# ATMS Implementation System for Identifying Traffic Conditions Leading to Potential Crashes

Mohamed Abdel-Aty and Anurag Pande

***Abstract*—Predicting a crash occurrence is the key to traffic safety. Real-time identification of freeway segments with high crash potential is addressed in this paper. For this study, historical crashes and corresponding traffic-surveillance data from loop detectors were gathered from a 36-mi corridor of Interstate 4 for 4 years. Following an exploratory analysis, two types of logistic-regression models (i.e., simple and multivariate) were developed. It was observed that, although the simple models have the advantage of being tolerant in their data requirements, their classification accuracy was inferior to that of the final multivariate model. Hence, the simple models were used to deduce time–space patterns of variation in crash risk while the multivariate model was chosen for final classification of traffic patterns. As a suggested application for the simple models, their output may be used for the preliminary assessment of the crash risk. If there is an indication of high crash risk, then the multivariate model may be employed to explicitly classify the data patterns as leading or not leading to a crash occurrence. A demonstration of this two-stage real-time application strategy, based on simple and multivariate models, is provided in the paper. The output from these model-processing real-time loop-detector data may be utilized by traffic-management authorities for developing proactive traffic-management strategies.**

***Index Terms*—Advanced traffic management, advanced traffic management system (ATMS), crash prediction, crash risk, real-time implementation.**

## I. Introduction

TRAFFIC safety studies are categorized into two groups according to Golob *et al.* [1]. The first group is called the aggregate studies, where units of analysis represent counts of crashes or crash rates for specific time periods (typically months or years) and for specific spaces (specific roads or networks) and the traffic flow is represented by parameters of the statistical distributions of traffic flow for similar time and space. Disaggregate studies belong to the second group, in which units of analysis are crashes themselves and the traffic flow is represented by parameters of the traffic flow at the time and the location of each crash.

Traditional traffic-safety literature [2], [3] has been more or less focused on crash frequency/rate estimation and hence belongs to the former category. However, the approach is not sufficient to "predict" crashes in real time using traffic-flow variables measured from loop detectors in an advanced traffic management system (ATMS) environment. There is a need to estimate models that use dynamic flow variables as inputs and determine whether or not they would lead to a crash occurrence. This approach belongs to the later category (i.e., the disaggregate studies), which are relatively new and are made possible by the proliferation of data collection and analysis capabilities in the field of intelligent transportation systems (ITS).

The essential premise of this approach involves identifying patterns in the traffic-surveillance data observed prior to historical crashes. The traffic-surveillance systems may then be enhanced to detect the identified patterns in real-time data. A reliable identification of such patterns could pave the way for developing proactive strategies to avoid crashes, such as, warning(s) to the motorists and variable speed limits. However, in this paper, the scope has been limited to show the potential of statistical models for reliable identification of these crash-prone conditions on the freeway.

These models would be a substantial advancement in the field of traffic management due to their potential contribution towards traffic safety as well as freeway operations. In this regard, a crash-prediction model was developed for the 13-mi central corridor of Interstate 4 in Orlando in one of our previous studies [4]. The model achieved satisfactory crash identification and demonstrated the feasibility of predicting crashes in real time. The model was developed using data from a small urban segment of the freeway with the crash data spanning a short period of time (8 mo).

For this study, the crash data were expanded to include 3755 crashes that occurred during a 4-year period (from 1999 through 2002) on a 36-mi instrumented corridor of Interstate 4 in Orlando metropolitan region. Out of these 3755 crashes, the corresponding loop data were available for 2046 crashes. A matched case-control dataset consisting of traffic data corresponding to the crash (case) and five matched noncrashes (controls) were created as per requirements of the analysis technique adopted. The idea of matched case-control analysis is to explore the effects of independent variables of interest on the binary outcome while controlling other confounding variables through the design of study. In the context of this research, crash versus noncrash is the binary outcome with traffic parameters being the independent variables. The design of the study controls the external factors such as geometric design of the freeway, time of the day, day of the week, etc. Simple (one covariate) and multivariate logistic-regression models were developed based on this attractive sampling technique. Based on the results from

Manuscript received February 17, 2005; revised July 14, 2005, September 23, 2005, and November 1, 2005. This work was supported in part by the Florida Department of Transportation (FDOT). The Associate Editor for this paper was B. K. Johnson.

The authors are with the Department of Civil and Environmental Engineering, University of Central Florida (UCF), Orlando, FL 32816 USA (e-mail: mabdel@mail.ucf.edu; anurag@mail.ucf.edu).

Digital Object Identifier 10.1109/TITS.2006.869612

these models, a two-stage implementation plan for the reliable real-time identification of crash-prone conditions is proposed.

## II. Background

The study by Hughes and Council [5] was among the first studies aiming at real-time preemptive crash prediction. The relationship between freeway safety and peak period operations was explored using loop-detector data. Traffic-flow consistency, as perceived by the drivers, was identified as one of the factors associated with a crash occurrence. Lee *et al.* [6] developed a log-linear model to predict crashes through the estimation of crash precursors from loop-detector data. In a later study by the same authors [7], the aforementioned model was refined. The coefficient of temporal variation in speed was shown to have a relatively longer term effect on the crash potential than the density, while the effect of average variation in speed across adjacent lanes was found to be insignificant.

Oh *et al.* [8] developed Bayesian classifiers to classify patterns as leading or not leading to crash and argued that a 5-min standard deviation of speed was the best indicator of "disruptive" traffic conditions leading to a crash as opposed to "normal" freeway traffic. In our previous study [9], we used a probabilistic neural network (PNN) as the classification algorithm and demonstrated the feasibility of "predicting" crashes at least 10 min in advance.

In some of the more detailed recent studies, Golob and Recker [10] and Golob *et al.* [1] concluded that the collision type is the best explained crash characteristic and that it is related to the median speed and left and interior-lane variations in speed. Moreover, it was observed that the severity of crashes tracks the inverse of the traffic volume and is influenced more by the volume than by the speed. Based on these results, in one of their later studies, Golob *et al.* [11] used loop data corresponding to more than 1000 crashes over six major freeways in Orange County, California, and developed a software tool called Flow Impacts on Traffic Safety (FITS) to forecast the type of crashes that are most likely to occur under the traffic conditions being monitored. A case study application of this tool on a section of SR-55 was also demonstrated. Findings from the aforementioned studies point towards the potential application of real-time traffic data in the field of traffic safety. However, crashes usually involve a complex interaction between traffic, geometric, and environmental factors. It is difficult to explicitly account for wide range of these factors in any of the modeling frameworks proposed by the aforementioned studies.

In one of our earlier studies [4], we argued that the accuracy of real-time crash identification may be increased if the model utilizes information on traffic-flow characteristics for both crash and noncrash cases while controlling other external factors (thereby implicitly accounting for factors such as the geometry and the location). This is known as matched case-control analysis, where each case refers to a crash and control refers to a noncrash case. The 5-min average occupancy measured upstream and the coefficient of variation in speed measured downstream of the crash location were identified to be the most significant crash precursors in the study. The logistic-regression model developed using these two parameters as inputs achieved satisfactory classification accuracy [4].

Despite this attractive modeling approach, the study was limited in scope due to insufficient data. Only 8 mo worth of crash data were collected for a small largely urban corridor. Due to largely uniform traffic and crash characteristics on the freeway segment, the transferability of the model remained suspect. In this study, the database has been expanded to include crashes spanning 4 years on a 36-mi freeway corridor. Moreover, a two-stage online application strategy has been proposed in order to identify real-time "black spots" on the freeway corridor under consideration.

## III. Methodology

The purpose of the proposed matched crash–noncrash analysis is to explore the effects of traffic-flow variables while controlling for the effects of other confounding variables through the design of the study. In this section, a brief description of sampling and modeling methodologies is provided in the context of the present research problem.

### A. Sampling Technique

Under a matched crash–noncrash study design, all crashes are selected first. For each crash, parameters such as location, time of day, day of the week, etc., associated with it are selected as matching factors. A subpopulation of noncrashes is then identified using these matching factors. For example, for a crash at certain freeway location on a Monday, a subpopulation of noncrash cases would consist of observations on traffic-flow variables obtained from the same location at the same time but over all other Mondays of the same year. A total of $m$ noncrash cases are then selected at random from each subpopulation of noncrash cases. The $m+1$ observations (1 crash and $m$ noncrash cases) form one stratum. Within the stratum, differences between crash and noncrash traffic characteristics may then be utilized for estimation of statistical model(s) for the binary target. This is accomplished under the conditional likelihood principle of the statistical theory.

