUS2_3
        SVDS2_3 AOUS2_3         AODS2_3 SOUS2_3       SODS2_3

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_3_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_3_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_4      AVDS1_4SVUS1_4        SVDS1_4 AOUS1_4
        AODS1_4 SOUS1_4         SODS1_4 AVUS2_4       AVDS2_4 SVUS2_4
        SVDS2_4 AOUS2_4         AODS2_4 SOUS2_4       SODS2_4

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_4_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_4_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;
```

```
where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

    model y(event='1')=AVUS1_0     AVDS1_0 SVUS1_0        SVDS1_0 AOUS1_0
        AODS1_0     SOUS1_0     SODS1_0 AVUS2_0        AVDS2_0 SVUS2_0
        SVDS2_0 AOUS2_0         AODS2_0            SOUS2_0
        SODS2_0 AVUS3_0         AVDS3_0 SVUS3_0        SVDS3_0 AOUS3_0
        AODS3_0           SOUS3_0            SODS3_0

        / selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_0_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_0_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

    model y(event='1')=AVUS1_1     AVDS1_1 SVUS1_1        SVDS1_1 AOUS1_1
        AODS1_1     SOUS1_1     SODS1_1 AVUS2_1        AVDS2_1 SVUS2_1
        SVDS2_1 AOUS2_1         AODS2_1     SOUS2_1     SODS2_1 AVUS3_1
        AVDS3_1 SVUS3_1         SVDS3_1 AOUS3_1        AODS3_1
        SOUS3_1     SODS3_1

        / selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_1_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_1_vo;

by descending IP_1;
```

```
run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

        model y(event='1')=AVUS1_2       AVDS1_2 SVUS1_2        SVDS1_2 AOUS1_2
                AODS1_2        SOUS1_2    SODS1_2 AVUS2_2        AVDS2_2 SVUS2_2
                SVDS2_2 AOUS2_2           AODS2_2        SOUS2_2 SODS2_2 AVUS3_2
                AVDS3_2 SVUS3_2           SVDS3_2 AOUS3_2        AODS3_2
                SOUS3_2        SODS3_2

        / selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_2_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_2_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

        model y(event='1')=AVUS1_3       AVDS1_3 SVUS1_3        SVDS1_3 AOUS1_3
                AODS1_3        SOUS1_3    SODS1_3 AVUS2_3        AVDS2_3 SVUS2_3
                SVDS2_3 AOUS2_3           AODS2_3        SOUS2_3 SODS2_3 AVUS3_3
                AVDS3_3 SVUS3_3           SVDS3_3 AOUS3_3        AODS3_3
                SOUS3_3        SODS3_3

        / selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_3_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;
```

```
proc sort data=dayonly.pred3_3_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

        model y(event='1')=AVUS1_4      AVDS1_4 SVUS1_4      SVDS1_4 AOUS1_4
            AODS1_4     SOUS1_4     SODS1_4 AVUS2_4      AVDS2_4 SVUS2_4
            SVDS2_4 AOUS2_4      AODS2_4     SOUS2_4     SODS2_4 AVUS3_4
            AVDS3_4 SVUS3_4      SVDS3_4 AOUS3_4      AODS3_4
            SOUS3_4     SODS3_4

        / selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_4_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_4_vo;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_0      ASDS1_0SSUS1_0          SSDS1_0

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_0_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;
```

```
proc sort data=dayonly.pred1_0_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_1     ASDS1_1SSUS1_1         SSDS1_1

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_1_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_1_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_2     ASDS1_2SSUS1_2         SSDS1_2

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_2_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_2_s;

by descending IP_1;

run;
```

```
proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_3     ASDS1_3SSUS1_3        SSDS1_3

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_3_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_3_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_4     ASDS1_4SSUS1_4        SSDS1_4

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred1_4_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred1_4_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_0     ASDS1_0SSUS1_0        SSDS1_0 ASUS2_0
```

```
        ASDS2_0     SSUS2_0     SSDS2_0

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_0_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_0_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_1     ASDS1_1SSUS1_1          SSDS1_1 ASUS2_1
        ASDS2_1     SSUS2_1     SSDS2_1

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_1_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_1_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_2     ASDS1_2SSUS1_2          SSDS1_2 ASUS2_2
        ASDS2_2     SSUS2_2     SSDS2_2
```

```
/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_2_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_2_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_3    ASDS1_3SSUS1_3        SSDS1_3 ASUS2_3
        ASDS2_3    SSUS2_3    SSDS2_3

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_3_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_3_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_4    ASDS1_4SSUS1_4        SSDS1_4 ASUS2_4
        ASDS2_4    SSUS2_4    SSDS2_4

/ selection=stepwise
```

```
slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred2_4_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred2_4_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_0    ASDS1_0SSUS1_0        SSDS1_0 ASUS2_0
      ASDS2_0    SSUS2_0    SSDS2_0 ASUS3_0        ASDS3_0 SSUS3_0
      SSDS3_0

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_0_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_0_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_1    ASDS1_1SSUS1_1        SSDS1_1 ASUS2_1
      ASDS2_1    SSUS2_1    SSDS2_1 ASUS3_1        ASDS3_1SSUS3_1
      SSDS3_1

/ selection=stepwise
```

```
slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_1_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_1_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_2     ASDS1_2 SSUS1_2        SSDS1_2 ASUS2_2
        ASDS2_2    SSUS2_2    SSDS2_2 ASUS3_2        ASDS3_2 SSUS3_2
        SSDS3_2

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_2_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_2_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_3     ASDS1_3 SSUS1_3        SSDS1_3 ASUS2_3
        ASDS2_3    SSUS2_3    SSDS2_3 ASUS3_3        ASDS3_3 SSUS3_3
        SSDS3_3

/ selection=stepwise
```

```
slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_3_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_3_s;

by descending IP_1;

run;

proc logistic data=sas_sjsu.us101nb_crash_nocrash;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=ASUS1_4    ASDS1_4SSUS1_4        SSDS1_4 ASUS2_4
        ASDS2_4    SSUS2_4    SSDS2_4 ASUS3_4        ASDS3_4SSUS3_4
        SSDS3_4

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=dayonly.pred3_4_s p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

proc sort data=dayonly.pred3_4_s;

by descending IP_1;

run;
```

COMPARE MODELS TO FIND BEST THREE

```
%inc "E:\code\gainlift_mac.sas";

ods graphics on;

%GainLift(data=dayonly.pred1_0_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
```

```
event=1,out=dayonly.pctile_pred1_0_vo);

datadayonly.pctile_pred1_0_vo; set dayonly.pctile_pred1_0_vo; modelname='pred1_0_
vo'; run;

%GainLift(data=dayonly.pred1_1_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_1_vo);

datadayonly.pctile_pred1_1_vo; set dayonly.pctile_pred1_1_vo; modelname='pred1_1_
vo'; run;

%GainLift(data=dayonly.pred1_2_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_2_vo);

datadayonly.pctile_pred1_2_vo; set dayonly.pctile_pred1_2_vo; modelname='pred1_2_
vo'; run;

%GainLift(data=dayonly.pred1_3_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_3_vo);

datadayonly.pctile_pred1_3_vo; set dayonly.pctile_pred1_3_vo; modelname='pred1_3_
vo'; run;

%GainLift(data=dayonly.pred1_4_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_4_vo);

datadayonly.pctile_pred1_4_vo; set dayonly.pctile_pred1_4_vo; modelname='pred1_4_
vo'; run;

%GainLift(data=dayonly.pred2_0_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_0_vo);

datadayonly.pctile_pred2_0_vo; set dayonly.pctile_pred2_0_vo; modelname='pred2_0_
vo'; run;

%GainLift(data=dayonly.pred2_1_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_1_vo);

datadayonly.pctile_pred2_1_vo; set dayonly.pctile_pred2_1_vo; modelname='pred2_1_
vo'; run;

%GainLift(data=dayonly.pred2_2_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_2_vo);

datadayonly.pctile_pred2_2_vo; set dayonly.pctile_pred2_2_vo; modelname='pred2_2_
vo'; run;

%GainLift(data=dayonly.pred2_3_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_3_vo);

datadayonly.pctile_pred2_3_vo; set dayonly.pctile_pred2_3_vo; modelname='pred2_3_
vo'; run;

%GainLift(data=dayonly.pred2_4_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_4_vo);
```

```
datadayonly.pctile_pred2_4_vo; set dayonly.pctile_pred2_4_vo; modelname='pred2_4_
vo'; run;

%GainLift(data=dayonly.pred3_0_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_0_vo);

datadayonly.pctile_pred3_0_vo; set dayonly.pctile_pred3_0_vo; modelname='pred3_0_
vo'; run;

%GainLift(data=dayonly.pred3_1_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_1_vo);

datadayonly.pctile_pred3_1_vo; set dayonly.pctile_pred3_1_vo; modelname='pred3_1_
vo'; run;

%GainLift(data=dayonly.pred3_2_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_2_vo);

datadayonly.pctile_pred3_2_vo; set dayonly.pctile_pred3_2_vo; modelname='pred3_2_
vo'; run;

%GainLift(data=dayonly.pred3_3_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_3_vo);

datadayonly.pctile_pred3_3_vo; set dayonly.pctile_pred3_3_vo; modelname='pred3_3_
vo'; run;

%GainLift(data=dayonly.pred3_4_vo, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_4_vo);

datadayonly.pctile_pred3_4_vo; set dayonly.pctile_pred3_4_vo; modelname='pred3_4_
vo'; run;