### B. Modeling Technique

Suppose there are $N$ strata with one crash and $m$ noncrash cases in stratum $j$, $j = 1, 2, \ldots, N$. Let the probability of the $i$th observation in the $j$th stratum being a crash be $p_j(x_{ij})$; where $x_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{kij})$ is the vector of $k$ traffic-flow variables $x_1, x_2, \ldots, x_k$; $i = 0, 1, 2, \ldots, m$; and $j = 1, 2, \ldots, N$. The probability $p_j(x_{ij})$ may be modeled using a linear logistic-regression model as follows:

$$\text{logit}\left(p_j(x_{ij})\right) = \alpha_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij}. \quad (1)$$

Note that the intercept term would be different for different strata. It summarizes the effect of parameters used to form the strata on the probability of a crash occurrence. In order to account for the stratification in the analysis, one may construct

a conditional likelihood. This conditional likelihood function is the product of $N$ terms, each of which is the conditional probability that the crash in the $j$th strata is the observation that involves a vector of explanatory variables $x_{0j}$ where $x_{0j}, x_{1j}, \ldots, x_{mj}$ are the vectors of explanatory variables in the $j$th stratum. The mathematical derivation of the relevant likelihood function is quite complex and is omitted here. The reader may consult [12] for a full derivation of the conditional likelihood function that can be expressed as

$$L(\beta) = \prod_{j=1}^{N} \left[ 1 + \sum_{i=1}^{m} \exp \left\{ \sum_{u=1}^{k} \beta_u (x_{uij} - x_{u0j}) \right\} \right]^{-1} \quad (2)$$

where parameters $\beta$ are the same as in (1). The likelihood function $L(\beta)$ is independent of the intercept terms $\alpha_1, \alpha_2, \ldots, \alpha_N$. Therefore, the effects of matching variables cannot be estimated, and (1) cannot be used to estimate crash probabilities. However, note that the values of the $\beta$ parameters that would maximize the likelihood function [given by (2)] are also estimates of $\beta$ coefficients in (1). These estimates can be used to approximate the relative risk of a crash occurrence in terms of the log-odds ratio.

The odds ratios can also be used to classify observations under this matched case-control framework [13]. To see this, consider two observation vectors $x_{1j} = (x_{11j}, x_{21j}, \ldots, x_{k1j})$ and $x_{2j} = (x_{12j}, x_{22j}, \ldots, x_{k2j})$ from the $j$th strata. From (1), one may verify that the log-odds ratio of a crash occurrence due to traffic-flow vector $x_{1j}$ relative to vector $x_{2j}$ would be

$$\log \left\{ \frac{\frac{p(x_{1j})}{[1-p(x_{1j})]}}{\frac{p(x_{2j})}{[1-p(x_{2j})]}} \right\} = \beta_1 (x_{11j} - x_{12j}) + \beta_2 (x_{21j} - x_{22j}) + \cdots + \beta_k (x_{k1j} - x_{k2j}). \quad (3)$$

The right-hand side (RHS) of this log-odds ratio is independent of $\alpha_j$ and can be estimated using the estimates for $\beta$ coefficients. We may utilize the relative log-odds ratio [from (3)] for the classification of individual observations by replacing $x_{2j}$ with the vector of values for the traffic-flow variables representing "normal" traffic conditions in the $j$th stratum. One may conveniently use a simple average of all noncrash observations within the stratum for each variable. If we let $\overline{x}_{2j} = (\overline{x}_{12j}, \overline{x}_{22j}, \overline{x}_{32j}, \ldots, \overline{x}_{k2j})$ denote the vector of average values for the $k$ variables over $m$ noncrash cases within the $j$th stratum, then the log odds of a crash relative to a noncrash may be approximated by

$$\log \left\{ \frac{\frac{p(x_{1j})}{[1-p(x_{1j})]}}{\frac{p(\overline{x}_{2j})}{[1-p(\overline{x}_{2j})]}} \right\} = \beta_1 (x_{11j} - \overline{x}_{12j}) + \beta_2 (x_{21j} - \overline{x}_{22j}) + \cdots + \beta_p (x_{k1j} - \overline{x}_{k2j}). \quad (4)$$

The above log-odds ratio can then be used to separate "normal" conditions from crash-prone conditions by establishing an appropriate threshold.

## IV. Data Collection and Preparation

Crash data and the corresponding traffic data collected through underground sensors on Interstate 4 (I-4) are used in this study. These sensors record and archive following traffic-flow parameters every 30 s from three lanes of the freeway in each direction: average vehicle counts, average speed, and lane detector occupancy. These data are collected on I-4 from 69 stations spaced at approximately 1/2 mi on a 36-mi stretch of the freeway. The crash data were collected from the Florida Department of Transportation (FDOT) crash database for the years 1999–2002.

First, the location for each crash that occurred in the study area during this period was identified. For every crash, the loop-detector station nearest to its location was determined and referred to as the station of the crash. The precrash loop-detector data from stations surrounding the crash location were collected based on the adjusted time of historical crashes estimated through a shock wave and the rule-based methodology [14]. Traffic data were extracted for each crash and noncrash cases corresponding to each crash. The correspondence here means that, for example, if a crash occurred on April 12, 2002 (Monday) 6:00 P.M. on I-4 in the eastbound direction and the nearest loop detector was at station 30, data were extracted from station 30 at four-loops upstream and two-loops downstream of station 30 for a 30-min period prior to the estimated time of the crash for all Mondays of the year at the same time. Thus, this case will have a loop-data table consisting of the speed, the volume, and the lane-occupancy (percent of time the loop is occupied by vehicles) values for all three lanes from the loop stations 26–32 (on the eastbound direction) between 5:30–6:00 P.M. for all the Mondays of the year 2002 including the day of the crash (crash case). This matched sampling essentially controls for external factors affecting the crash occurrence such as driver population, time of day, day of week, location on the freeway, etc. (thus implicitly accounting for these factors). More details on this sampling technique, the application of the methodology, and data cleaning may be found in our earlier study [4].

The raw 30-s loop data have random noises and are difficult to work with in a modeling framework. Therefore, the 30-s raw data were combined in the forms of 5-min averages and standard deviations. Thus, the 30-min period was divided into six 5-min time slices (a 3-min aggregation was also attempted, but 5-min was preferred—see [14]). The series of seven stations was named "B"–"H," respectively, with "B" being farthest station upstream and so on. It should be noted that "F" would be the station closest to the location of the crash with "G" and "H" being the stations downstream of the crash location. Similarly, six 5-min time slices were denoted 1–6. The interval between the time of the crash and 5-min prior to the crash was named as time slice 1, the interval between 5–10-min prior to the crash as time slice 2, the interval between 10–15 min prior to the crash as time slice 3, and so on.

Using data from only the specific lane of the crash would have reduced the size of the dataset to about 30% of the original crash sample due to the fact that the loop data from the specific lane of the crash were often missing. Therefore, in the final dataset, averages and standard deviations were obtained using parameter (speed, volume, and occupancy) values over the three lanes. Hence, the averages (and standard deviations) at the 5-min level were calculated using 30 (3 lanes $*$ 5 min $*$ 2 observations/min) observations. Therefore, even if at a location where the loop detector from a certain lane was not reporting data, there would be observations (either 10 from one lane or 20 from two lanes) available to obtain a measure of the traffic conditions at that location. This not only increased the sample size to more than 2000 crashes but also helped in developing a system more robust for loop failures, since all three lanes at loop-detector stations are less likely to be simultaneously unavailable. Another advantage is that these aggregated measures not only capture temporal variations (or lack there of) of parameters on the freeway, but their variations across the three lanes as well. This issue has been investigated thoroughly in [14].

This dataset with 2046 matched strata included all types of crashes. The type of crash information available in the FDOT crash database was utilized to retain only multivehicle crashes. The ambient traffic characteristics are more likely to affect crashes involving interaction among vehicles rather than the single-vehicle crashes (which were removed from the dataset, since single-vehicle crashes are more likely to be caused by errors on the part of individual drivers). The resulting dataset had 1528 matched strata available for analysis. For each of the seven loop detectors (B–H) and six time slices (1–6) mentioned above, the values of means (AS, AV, AO) and standard deviations (SS, SV, SO) of speed, volume, and occupancy, respectively, were available for all crashes and the corresponding noncrash cases in every strata. Due to data availability issues, there were different numbers of controls (noncrash cases) for each case (crash). To carry out matched case–control analysis, a symmetric dataset was created (such that each crash case in the dataset has the same number of corresponding noncrash cases as controls) by randomly selecting five noncrash cases for each crash.