%GainLift(data=dayonly.pred1_0_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_0_s);

datadayonly.pctile_pred1_0_s; set dayonly.pctile_pred1_0_s; modelname='pred1_0_s';
run;

%GainLift(data=dayonly.pred1_1_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_1_s);

datadayonly.pctile_pred1_1_s; set dayonly.pctile_pred1_1_s; modelname='pred1_1_s';
run;

%GainLift(data=dayonly.pred1_2_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_2_s);

datadayonly.pctile_pred1_2_s; set dayonly.pctile_pred1_2_s; modelname='pred1_2_s';
run;

%GainLift(data=dayonly.pred1_3_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_3_s);

datadayonly.pctile_pred1_3_s; set dayonly.pctile_pred1_3_s; modelname='pred1_3_s';
```

```
run;

%GainLift(data=dayonly.pred1_4_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred1_4_s);

datadayonly.pctile_pred1_4_s; set dayonly.pctile_pred1_4_s; modelname='pred1_4_s';
run;

%GainLift(data=dayonly.pred2_0_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_0_s);

datadayonly.pctile_pred2_0_s; set dayonly.pctile_pred2_0_s; modelname='pred2_0_s';
run;

%GainLift(data=dayonly.pred2_1_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_1_s);

datadayonly.pctile_pred2_1_s; set dayonly.pctile_pred2_1_s; modelname='pred2_1_s';
run;

%GainLift(data=dayonly.pred2_2_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_2_s);

datadayonly.pctile_pred2_2_s; set dayonly.pctile_pred2_2_s; modelname='pred2_2_s';
run;

%GainLift(data=dayonly.pred2_3_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_3_s);

datadayonly.pctile_pred2_3_s; set dayonly.pctile_pred2_3_s; modelname='pred2_3_s';
run;

%GainLift(data=dayonly.pred2_4_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred2_4_s);

datadayonly.pctile_pred2_4_s; set dayonly.pctile_pred2_4_s; modelname='pred2_4_s';
run;

%GainLift(data=dayonly.pred3_0_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_0_s);

datadayonly.pctile_pred3_0_s; set dayonly.pctile_pred3_0_s; modelname='pred3_0_s';
run;

%GainLift(data=dayonly.pred3_1_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_1_s);

datadayonly.pctile_pred3_1_s; set dayonly.pctile_pred3_1_s; modelname='pred3_1_s';
run;

%GainLift(data=dayonly.pred3_2_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_2_s);

datadayonly.pctile_pred3_2_s; set dayonly.pctile_pred3_2_s; modelname='pred3_2_s';
run;
```

```
%GainLift(data=dayonly.pred3_3_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_3_s);

datadayonly.pctile_pred3_3_s; set dayonly.pctile_pred3_3_s; modelname='pred3_3_s';
run;

%GainLift(data=dayonly.pred3_4_s, groups=10, oneplot=CCAPT , response=y, p=IP_1,
event=1,out=dayonly.pctile_pred3_4_s);

datadayonly.pctile_pred3_4_s; set dayonly.pctile_pred3_4_s; modelname='pred3_4_s';
run;

datadayonly.final_compare;

set dayonly.pctile_pred1_0_s

dayonly.pctile_pred1_1_s

dayonly.pctile_pred1_2_s

dayonly.pctile_pred1_3_s

dayonly.pctile_pred1_4_s

dayonly.pctile_pred2_0_s

dayonly.pctile_pred2_1_s

dayonly.pctile_pred2_2_s

dayonly.pctile_pred2_3_s

dayonly.pctile_pred2_4_s

dayonly.pctile_pred3_0_s

dayonly.pctile_pred3_1_s

dayonly.pctile_pred3_2_s

dayonly.pctile_pred3_3_s

dayonly.pctile_pred3_4_s

dayonly.pctile_pred1_0_vo

dayonly.pctile_pred1_1_vo

dayonly.pctile_pred1_2_vo

dayonly.pctile_pred1_3_vo

dayonly.pctile_pred1_4_vo

dayonly.pctile_pred2_0_vo

dayonly.pctile_pred2_1_vo

dayonly.pctile_pred2_2_vo
```

dayonly.pctile_pred2_3_vo

dayonly.pctile_pred2_4_vo

dayonly.pctile_pred3_0_vo

dayonly.pctile_pred3_1_vo

dayonly.pctile_pred3_2_vo

dayonly.pctile_pred3_3_vo

dayonly.pctile_pred3_4_vo;

run;

procgplot data=dayonly.final_compare;

whereSelectedPct=30;

plotCumPctCaptured*modelname;

run;

SCORING US-101 SB AND I-880 DATA FOR BEST 1 VDS MODEL

proc logistic data=SAS_SJSU.us101nb_crash_nocrashoutmodel=results2.pred1_4_vo_ model;

where (1<weekday(crashday)<7 and 18000<=crashtime<=79200);

model y(event='1')=AVUS1_4   AVDS1_4SVUS1_4 SVDS1_4 AOUS1_4 AODS1_4 SOUS1_4   SODS1_4

/ selection=stepwise

slentry=0.3

slstay=0.35

details

lackfit;

output out=results2.pred1_4_vo p=phat lower=lcl upper=ucl

predprob=(individual crossvalidate);

run;

/*pred1_2 name convention for the input to the model*/

proc logistic inmodel=results2.pred1_4_vo_model;

score data=sas_sjsu.crash_nocrash_us101sb out=results2.us101sb_pred1_4_vo;

run;