In addition to the aforementioned dataset, we also created a pseudo case-control dataset in which six random noncrash cases in each stratum were selected, and one of them was assigned as a (pseudo) crash while all the real crash cases were dropped. The results from this dataset were analyzed in order to delineate the differences between real and pseudo case-control datasets.

## V. Multivariate Logistic-Regression Model

### A. Data Analysis

From each of the seven loop detectors (B–H) and six time slices (1–6) mentioned above, the values of averages (AS, AV, AO) and standard deviations (SS, SV, SO) of speed, volume, and occupancy, respectively, were used one at a time as the a risk factor (i.e., the independent variable) in a logistic-regression model. In the logistic-regression setting, the output of these simple models would be the hazard ratio for the parameter used as a covariate in the model. The hazard ratio for an explanatory variable with a regression coefficient $\beta$ is defined as $\exp(\beta)$ [15].

These hazards ratios, computed by exponentiating the parameter estimates, are useful in interpreting the results of the analysis. If the hazards ratio of a prognostic factor is greater than 1, an increment in the factor increases the hazard rate. If the hazard ratio is less than 1, an increment in the factor decreases the hazard rate [13].

An exploratory analysis with these 5-min averages and standard deviations of speed showed that the hazard ratio for the standard deviation of speed were all greater than unity, while they were all less than one for the average speeds at stations B–H and time slices 1–6. Thus, the coefficients of the variation in speed, if used as independent variables, were expected to result in hazard ratios substantially greater than 1. Therefore, we combined averages and standard deviations of speed, occupancy, and volume into the variables CVS, CVO, and CVV (coefficients of variation of speed occupancy, and volume, respectively, expressed in percentage as $(\mathrm{SS}/\mathrm{AS}) * 100$, $(\mathrm{SO}/\mathrm{AO}) * 100$, and $(\mathrm{SV}/\mathrm{AV}) * 100$). A logarithmic transformation was applied to these coefficients of variation due to the skewed nature of their distributions. It was found that the variables LogCVS, AO, and SV had the most significant hazard ratios.

The results of stratified conditional simple (involving one covariate) logistic-regression models were further examined for these three variables (LogCVS, AO, and SV) over each of the seven loop detectors and six time slices to identify the time duration(s) and the location of loop detector(s) whose traffic characteristics are significantly correlated with the binary target (crash versus noncrash). This was accomplished by estimating the hazard ratio using a proportional hazard regression analysis [proportional hazard regression (PHREG) of statistical analysis system (SAS)] for each of the 126 (seven stations $*$ six time slices $*$ three parameters, i.e., LogCVS, AO, and SV) parameters. The outputs of 126 models were the hazard ratios for these variables at various stations and time slices along with the $p$-values for the test indicating whether the values are significantly different from unity. A hazard ratio is an estimate of the expected change in the odds of having a crash. Therefore, if the output hazard ratio of a variable is significantly different from one and, for example, is equal to two, then increasing the value of this variable by one unit would double the risk of observing a crash at station F (station of the crash). The arrangement used for stations (B–H) and time slices (1–6) is crucial for generating the patterns of a crash risk and its "propagation" in a time–space framework. These 126 single covariate models were estimated for corresponding hazard ratios using the pseudo matched case-control dataset as well.

### B. Results

It was noticed that the hazard ratio for LogCVS and AO increases as we approach the station of the crash (station F) and the time of the crash (slice 1). The values of the hazard ratio for AO were low (i.e., closer to 1.0) yet statistically very significant (indicated by the chi-square statistic and the $p$-value). The

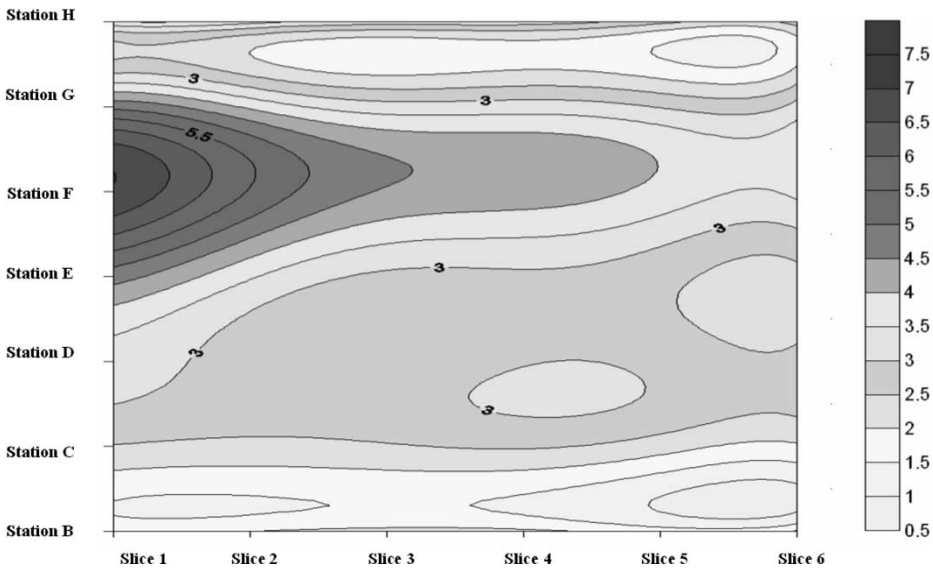


Fig. 1. Time–space pattern of the hazard ratio for LogCVS obtained from a 5-min combined lane dataset for multivehicle crashes.

reason for the low value is that the occupancy usually changes by 1% quite frequently on freeways. Therefore, it is more meaningful to represent the increased risk of observing a crash resulting from a 10% increase in occupancy. This modified risk ratio can be obtained by raising the hazard ratio to the power 10. For all SV parameters, hazard ratios were found to be less than one and appeared to be decreasing as the time and the station of a crash are approached from the downstream direction. Note that the value of the hazard ratio significantly different from 1 (and not necessarily a high value) makes the variable a better crash precursor. In this regard, hazard ratios for SV indicates that, as this parameter becomes smaller at certain freeway sections, the crash risk apparently increases at locations upstream of these sections. Generally, it can be argued that a higher LogCVS, AO value and a lower SV value increase the likelihood of crashes. For LogCVS, this trend is observed starting at about 1–1.5-mi upstream of the crash location (from station D); it is considerably clear at about 0.5 mi upstream and also downstream. Based on the temporal variation trends in hazard ratios, it was concluded that the "ingredients" for an impending crash may be observed about 15 min before its occurrence.

To ascertain the fact that these results are depicting associations of certain traffic parameters with a crash occurrence and not some random patterns in the data, we analyzed hazard ratios from the pseudo case-control dataset. As expected, the trends in the resultant hazard ratio were either nonexistent (as was the case with LogCVSs and SVs with hazard ratios not significantly different than unity) or reversed (as was the case with AOs with hazard ratios significantly less than unity).

To understand the patterns of the crash risk with respect to the time and the location of the crash in a time–space framework, we generated contour plots of hazard ratios corresponding to three parameters (LogCVS, AO, and SV). One such plot, with hazard ratios for LogCVS at various time-slice–station combinations as the contour variables, is shown in Fig. 1. These hazard ratios essentially represent the risk of observing a multivehicle crash at station F attributable to the value of 5-min LogCVS recorded at surrounding stations during the period leading to the crash. According to the color scale provided alongside the plot, it may be seen that the dark colored regions represent high hazard ratios and thereby indicating more risk. Note that these hazard ratios were generated from the 1:5 dataset including one crash and five noncrashes in each stratum. As mentioned earlier, these trends were expectantly nonexistent when the pseudo matched dataset was used to estimate the hazard ratios. The contour plots for the hazard ratios corresponding to the LogCVS values from the pseudo dataset essentially represent "normal" traffic conditions on freeways.