```
proc logistic inmodel=results2.pred1_4_vo_model;

score data=sas_sjsu.crash_nocrash_880nb out=results2.i880nb_pred1_4_vo;

run;

proc logistic inmodel=results2.pred1_4_vo_model;

score data=sas_sjsu.crash_nocrash_880sb out=results2.i880sb_pred1_4_vo;

run;
```

COMPARING BEST MODELS FOR EACH DATASET

```
%inc "E:\code\gainlift_mac.sas";

ods graphics on;

%GainLift(data=results2.us101sb_pred1_4_vo, groups=10, oneplot=CCAPT ,
response=y, p=P_1, event=1,out=results2.pctile_us101sb_pred1_4_vo);

dataresults2.pctile_us101sb_pred1_4_vo; set results2.pctile_us101sb_pred1_4_vo;
modelname='us101sb_pred1_4_vo'; run;

%GainLift(data=results2.i880nb_pred1_4_vo, groups=10, oneplot=CCAPT , response=y,
p=P_1, event=1,out=results2.pctile_i880nb_pred1_4_vo);

dataresults2.pctile_i880nb_pred1_4_vo; set results2.pctile_i880nb_pred1_4_vo;
modelname='i880nb_pred1_4_vo'; run;

%GainLift(data=results2.i880sb_pred1_4_vo, groups=10, oneplot=CCAPT , response=y,
p=P_1, event=1,out=results2.pctile_i880sb_pred1_4_vo);

dataresults2.pctile_i880sb_pred1_4_vo; set results2.pctile_i880sb_pred1_4_vo;
modelname='i880sb_pred1_4_vo'; run;

%GainLift(data=results2.us101sb_pred2_1_vo, groups=10, oneplot=CCAPT ,
response=y, p=P_1, event=1,out=results2.pctile_us101sb_pred2_1_vo);

dataresults2.pctile_us101sb_pred2_1_vo; set results2.pctile_us101sb_pred2_1_vo;
modelname='us101sb_pred2_1_vo'; run;

%GainLift(data=results2.i880nb_pred2_1_vo, groups=10, oneplot=CCAPT , response=y,
p=P_1, event=1,out=results2.pctile_i880nb_pred2_1_vo);

dataresults2.pctile_i880nb_pred2_1_vo; set results2.pctile_i880nb_pred2_1_vo;
modelname='i880nb_pred2_1_vo'; run;

%GainLift(data=results2.i880sb_pred2_1_vo, groups=10, oneplot=CCAPT , response=y,
p=P_1, event=1,out=results2.pctile_i880sb_pred2_1_vo);

dataresults2.pctile_i880sb_pred2_1_vo; set results2.pctile_i880sb_pred2_1_vo;
modelname='i880sb_pred2_1_vo'; run;

%GainLift(data=results2.us101sb_pred3_1_vo, groups=10, oneplot=CCAPT ,
```

```
response=y, p=P_1, event=1,out=results2.pctile_us101sb_pred3_1_vo);

dataresults2.pctile_us101sb_pred3_1_vo; set results2.pctile_us101sb_pred3_1_vo;
modelname='us101sb_pred3_1_vo'; run;

%GainLift(data=results2.i880nb_pred3_1_vo, groups=10, oneplot=CCAPT , response=y,
p=P_1, event=1,out=results2.pctile_i880nb_pred3_1_vo);

dataresults2.pctile_i880nb_pred3_1_vo; set results2.pctile_i880nb_pred3_1_vo;
modelname='i880nb_pred3_1_vo'; run;

%GainLift(data=results2.i880sb_pred3_1_vo, groups=10, oneplot=CCAPT , response=y,
p=P_1, event=1,out=results2.pctile_i880sb_pred3_1_vo);

dataresults2.pctile_i880sb_pred3_1_vo; set results2.pctile_i880sb_pred3_1_vo;
modelname='i880sb_pred3_1_vo'; run;

data results2.final_compare_3_best;

set results2.pctile_us101sb_pred1_4_vo

results2.pctile_i880nb_pred1_4_vo

results2.pctile_i880sb_pred1_4_vo

results2.pctile_us101sb_pred2_1_vo

results2.pctile_i880nb_pred2_1_vo

results2.pctile_i880sb_pred2_1_vo

results2.pctile_us101sb_pred3_1_vo

results2.pctile_i880nb_pred3_1_vo

results2.pctile_i880sb_pred3_1_vo;