It may be seen in Fig. 1 that the region around station F remains fairly dark (i.e., crash prone) for about a 20-min period, while upstream and downstream sites (stations E and G, respectively) also show a high risk for about a 15–20-min period before observing a crash. These results are significant since they allow leverage in terms of time to anticipate an impending crash. It is also important to note that the clearest trends in the hazard ratio were depicted by the contour plot corresponding to LogCVS, with a stark contrast between the locations of crash and the surrounding stations. The contour plots corresponding to hazard ratios for parameters SV and AO (not shown here) also exhibited similar trends.

## VI. Multivariate Logistic-Regression Model

### A. Data Analysis

The results from the exploratory analysis showed that the three parameters, namely, LogCVS, SV, and AO are most significantly associated with a crash occurrence. These three parameters correspond to 126 potential independent variables (three parameters measured from seven stations during six time slices) for the final model. Also, based on the results from the previous section, we can discard parameters from stations B, C, and D. Even though the hazard ratios for parameters from these stations were significantly different from unity, they were less significant than their counterparts belonging to stations E,

TABLE I
FINAL MODEL DESCRIPTION

| Variable | DF | Parameter Estimate | Standard Error | Chi-square | Pr > Chi-square (p-value) | Hazard Ratio |
|---|---|---|---|---|---|---|
| *LogCVSF2* | 1 | 1.21405 | 0.15548 | 60.9729 | <.0001 | 3.367 |
| *AOG2* | 1 | 0.02466 | 0.00571 | 18.6747 | <.0001 | 1.025 |
| *SVG2* | 1 | -0.19124 | 0.04569 | 17.5216 | <.0001 | 0.826 |

F, G, and H. Essentially, it means that any variable selection procedure examining these factors together (from stations B, C, D, E, F, G, and H) would invariably show the factors from the three upstream stations (stations B, C, and D) as insignificant.

One might argue that, even if that is the case, we should still examine both full and reduced models and make the decision about critical stations based on the classification accuracy. This would not be a good idea since the modeling procedure requires all variables used in the model to be nonmissing (i.e., complete-case analysis) in order to use any observation from the dataset for model building. It should be understood that the data from seven stations would not be simultaneously available at all times due to intermittent hardware failures. It means that some independent variables will be missing in certain observations. The number of observations, which have no variables missing and hence may be used for model building, would be reduced drastically if a lot of independent variables are examined. To illustrate this point, let us assume that each of the $k$ variables can be missing completely at random with a probability $\alpha$; then, the expected proportion of complete cases will be $(1-\alpha)^k$. For example, 1% missing values (missing completely at random) in each of the 126 potential input variables would leave only 28% of complete cases on an average. Note that, with this example, we are not trying to estimate the cases available for a complete-case analysis, but it is provided to illustrate the reduction in the sample size as the number of potential input variables increases.

Also, even though time duration 1 (0–5 min) prior to a crash exhibited significant hazard ratios, it is too close to the actual time of the crash and thus not useful in practice for crash-prediction models. This time duration is thus ignored from further considerations.

For each of the remaining five time slices (time slice 2–6), we have 12 traffic-flow variables LogCVS, SV, and AO from four loop-detector stations (stations E, F, G, and H). To identify the most significant variables from each time slice among the set of these 12 variables, the binary variable "$Y$" is modeled using stratified conditional logistic regression. The SAS procedure PHREG allows one to identify significant variables within this framework using standard automatic search techniques: stepwise, forward, and backward. A full description of the three automatic search procedures can be found in [16]. The $\beta$ coefficients are obtained for significant variables found by these three search procedures.

These procedures resulted in three significant variables for time slice 2 (5–10 min before a crash occurrence): $\mathrm{LogCVSF2} = \log_{10}(\mathrm{CVS})$ at station F (the station of the crash), $\mathrm{AOG2} = \mathrm{AO}$ at station G (the downstream station), and $\mathrm{SVG2} = \mathrm{SV}$ at station G (the downstream station). All other variables were found to be statistically insignificant. Similar search procedures from subsequent time slices resulted in slightly different models involving variables measured during time slice 3, 4, and so on. The model belonging to which time slice should be used was decided based on the classification accuracy achieved from the models. The model with input parameters from time slice 2 was found to be the best in this regard.

The final logistic-regression model included three variables: LogCVSF2, AOG2, and SVG2. The details of the final model are provided in Table I. The table provides model degrees of freedom, estimates for the model coefficients and the standard error, chi-square test statistic along with the $p$-value, and the corresponding hazard ratio for each parameter. Although the model is developed for binary classification, it is essential to establish links between the factors entered in the model and the crash occurrence. The premise of the approach adopted here involves identifying patterns in the loop-detector data that are observed prior to historical crashes. Without the existence of a real relationship between model parameters and crashes, it can be argued that the identified patterns in the loop data have a mere correlation with the crash occurrence rather than a causal relationship. This argument would mean that the estimated statistical model(s) cannot be used to identify crash-prone conditions in the future.

However, it is not the case here, since significant traffic parameters and their model coefficients could be traced as contributing to crash-prone conditions on the freeway. A positive coefficient for LogCVSF2 indicates that the high coefficient of variation at a certain freeway location leads to frequently forming and dissipating queues and in turn leading to conditions in which drivers need to slow down and speed up quite often and be very attentive in following the vehicles ahead of them. SVG2 has a negative $\beta$ coefficient implying increasing odds of a crash as this parameter decreases. The signs of the coefficients indicate (positive for LogCVSF2 and negative for SVG2) that a high variation in speeds with little or no difference in volume across lanes might cause drivers in the slow lane to make lane changes; resulting in increased odds of experiencing a crash. The other factor AOG2 in the model also has a positive coefficient, indicating that a high occupancy at the station downstream of the crash site 5–10 min before the crash increases the odds of a crash occurrence. This observation may

TABLE II
(a) CLASSIFICATION RESULTS FROM THE MULTIVARIATE MODEL ON THE DATASET USED TO DEVELOP THE MODEL. (b) CLASSIFICATION RESULTS FROM THE SIMPLE MODEL WITH LogCVSF2 AS THE COVARIATE ON THE DATASET USED TO DEVELOP THE MODEL

(a)

| | | Predicted | | |
|---|---|---|---|---|
| | | 0(non-crash) | 1(crash) | Total |
| Actual | 0(non-crash) | Frequency = 2719<br>Percent = 43.63<br>Row Pct = 52.69<br>Col Pct = 87.09 | 2441<br>39.17<br>47.31<br>78.49 | 5160<br>82.80 |
| | 1(crash) | 403<br>6.47<br>37.59<br>12.91 | 669<br>10.73<br>62.41<br>21.51 | 1072<br>17.20 |
| | Total | 3122<br>50.10 | 3110<br>49.90 | 6232<br>100.00 |

(b)

| | | Predicted | | |
|---|---|---|---|---|
| | | 0(non-crash) | 1(crash) | Total |
| Actual | 0(non-crash) | Frequency = 3198<br>Percent = 43.19<br>Row Pct = 52.32<br>Col Pct = 85.83 | 2914<br>39.36<br>47.68<br>79.23 | 6112<br>82.55 |
| | 1(crash) | 528<br>7.13<br>40.87<br>14.17 | 764<br>10.32<br>59.13<br>20.77 | 1292<br>17.45 |
| | Total | 3726<br>50.32 | 3678<br>49.68 | 7404<br>100.00 |

be associated with the backward propagation of a congested flow regime, which in turn could increase the probability of observing a rear-end crash.

### B. Classification Accuracy of the Models

As previously explained in the modeling-methodology section, the odd ratio given by (4) may be used to classify crash and noncrash cases. We first calculated the averages of the three variables LogCVSF2, AOG2 and SVG2 over five noncrashes within each of the 1528 matched strata in the dataset. For the $j$th-matched stratum, the vector $x_{2j}$ in (4) may be replaced by the vector of these noncrash means. The odds ratios for each observation in the dataset are then calculated based on (4), utilizing $\beta$ coefficients from Table I with the vector $x_{1j}$ being the observation from the dataset. An observation may be classified as a crash if the corresponding odds ratio is greater than 1 and as a noncrash if the ratio is less than or equal to 1. The classification table resulting from this rule for the 1:5 matched dataset is shown in Table II(a). It may be observed that more than 62.41% of crashes are identified using this threshold for the odd ratio. Table II(b) depicts the classification performance of one of the simple models (i.e., the model with LogCVSF2 as the only covariate) at the same threshold for the odds ratio. The simple model with LogCVSF2 was chosen to generate the comparison classification table, since it was the single most significant model of all one-covariate models. By comparing the two tables, it may be seen that the misclassification rate is higher on crash as well as noncrash cases for the simple model. While simple models have the advantage due to their data requirement, the decision regarding the selection of models must be made based on the classification accuracy. Therefore, the multivariate model is recommended for a reliable classification of the patterns.