run;

procgplot data=results2.final_compare_3_best;

whereSelectedPct=30;

plotCumPctCaptured*modelname;

run;
```

BEST CLASSIFICATION TREE MODEL RULES

```
IF  ASDS2_3  <      13.64

AND 27.750069233 <= SSDS2_3

THEN

 NODE  :    6
```

N : 29

0 : 79.3%

1 : 20.7%

IF 34.787389945<= SSDS1_3

AND 13.64 <= ASDS2_3

AND 27.750069233 <= SSDS2_3

THEN

NODE : 13

N : 590

0 : 98.8%

1 : 1.2%

IF 14.902777778<= ASUS2_3

AND SSDS2_3 < 3.2564497325

AND ASDS2_3 < 62.4125

THEN

NODE : 15

N : 644

0 : 94.6%

1 : 5.4%

IF 62.4125 <= ASDS2_3 < 76.825

AND SSDS2_3 < 5.2618221829

THEN

NODE : 18

N : 123

0 : 85.4%

1 : 14.6%

IF 76.825 <= ASDS2_3

AND SSDS2_3 < 5.2618221829

THEN

NODE : 19

N : 85

```
 0     :  65.9%

 1     :  34.1%

IF  46.647222222<= ASUS2_3

AND SSDS1_3  < 34.787389945

AND        13.64 <= ASDS2_3

AND 27.750069233 <= SSDS2_3

THEN

 NODE   :    23

 N     :   1017

 0     :  95.8%

 1     :   4.2%

IF  SSUS1_3  < 30.757920412

AND ASUS2_3  < 14.902777778

AND SSDS2_3  < 3.2564497325

AND ASDS2_3  <     62.4125

THEN

 NODE   :    26

 N     :    191

 0     :  77.0%

 1     :  23.0%

IF  30.757920412<= SSUS1_3

AND ASUS2_3  < 14.902777778

AND SSDS2_3  < 3.2564497325

AND ASDS2_3  <     62.4125

THEN

 NODE   :    27

 N     :    120

 0     :  92.5%

 1     :   7.5%

IF      51.675 <= ASDS2_3  <     62.4125

AND 21.013899321 <= SSDS2_3  < 27.750069233
```

THEN

  NODE   :   33

  N   :   327

  0   : 90.5%

  1   :   9.5%

IF SSDS1_3 < 6.7492863341

AND 16.562533892 <= SSUS1_3

AND 5.2618221829 <= SSDS2_3 < 27.750069233

AND   62.4125 <= ASDS2_3

THEN

  NODE   :   38

  N   :   252

  0   : 99.6%

  1   :   0.4%

IF ASDS1_3 <   34.65

AND ASUS2_3 < 46.647222222

AND SSDS1_3 < 34.787389945

AND   13.64 <= ASDS2_3

AND 27.750069233 <= SSDS2_3

THEN

  NODE   :   40

  N   :   548

  0   : 94.0%

  1   :   6.0%

IF ASDS2_3 <   11.6375

AND ASUS1_3 <   31.6375

AND 3.2564497325 <= SSDS2_3 < 21.013899321

THEN

  NODE   :   50

  N   :   41

```
  0    :  78.0%

  1    :  22.0%

IF      11.6375 <= ASDS2_3 <      62.4125

AND ASUS1_3 <      31.6375

AND 3.2564497325 <= SSDS2_3 < 21.013899321

THEN

 NODE   :    51

 N    :    125

  0    :  48.0%

  1    :  52.0%

IF  ASDS1_3 <      25.185

AND     31.6375 <= ASUS1_3

AND 3.2564497325 <= SSDS2_3 < 21.013899321

AND ASDS2_3 <      62.4125

THEN

 NODE   :    52

 N    :    67

  0    :  89.6%

  1    :  10.4%

IF      25.185 <= ASDS1_3

AND     31.6375 <= ASUS1_3

AND 3.2564497325 <= SSDS2_3 < 21.013899321

AND ASDS2_3 <      62.4125

THEN

 NODE   :    53

 N    :    228

  0    :  71.9%

  1    :  28.1%

IF  ASDS2_3 <      20.2625

AND 21.013899321 <= SSDS2_3 < 27.750069233

THEN
```

```
 NODE   :    54

 N    :    63

 0    :  95.2%

 1    :   4.8%

IF     20.2625 <= ASDS2_3  <      51.675

AND 21.013899321 <= SSDS2_3  < 27.750069233

THEN

 NODE   :    55

 N    :    124

 0    :  71.8%

 1    :  28.2%

IF  SSUS1_3  < 5.4405194116

AND ASUS1_3  < 66.672222222

AND 5.2618221829 <= SSDS2_3  < 27.750069233

AND     62.4125 <= ASDS2_3

THEN

 NODE   :    58

 N    :    449

 0    :  89.5%

 1    :  10.5%

IF  5.4405194116<= SSUS1_3  < 16.562533892

AND ASUS1_3  < 66.672222222

AND 5.2618221829 <= SSDS2_3  < 27.750069233

AND     62.4125 <= ASDS2_3

THEN

 NODE   :    59

 N    :    339

 0    :  77.9%

 1    :  22.1%

IF  SSDS1_3  < 20.613077874

AND 66.672222222 <= ASUS1_3
```

```
AND SSUS1_3  < 16.562533892

AND 5.2618221829 <= SSDS2_3  < 27.750069233

AND     62.4125 <= ASDS2_3

THEN

 NODE   :    60

 N     :   1586

 0     :  93.4%

 1     :   6.6%

IF  20.613077874<= SSDS1_3

AND 66.672222222 <= ASUS1_3

AND SSUS1_3  < 16.562533892

AND 5.2618221829 <= SSDS2_3  < 27.750069233

AND     62.4125 <= ASDS2_3

THEN

 NODE   :    61

 N     :    433

 0     :  86.4%

 1     :  13.6%

IF  5.2618221829<= SSDS2_3  < 5.7513568455

AND 6.7492863341 <= SSDS1_3

AND 16.562533892 <= SSUS1_3

AND     62.4125 <= ASDS2_3

THEN

 NODE   :    64

 N     :    20

 0     :  80.0%

 1     :  20.0%

IF  5.7513568455<= SSDS2_3  < 27.750069233

AND 6.7492863341 <= SSDS1_3

AND 16.562533892 <= SSUS1_3

AND     62.4125 <= ASDS2_3
```

```
THEN

 NODE   :    65

 N      :   1074

 0      :  95.0%

 1      :   5.0%

IF      13.64 <= ASDS2_3  <     39.7875

AND       34.65 <= ASDS1_3

AND ASUS2_3  < 46.647222222

AND SSDS1_3  < 34.787389945

AND 27.750069233 <= SSDS2_3

THEN

 NODE   :    68

 N      :     66

 0      :  80.3%

 1      :  19.7%

IF      39.7875 <= ASDS2_3

AND       34.65 <= ASDS1_3

AND ASUS2_3  < 46.647222222

AND SSDS1_3  < 34.787389945

AND 27.750069233 <= SSDS2_3

THEN

 NODE   :    69

 N      :     85

 0      :  94.1%

 1      :   5.9%
```

ABBREVIATIONS AND ACRONYMS

ART2	Adaptive Resonance Theory 2
AADT	Annual Average Daily Traffic
ANN	Artificial Neural Network Model
XML	Extensible Markup Language
FITS	Flow Impacts on Traffic Safety
ART	Adaptive Resonance Theory
HOV	High-occupancy Vehicle
ID	Identification
ITS	Intelligent Transportation System
MLF	Multilayer Feed Forward
MLP	Multilayer Perceptron
NB	Northbound
NLCCA	Non-linear (Non-Parametric) Canonical Correlation Analysis
PeMS	Performance Measurement System
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
PNN	Probabilistic Neural Network
RBF	Radial Basis Function
RTMC	Regional Transportation Management Center
RMSE	Root Mean Squared Error
SB	Southbound
SOFM	Self-organizing Feature Map
SOM	Self-organizing Map
SR	State Route
VSL	Variable Speed Limit
VDS	Vehicle-detector Station

BIBLIOGRAPHY

Abdel-Aty, M., and F. Abdalla. 2004. "Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes using generalized estimating equations for correlated data." Presented at the 83rd annual meeting of Transportation Research Board, Washington, D.C.

Abdel-Aty, Mohamed, and Anurag Pande. 2005. "Identifying crash propensity using specific traffic speed conditions." *Journal of Safety Research* 36 (1): 97–108.

Abdel-Aty, Mohamed, Nizam Uddin, and Anurag Pande. 2005. "Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways." *Transportation Research Record: Journal of the Transportation Research Board* 1908 (-1) (January 1): 51–58.

Abdel-Aty, Mohamed, Nizam Uddin, Anurag Pande, Fathy Abdalla, and Liang Hsia. 2004. "Predicting freeway crashes from loop detector data by matched case-control logistic regression." *Transportation Research Record: Journal of the Transportation Research Board* 1897 (-1) (January 1): 88–95.

Abdelwahab, Hassan, and Mohamed Abdel-Aty. 2001. "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections." *Transportation Research Record: Journal of the Transportation Research Board* 1746 (-1) (January 1): 6–13.

———. 2002. "Artificial neural networks and logit models for traffic safety analysis of toll plazas." *Transportation Research Record: Journal of the Transportation Research Board* 1784 (-1) (January 1): 115–125.