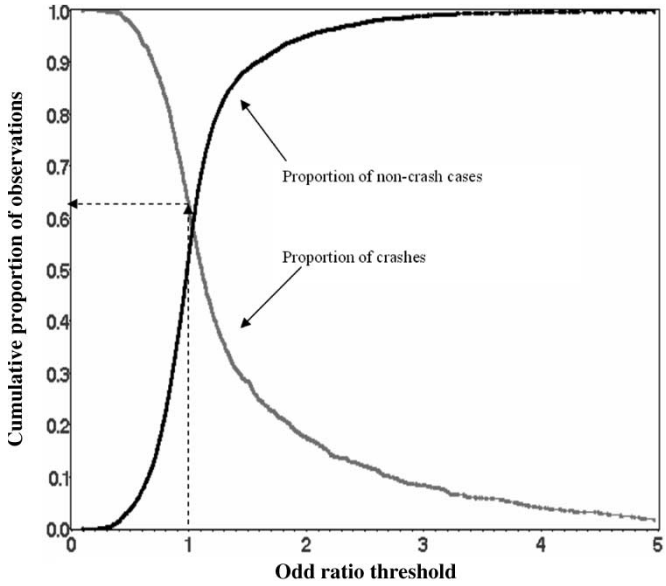


Fig. 2. Classification performance of the multivariate model: Cumulative proportion of crashes above and noncrash cases below a range of odds-ratio-threshold values (the gray curve denotes proportion of crashes and the black curve denotes proportion of noncrash cases).

The threshold (chosen to be equal to 1 here) on odds ratio may be varied in order to achieve a desirable classification given the tradeoff between the overall classification accuracy (crash and noncrash) and the crash identification. Cumulative proportions of crashes above and noncrashes equal or below a range of these odd ratios were determined and plotted against odd ratios in Fig. 2 (only odds-ratio threshold less than or equal to 5 are shown on the horizontal axis). The figure provides the proportion of crashes and noncrashes correctly classified as a function of the chosen odds-ratio threshold. It may be seen that, on this dataset, the threshold of unity provides a reasonable balance between the two conflicting attributes (i.e., overall classification and crash identification) and hence is recommended as the cutoff value. However, in a real-time application, this threshold might be altered based on considerations such as time of day, day of the week, or freeway operation regime. For example, during a free-flow operation (characterized by high speeds), a lower value of odds ratio may be used as the threshold so that most of the crashes are identified even if that increases the number of "false-alarms" because speed is known to be positively associated with the severity of crashes.

## VII. IMPLEMENTATION PLAN

### A. Phase 1—Simple Model(s) Implementation: Procedure and Data Requirement

The single-covariate (i.e., simple) models need information from only one loop-detector station at a time. It makes these

TABLE III
HAZARD RATIOS FROM SINGLE COVARIATE MODELS CONSISTING OF LogCVS FROM FIVE STATIONS AND SIX TIME SLICES

| Hazard ratio corresponding to station | Hazard ratio to asses the crash risk within next….. | | | | | |
|---|---|---|---|---|---|---|
| | 0-5 minutes (slice 1) | 5-10 minutes (slice 2) | 10-15 minutes (slice 3) | 15-20 minutes (slice 4) | 20-25 minutes (slice 5) | 25-30 minutes (slice 6) |
| D | 3.331 | 3.132 | 2.430 | 3.074 | 2.735 | 2.499 |
| E | 4.436 | 3.335 | 3.025 | 3.257 | 2.664 | 2.426 |
| F | 7.237 | 5.580 | 4.485 | 3.801 | 3.654 | 3.809 |
| G | 4.705 | 3.899 | 3.037 | 3.519 | 3.209 | 2.964 |
| H | 3.976 | 3.635 | 3.476 | 3.139 | 2.623 | 2.871 |

models particularly attractive given the intermittent failures of magnetic loops on certain stations. It also makes the data processing faster for the online application. The output for each of the simple models developed was the hazard ratio for the corresponding covariate. According to its definition, the hazard ratio multiplied by the value of corresponding covariate would provide the measure of a crash risk relative to the situation if the value of the covariate was zero.

For a real-time application, the instrumented freeway corridor can be divided into 69 (which is the total number of loop-detector stations) sections of approximately 0.5 mi in each direction, such that each loop detector remains at the center of each such section. It is clear that for crashes occurring on any of these sections, the corresponding station would be analogous to station F (station of the crash), as defined earlier in the paper. The series of 69 loop detectors on the corridor may then be divided into sets of five stations as (1–5), (2–6), (3–7), and so on up to (65–69). The sets of five detectors are chosen because these stations would correspond to stations D–H (two upstream stations, station F, and two downstream stations, respectively). Note that hazard ratios for parameters from stations B and C, the two stations located farthest upstream of the station of the crash, were not as critically associated with a crash occurrence as those from stations D–H. Therefore, the sets of loop detectors chosen for the implementation strategy consist of five stations (D–H) as opposed to the seven stations (B–H), which contributed input parameters for simple models. Among the three parameters (i.e., LogCVS, SV, and AO) LogCVS was chosen for the preliminary assessment of the crash risk because the contour plot depicting the time–space variation of the crash risk (Fig. 1) based on LogCVS showed a stark contrast between the location that experienced a crash and the locations that did not. The values of the hazard ratios corresponding to LogCVSs measured at these five stations (D–H) during the six time slices are shown in Table III.

With the hazard ratios for LogCVS from station D–H (shown in Table III), one can observe the change in the crash risk on the basis of changes in LogCVS and update it in real time. The update may be done on a continuous basis as soon as new observations are recorded in real time. For example, we first calculate the LogCVS based on ten most recent observations available. Then, after 30 s as the latest observation (since loop data are updated every 30 s) come in, the data may be included in the calculation of LogCVS, replacing the farthest observation. The LogCVS measured at different stations may be multiplied by the corresponding hazard ratios to obtain the measure of a crash risk for a period up to the next 30 min. Hazard ratios corresponding to station D would be chosen if the station is the most upstream of the set of five, to station G if it is the most downstream, and to station F if it is the station belonging to that particular section and so on. The decision of selecting the time slice depends upon how much time in advance are we trying to assess the crash risk. For example, to obtain the risk of observing a crash within the next 10–15-min period, the hazard ratio(s) belonging to time slice 3 should be used; for the next 5–10 min, the hazard ratio(s) from time slice 2 may be used. The measure of a crash risk may then be plotted as a contour variable in a time–space framework. Based on the patterns depicted by the continuously updated plots, freeway locations with a high crash risk may be identified in real time.

### *B. Implementation of Simple Models: Illustration*

In this section, we illustrate how the crash risk and its variation on a freeway location may be observed through continuously updated contour plots. The application is demonstrated using historical loop-detector data belonging to a crash and a noncrash case. Table IV shows a sample of LogCVS calculated as a moving average from the actual historical traffic speed data from a set of five detectors, starting at 15 min prior to the time of crash. These data were collected prior to a real crash that occurred on April 6, 1999 near station 34 at 4:35 P.M. on Interstate 4 in the eastbound direction. The formulation of LogCVS remains the same as in the modeling phase.