Abdulhai, Baher, and Stephen G. Ritchie. 1999. "Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network." *Transportation Research Part C: Emerging Technologies* 7 (5) (October): 261–280.

Al-Deek, H., A. Garib, and A. Radwan. 1998. A new method for estimating freeway incident congestion. Text. http://cat.inist.fr/?aModele=afficheN&cpsidt=2958605.

Al-Deek, H., S. Ishak, and A. Khan. 1996. Impact of freeway geometric and incident characteristics on incident detection. Text. http://cat.inist.fr/?aModele=afficheN&cpsidt=2477103.

Awad, Wael, and Bruce Janson. 1998. "Prediction models for truck accidents at freeway ramps in Washington State using regression and artificial intelligence techniques." *Transportation Research Record: Journal of the Transportation Research Board* 1635 (-1) (January 1): 30–36.

Breiman, L., J. H. Freidman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Cheu, Ruey L., and Stephen G. Ritchie. 1995. "Automated detection of lane-blocking freeway incidents using artificial neural networks." *Transportation Research Part C: Emerging Technologies* 3 (6) (December): 371–388.

Collett, D. 1991. *Modelling Binary Data*. Chapman and Hall.

Gayah, V. V., C. Dos Santos, M. Abdel-Aty, A. Dhindsa, and J. Dilmore. 2006. Evaluating ITS strategies for real-time freeway safety improvement. In *Intelligent Transportation Systems Conference, 2006*. IEEE, 1114–1119.

Golob, Thomas F., and Wilfred W. Recker. 2003. "Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions." *Journal of Transportation Engineering* 129 (4) (July): 342–353.

———. 2004. "A method for relating type of crash to traffic flow characteristics on urban freeways." *Transportation Research Part A: Policy and Practice* 38 (1) (January): 53–80.

Golob, Thomas F., Wilfred W. Recker, and Veronica M. Alvarez. 2004a. "Tool to evaluate safety effects of changes in freeway traffic flow." *Journal of Transportation Engineering* 130 (2) (March): 222–230.

———. 2004b. "Freeway safety as a function of traffic flow." *Accident Analysis & Prevention* 36 (6) (November): 933–946.

Hand, D. J., Heikki Mannila, and Padhraic Smyth. 2001. *Principles of data mining.* Cambridge, MA: MIT Press, August 1.

Hosner, D. W., and S. Lemeshow. 1989. *Applied Logistic Regression*. Wiley & Sons.

Hughes, R., and F. Council. 1999. "On establishing relationship(s) between freeway safety and peak period operations: Performance measurement and methodological considerations." Presented at the 78th annual meeting of Transportation Research Board, Washington, D.C.

Ishak, Sherif, and Haitham Al-Deek. 1999. "Performance of automatic ANN-based incident detection on freeways." *Journal of Transportation Engineering* 125 (4) (July): 281–290.

Ishak, S., and C. Alecsandru. 2005. "Analysis of freeway pre-incident, post-incident, and non-incident conditions using second-order spatio-temporal traffic performance measures." Presented at the 84th annual meeting of Transportation Research Board, Washington, D.C.

Kockelman, K., and J. Ma. 2004. "Freeway speeds and speed variations preceding crashes, within and across lanes." *Journal of the Transportation Research Forum* 46 (1): 43–62 (2007).

Lee, Chris, Bruce Hellinga, and Frank Saccomanno. 2003. "Real-time crash prediction model for application to crash prevention in freeway traffic." *Transportation Research Record: Journal of the Transportation Research Board* 1840 (-1) (January 1): 67–77.

———. 2004. "Assessing safety benefits of variable speed limits." *Transportation Research Record: Journal of the Transportation Research Board* 1897 (-1) (January 1): 183–190.

Lee, Chris, Frank Saccomanno, and Bruce Hellinga. 2002. "Analysis of crash precursors on instrumented freeways." *Transportation Research Record: Journal of the Transportation Research Board* 1784 (-1) (January 1): 1–8.

Madanat, S., and P. Liu. 1995. *A Prototype System For Real-Time Incident Likelihood Prediction*. IDEA Project Final Report (ITS-2). Washington, D.C.: Transportation Research Board. http://pubsindex.trb.org/view.aspx?id=465172.

Mussone, Lorenzo, Andrea Ferrari, and Marcello Oneta. 1999. "An analysis of urban collisions using an artificial intelligence model." *Accident Analysis & Prevention* 31 (6) (November): 705–718.

Nezamuddin, N., Nan Jiang, Jianming Ma, Ti Zhang, and S. Travis Waller. 2011. "Active traffic management strategies: implications for freeway operations and traffic safety." Presented at the 90th annual meeting of Transportation Research Board,Washington, D.C.

Oh, C., J. Oh, S. Ritchie, and M. Chang. 2001. "Real-time estimation of freeway accident likelihood." Presented at the 80th annual meeting of Transportation Research Board, Washington, D.C.

Pande, Anurag. 2003. *Classification of Real-Time Traffic Speed Patterns to Predict Crashes on Freeways*. Orlando, FL: University of Central Florida.

———. 2005. *Applying Hybrid Models for Real-Time Crash Risk Assessment on Freeways.* Orlando, FL: University of Central Florida.

Pande, Anurag, and Mohamed Abdel-Aty. 2006a. "Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways." *Transportation Research Record: Journal of the Transportation Research Board* 1953 (-1) (January 1): 31–40.

———. 2006b. "Assessment of freeway traffic parameters leading to lane-change related collisions." *Accident Analysis & Prevention* 38 (5) (September): 936–948.

———. 2007. "Multiple-model framework for assessment of real-time crash risk." *Transportation Research Record: Journal of the Transportation Research Board* 2019 (-1) (December 1): 99–107.

———. 2008. "A computing approach using probabilistic neural networks for instantaneous appraisal of rear-end crash risk." *Computer-Aided Civil and Infrastructure Engineering* 23 (7) (October): 549–559.

Pande, Anurag, Mohamed Abdel-Aty, and Liang Hsia. 2005. "Spatiotemporal variation of risk preceding crashes on freeways." *Transportation Research Record: Journal of the Transportation Research Board* 1908 (-1) (January 1): 26–36.

Park, S., and S. Ritchie. 2004. "Exploring the relationship between freeway speed variance, lane changing, and vehicle hetrogenity." Presented at the 83rd annual meeting of Transportation Research Board, Washington, D.C.

Pham, Minh-Hai, Nour-Eddin El Faouzi, and André-Gilles Dumont. 2011. "Real-time identification of risk-prone traffic patterns taking into account weather conditions". Presented at the 90th annual meeting of Transportation Research Board, Washington, D.C.

SAS Institute. 2001. *SAS*. Irvine, CA: SAS Institute.

Sayed, Tarek, and Walid Abdelwahab. 1998. "Comparison of fuzzy and neural classifiers for road accidents analysis." *Journal of Computing in Civil Engineering* 12 (1) (January): 42–47.

Sohn, S., and H. Shin. 2001. "Pattern recognition for road traffic accident severity in Korea." *Ergonomics* 44 (1): 107–117.

Songchitruksa, Praprut, and Kevin Balke. 2006. "Assessing weather, environment, and loop data for real-time freeway incident prediction." *Transportation Research Record: Journal of the Transportation Research Board* 1959 (-1) (January 1): 105–113.

Vorko, Ariana, and Franjo Jovic. 2000. "Multiple attribute entropy classification of school-age injuries." *Accident Analysis & Prevention* 32 (3) (May): 445–454.

Xu, Chengcheng, Liu Pan, Wei Wang, and Yu Chunjun. 2011. "Exploration and identification of hazardous traffic flow states before crash occurrences on freeways." Presented at the 90th annual meeting of Transportation Research Board, Washington, D.C.

Zhang, C., J. Ivan, W. El-Dessouki, and E. Anagnostou. 2005. "Relative risk analysis for studying the impact of adverse weather conditions and congestion on traffic accidents." Presented at the 84th annual meeting of Transportation Research Board, Washington, D.C.

Zhou, Min, and Virginia Sisiopiku. 1997. "Relationship between volume-to-capacity ratios and accident rates." *Transportation Research Record: Journal of the Transportation Research Board* 1581 (-1) (January 1): 47–52.

ABOUT THE AUTHORS

ANURAG PANDE, PH.D.

Anurag Pande, Ph.D., is an Assistant Professor of Civil Engineering at California Polytechnic State University, San Luis Obispo. In addition, he is the coordinator for the dual-degree program in civil engineering and city and regional planning. He is a Research Associate of the Mineta Transportation Institute.

CORNELIUS NUWORSOO, PH.D.

Cornelius Nuworsoo, Ph.D., is an Associate Professor of Transportation Planning at California Polytechnic State University, San Luis Obispo. He also serves as graduate programs coordinator in the Department of City and Regional Planning. He is a Research Associate of the Mineta Transportation Institute.

CAMERON SHEW

Cameron Shew is a Master's degree candidate in Civil and Environmental Engineering at California Polytechnic State University, San Luis Obispo. He is a certified engineer-in-training (EIT) and has interned with Fehr & Peers Transportation Consultants (June 2011–September 2011) and Kittelson & Associates, Inc. (June 2010–September 2010). His interests include transportation design and traffic simulation and safety.

PEER REVIEW

San José State University, of the California State University system, and the MTI Board of Trustees have agreed upon a peer review process required for all research published by MTI. The purpose of the review process is to ensure that the results presented are based upon a professionally acceptable research protocol.

Research projects begin with the approval of a scope of work by the sponsoring entities, with in-process reviews by the MTI Research Director and the Research Associated Policy Oversight Committee (RAPOC). Review of the draft research product is conducted by the Research Committee of the Board of Trustees and may include invited critiques from other professionals in the subject field. The review is based on the professional propriety of the research methodology.