Table III depicted the hazard ratios corresponding to stations D–H at all six time slices. In Table V(a)–(c), the process of calculating the values for the contour variables (measure of the crash risk obtained by multiplying LogCVS values with the corresponding hazard ratios) is shown. In the first row of Table V(a), 1.42 (the LogCVS value obtained from station 32, which corresponds to station D, during the 5-min period of 4:14:30 to 4:19:30 P.M.) is multiplied by the hazard ratios for station D corresponding to each of the six time slices to obtain

TABLE IV
SNAP SHOTS OF 5-min LogCVS (VALUES UPDATED EVERY 30 s) CALCULATED AS A MOVING AVERAGE STARTING AT 15 min PRIOR TO A CRASH OCCURRENCE

| Date-Time | Station | Station of Crash | LogCVS |
|---|---|---|---|
| 4/6/99 4:19:30 PM | 32 (D) | 34 | 1.42 |
| 4/6/99 4:20:00 PM | 32 (D) | 34 | 1.42 |
| 4/6/99 4:20:30 PM | 32 (D) | 34 | 1.45 |
| 4/6/99 4:19:30 PM | 33 (E) | 34 | 1.60 |
| 4/6/99 4:20:00 PM | 33 (E) | 34 | 1.65 |
| 4/6/99 4:20:30 PM | 33 (E) | 34 | 1.67 |
| 4/6/99 4:19:30 PM | 34 (F) | 34 | 1.42 |
| 4/6/99 4:20:00 PM | 34 (F) | 34 | 1.43 |
| 4/6/99 4:20:30 PM | 34 (F) | 34 | 1.52 |
| 4/6/99 4:19:30 PM | 35 (G) | 34 | 1.56 |
| 4/6/99 4:20:00 PM | 35 (G) | 34 | 1.57 |
| 4/6/99 4:20:30 PM | 35 (G) | 34 | 1.59 |
| 4/6/99 4:19:30 PM | 36 (H) | 34 | 1.71 |
| 4/6/99 4:20:00 PM | 36 (H) | 34 | 1.69 |
| 4/6/99 4:20:30 PM | 36 (H) | 34 | 1.74 |

the measures of a crash risk for the next 30 min. In the second row, 1.42 is replaced by 1.60, which happens to be the value of LogCVS from station 33 (i.e., station E) during the last 5-min period. The third, fourth, and fifth rows of the table are created by multiplying the hazard ratios corresponding to stations F, G, and H with the values of LogCVS at the corresponding stations.

Table V(b) is generated through a similar procedure; the only difference being that the values for LogCVS are now updated based on the most recent speed observations. In Table V(c), the values of the independent covariate LogCVS are further updated based on the most recent speed observations. in Table V(a), it may be noted that the values of LogCVS are highlighted in yellow (light color) to associate them with the observations from the same period of time (4:14:30–4:19:30 P.M.) in Table IV. Similarly, in Table V(b) and (c), the updated values for LogCVS are highlighted red (dark) and green (medium) to associate them with the respective 5-min periods during which these values were observed (Table IV).

Three contour plots depicting the variation in the crash risk generated from these data are shown in Fig. 3(a)–(c). It can clearly be seen that the region about station F remains dark, indicating a high risk for a crash occurrence. It may be noted that the values for the contour variable in Fig. 3(a) come from the corresponding cells of Table V(a). The plot is updated to Fig. 3(b) as soon as the new set of readings is recorded (after 30 s). The values from Table V(b) are used to generate the updated plot in Fig. 3(b), which eventually turns into Fig. 3(c) after 30 s based on Table V(c). The updated patterns do not differ much from their predecessor since most of the observations contributing to the calculation of LogCVS remain the same. Only three observations out of thirty (the number of total observations used for computing required averages and standard deviations) are updated after 30 s.

These figures may be contrasted with similar patterns generated for the same time of the day prior to a corresponding matched noncrash case (On April 27, 1999 from the same set of stations). These patterns are shown in Fig. 4(a)–(c). In a real-time application of the models, the measures of risk may be calculated continuously and the corresponding plots can be generated using the color scheme depicted on the side of each plot. According to the color scale, the dark (red) colors represent the regions of the contours where the measure of the crash risk exceeds 6.0. There is no such region in Fig. 4(a)–(c), which correspond to a matched noncrash case from the dataset. It should be noted that the difference between the crash and noncrash

TABLE V
(a) MEASURE FOR THE RISK OF OBSERVING A CRASH IN THE SEGMENT BELONGING TO STATION F WITHIN THE NEXT 30 min AT TIME 4:19:30 P.M. (b) MEASURE FOR THE RISK OF OBSERVING A CRASH IN THE SEGMENT BELONGING TO STATION F WITHIN THE NEXT 30 min AT TIME 4:20:00 P.M. (c) MEASURE FOR THE RISK OF OBSERVING A CRASH IN THE SEGMENT BELONGING TO STATION F WITHIN THE NEXT 30 min AT TIME 4:20:30 P.M.

(a)

| Measure of risk according to *Log CVS* from station | Measure of the crash risk with in next | | | | | |
|---|---|---|---|---|---|---|
| | 0-5 minutes (slice 1) | 5-10 minutes (slice 2) | 10-15 minutes (slice 3) | 15-20 minutes (slice 4) | 20-25 minutes (slice 5) | 25-30 minutes (slice 6) |
| (D) | 3.331*1.42 | 3.132*1.42 | 2.430*1.42 | 3.074*1.42 | 2.735*1.42 | 2.499*1.42 |
| (E) | 4.436*1.60 | 3.335*1.60 | 3.025*1.60 | 3.257*1.60 | 2.664*1.60 | 2.426*1.60 |
| (F) | 7.237*1.42 | 5.580*1.42 | 4.485*1.42 | 3.801*1.42 | 3.654*1.42 | 3.809*1.42 |
| (G) | 4.705*1.56 | 3.899*1.56 | 3.037*1.56 | 3.519*1.56 | 3.209*1.56 | 2.964*1.56 |
| (H) | 3.976*1.71 | 3.635*1.71 | 3.476*1.71 | 3.139*1.71 | 2.623*1.71 | 2.871*1.71 |

(b)

| Measure of risk according to *Log CVS* from station | Measure of the crash risk with in next | | | | | |
|---|---|---|---|---|---|---|
| | 0-5 minutes (slice 1) | 5-10 minutes (slice 2) | 10-15 minutes (slice 3) | 15-20 minutes (slice 4) | 20-25 minutes (slice 5) | 25-30 minutes (slice 6) |
| (D) | 3.331*1.42 | 3.132*1.42 | 2.430*1.42 | 3.074*1.42 | 2.735*1.42 | 2.499*1.42 |
| (E) | 4.436*1.65 | 3.335*1.65 | 3.025*1.65 | 3.257*1.65 | 2.664*1.65 | 2.426*1.65 |
| (F) | 7.237*1.43 | 5.580*1.43 | 4.485*1.43 | 3.801*1.43 | 3.654*1.43 | 3.809*1.43 |
| (G) | 4.705*1.57 | 3.899*1.57 | 3.037*1.57 | 3.519*1.57 | 3.209*1.57 | 2.964*1.57 |
| (H) | 3.976*1.69 | 3.635*1.69 | 3.476*1.69 | 3.139*1.69 | 2.623*1.69 | 2.871*1.69 |

(c)

| Measure of risk according to *Log CVS* from station | Measure of the crash risk with in next | | | | | |
|---|---|---|---|---|---|---|
| | 0-5 minutes (slice 1) | 5-10 minutes (slice 2) | 10-15 minutes (slice 3) | 15-20 minutes (slice 4) | 20-25 minutes (slice 5) | 25-30 minutes (slice 6) |
| (D) | 3.331*1.45 | 3.132*1.45 | 2.430*1.45 | 3.074*1.45 | 2.735*1.45 | 2.499*1.45 |
| (E) | 4.436*1.67 | 3.335*1.67 | 3.025*1.67 | 3.257*1.67 | 2.664*1.67 | 2.426*1.67 |
| (F) | 7.237*1.52 | 5.580*1.52 | 4.485*1.52 | 3.801*1.52 | 3.654*1.52 | 3.809*1.52 |
| (G) | 4.705*1.59 | 3.899*1.59 | 3.037*1.59 | 3.519*1.59 | 3.209*1.59 | 2.964*1.59 |
| (H) | 3.976*1.74 | 3.635*1.74 | 3.476*1.74 | 3.139*1.74 | 2.623*1.74 | 2.871*1.74 |

case is highlighted here to illustrate the application. However, in some other cases, the difference may not be as clear.

The simple models are proposed to be applied in the first phase of the proposed implementation plan. In the second and final phase, a multivariate model employing data from three stations would be applied to assess the crash risk for the next 5–10-min period. Keeping this in perspective, an effective online application strategy would be to critically examine the region in the contour plots (generated in the first phase of the strategy) where the abscissa encompasses time slice 3. It would provide a preliminary assessment for the risk of a crash occurrence within the next 10–15 min. If the sequential patterns of a crash risk appear hazardous, as is the case with those depicted in Fig. 3(a)–(c), then the multivariate model can be employed for a reliable classification of the patterns in loop data. The application for the multivariate model is described in the following section.

### C. Phase 2—Application of Multivariate Models: Procedure and Data Requirement

Following the detection of hazardous patterns through the contour plots in the first phase; a multivariate model may be applied for classification. As explained earlier, the odds ratio can be calculated using (4) to classify the patterns in crash and noncrash cases.

For this purpose, we first calculated the averages of the three covariates included in the final model: LogCVSF2, AOG2, and SVG2 over five noncrashes within each matched stratum of the 1:5 matched dataset. For the $j$th matched set, vector $\overline{x}_{k2j}$ in (4)

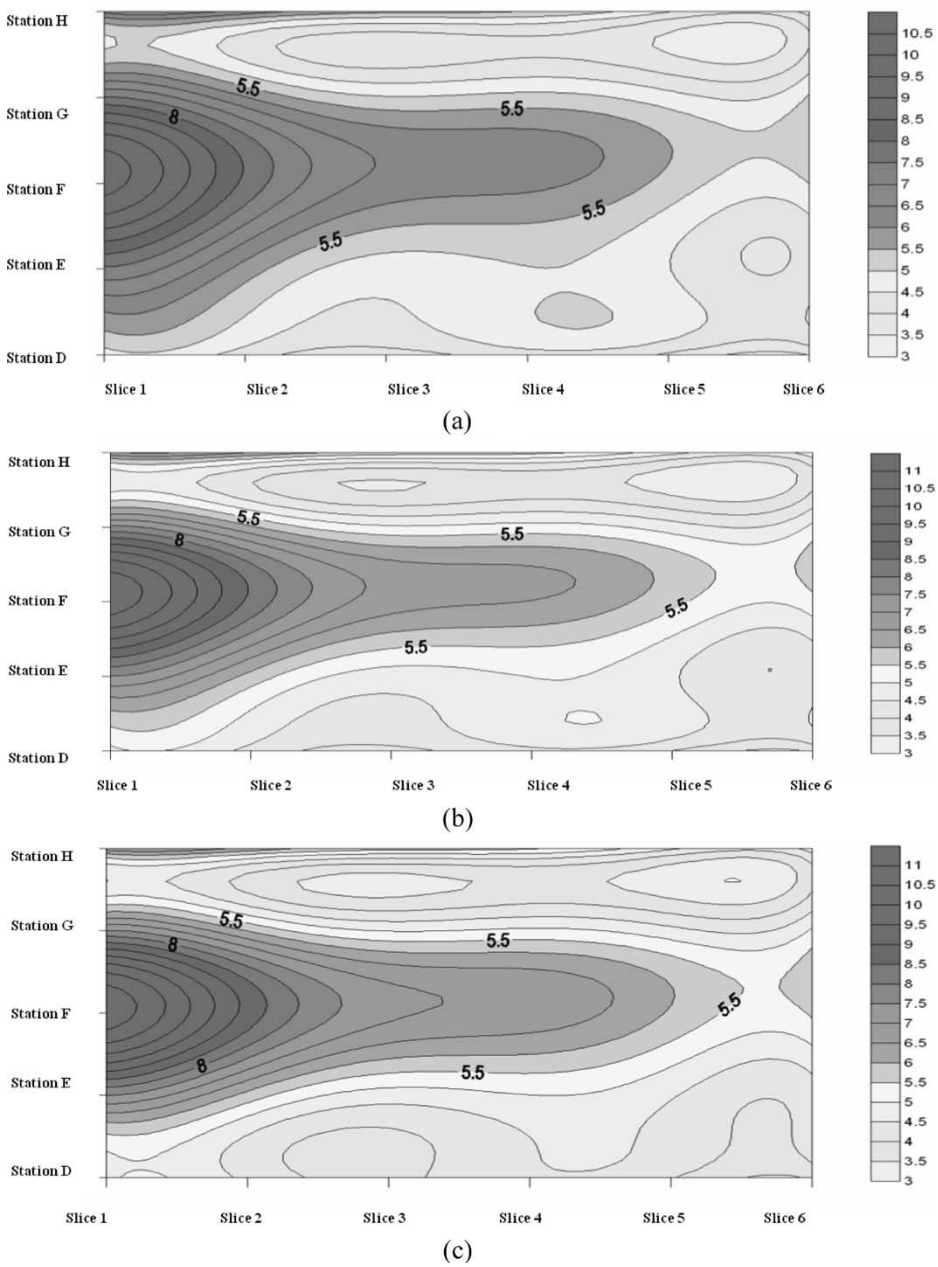


Fig. 3. (a)–(c) Illustrative pattern(s) of variation in measures for the risk of observing a crash in the segment belonging to station F updated every 30 s.

may be replaced by the vector of these noncrash means and the most current values of the three variables can be used as $x_{k1j}$ to calculate the odds ratio. Equation (4) with estimated values of the parameters could be rewritten as

$$\left\{\frac{\frac{p(x_{1j})}{[1-p(x_{1j})]}}{\frac{p(\overline{x}_{2j})}{[1-p(\overline{x}_{2j})]}}\right\} = \exp\left(1.21405(\text{LogCVSF2} - .95164)\right.$$
$$\left. + 0.02466(\text{AOG2} - 13.26) - 0.19124(\text{SVG2} - 2.56445)\right). \quad (5)$$

The RHS of (5) represents the odds ratio. Note that the $\beta_p$ (model coefficients) in (4) have been replaced with the estimates of model coefficients for LogCVSF2, AOG2, and SVG2, respectively. The values for the estimates of the three model coefficients were shown in Table I. The vector $\overline{x}_{k2j}$ has been replaced with the averages of the three covariates over the five noncrash cases of the stratum. The real-time values for the three independent variables (LogCVSF2, AOG2, and SVG2) may be used in (5) to obtain the odds ratio of having a crash versus not having a crash. If the resultant odds ratio exceeds unity, then the patterns may be classified as "crash prone." These odds ratios and the resultant classification may also be updated in a way similar to the contour plots. To update the odds ratios every 30 s, the oldest set of observations in the 5-min period may be replaced by the 30-s data most recently recorded.

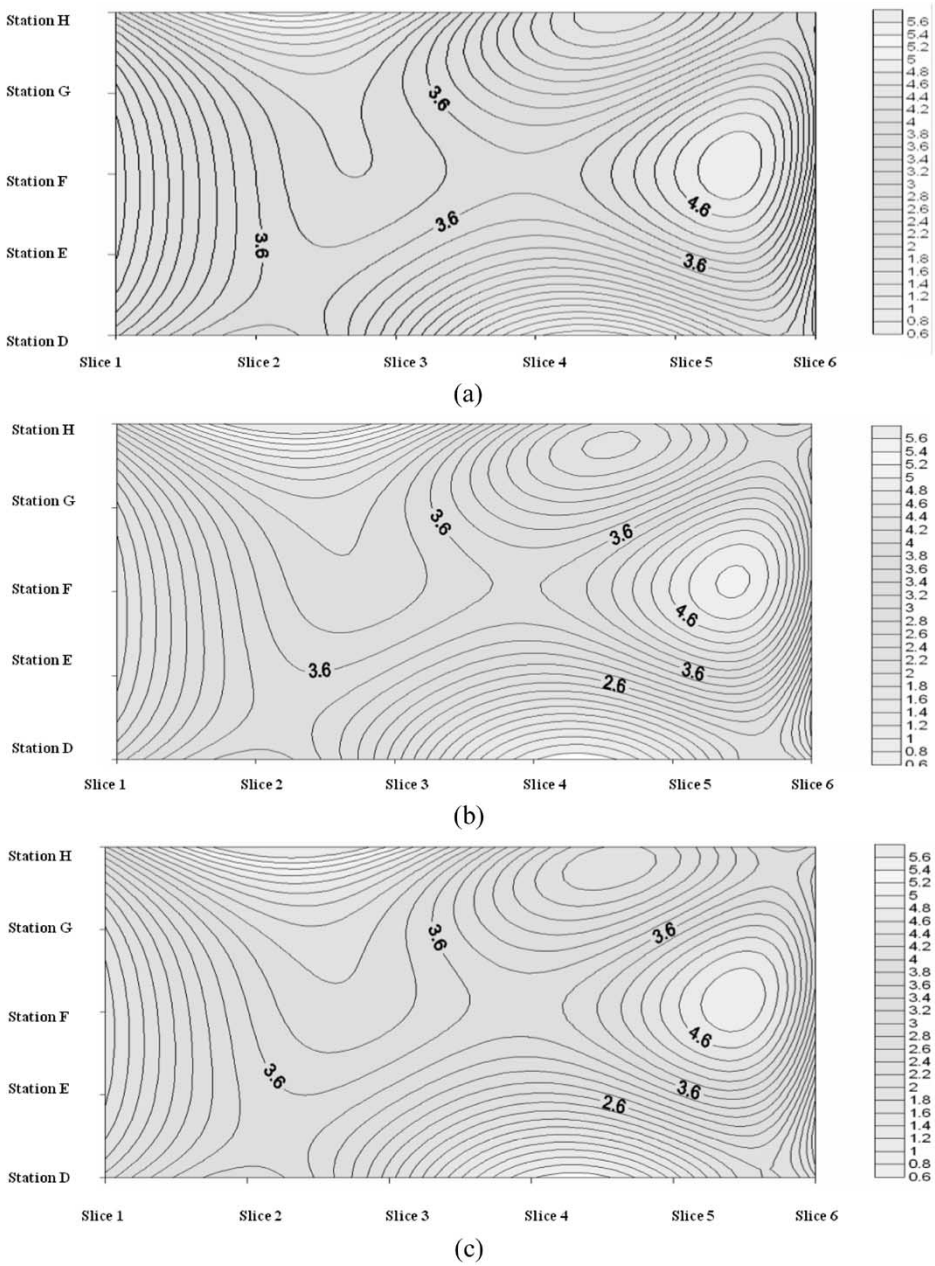


Fig. 4. (a)–(c) Illustrative pattern(s) of variation in measures for the risk of observing a crash in the segment belonging to station F updated every 30 s for a noncrash scenario.

### *D. Multivariate Model: Illustration*

Table VI shows the historical values for the three covariates included in the final model starting at 10 min prior to the same crash, which was used to illustrate the application of the simple models. These values are calculated on a continuous basis, i.e., the averages and standard deviations are calculated as moving averages. The procedure to obtain the input parameters through the moving average is same as that described in the implementation plan for the simple models. In the first row, three input parameters (5-min average occupancy, 5-min standard deviation of volume, and 5-min coefficient of variation in speed) are obtained using 30 latest observations (5 min $*$ two observations every min $*$ three lanes) from the corresponding stations. In subsequent rows, parameters are updated using the most recent speed, volume, and occupancy observations. The odds ratios are calculated based on (5). The odds ratios of having a crash versus not having a crash near station F for the observed values of the independent variables are also shown in the table. It may be seen that in all three instances, the odds ratio is greater than unity; hence, the model classifies the data patterns as "crash prone." This was expected, since it is already known that a crash did occur following these data patterns. Since the final model included the parameters from time slice 2, the odds of a crash occurrence within the next 5–10-min period are assessed. It should be noted that the application of the model just requires simple arithmetic calculations using traffic parameters and

TABLE VI
OUTPUT FROM THE FINAL MULTIVARIATE MODEL WHEN APPLIED ON THE HISTORICAL LOOP DATA PRIOR TO A CRASH

| *Date-Time* | *LogCVS (Station 34) (Station F)* | *SV (Station 35) (Station G)* | *AO (Station 35) (Station G)* | *Odds ratio* | *Decision for next 5-10 minute slice* |
|---|---|---|---|---|---|
| 4/6/99 4:25:00 PM | 1.69 | 2.44 | 19.97 | 2.97 | Crash prone |
| 4/6/99 4:25:30 PM | 1.64 | 2.07 | 19.77 | 2.96 | Crash prone |
| 4/6/99 4:26:00 PM | 1.55 | 2.21 | 20.07 | 2.62 | Crash prone |

estimated logistic-regression coefficients (from Table I). Hence, the model can be easily applied to real-time data.

## VIII. SUMMARY AND CONCLUSION

The objective of this research was to develop a strategy to identify crash-prone conditions on freeways in real time. A detailed database was assembled for all crashes that occurred on the instrumented corridor of Interstate 4 in the period 1999–2002. Statistical links between turbulent traffic conditions measured through loop detectors and crash occurrences were established. It was demonstrated that these links may be used for identification of freeway "black spots" in real time.

A logistic regression within a stratum matched case-control study design was used as the analysis technique. The matched design of the study implicitly accounts for external factors such as the freeway geometry, time of the day, and day of the week. Following an exploratory analysis, a series of simple (involving one covariate) models were estimated for the binary target (crash versus noncrash). A multivariate logistic-regression model was also estimated through a stepwise variable selection procedure. A 5-min coefficient of variation in speed at the loop-detector station closest to the crash location was found to affect the crash occurrence most significantly. In the final model, a 5-min average occupancy and a 5-min standard deviation of volume (observed at the loop detector downstream of the crash location) were also found significant. The final multivariate model with these three input variables can be used to calculate the odds ratio of observing a crash versus not observing a crash. A threshold value on this ratio may be established to determine whether the location has to be flagged as a potential "crash location." It was shown that using 1.0 as the threshold over 62% crashes can be identified by the model. It should be noted that, even though the simple models achieved a classification accuracy inferior to that of the final model, the advantage of using those models is that they have tolerant data requirements. In addition, it was shown that the results from simple one-covariate models may be used to obtain a time–space variation of the crash risk.

A two-stage real-time application plan for these models was also proposed in the paper. The proposed plan essentially involves a preliminary assessment of freeway traffic conditions through the contour plots generated by applying simple models. If these plots indicate a high risk of crash occurrence, the loop data may be subjected to the multivariate model for classification. If the classification model identifies traffic patterns from the loop detectors as "crash prone," then the traffic-management authorities can keep their crash mitigation squad on alert so that the impacts of the impending crash occurrence may be minimized. If these models trigger more warnings at certain freeway locations, then the traffic-management authorities may closely watch such locations through surveillance cameras. It will help in recognizing the problems associated with these locations, e.g., weaving sections, configuration of the ramps with respect to the freeway, etc.

It should be acknowledged that some parameters (e.g., the rainfall information and human factors), which could potentially impact the probability of a crash occurrence, have not been included in the analysis. It is expected that the effect of weather conditions (i.e., the rainfall) on traffic would arguably be captured by the parameters measured at the loop detectors. As for the human factors, the errors by individual driver(s) would play a critical role in a crash occurrence. However, there is no way to measure the behavior of all the drivers on a freeway section in real time. Hence, the goal in this study was to try and identify patterns observed in the loop-detector data before the historical crashes. These patterns were then explained as conditions in which crashes are more likely to occur and under such conditions drivers need to be more attentive in order to avoid crashes.

It should also be noted that the models developed here are calibrated for the Interstate 4 corridor in Orlando. Therefore, the same model coefficients may not be applicable for other corridors. However, the matched case-control sampling approach can be easily extended to any other instrumented corridor equipped with loop detectors.

Based on the results of this study and the understanding of the crash occurrence phenomena, more aggressive strategies, e.g., variable speed limits and warning the drivers through variable message signs etc., need to be explored. These strategies may be used by freeway management authorities to intervene and reduce the crash potential. However, the development of these proactive strategies, their application, and their impact on drivers are nontrivial issues and demand separate attention.

## ACKNOWLEDGMENT

The authors would like to thank L. Hsia of FDOT. The loop-detector data used in this paper were obtained from the University of Central Florida (UCF) data warehouse, and the crash data were provided by FDOT.

## REFERENCES

[1] T. F. Golob, W. W. Recker, and V. M. Alvarez, "Freeway safety as a function of traffic flow," *Accident Anal. Prev.*, vol. 36, no. 6, pp. 933–946, 2004.
[2] E. Hauer, "Statistical safety modeling," presented at the 83rd Annual Meeting Transportation Research Board (TRB), Washington, DC, 2004, Paper 04-2692.
[3] M. Zhou and V. P. Sisiopiku, "Relationship between volume-to-capacity ratios and accident rates," *Transp. Res. Rec.*, no. 1581, pp. 47–52, 1997.
[4] M. Abdel-Aty, N. Uddin, F. Abdalla, A. Pande, and L. Hsia, "Predicting freeway crashes based on loop detector data using matched case-control logistic regression," *Transp. Res. Rec.*, no. 1897, pp. 88–95, 2004.
[5] R. Hughes and F. Council, "On establishing relationship(s) between freeway safety and peak period operations: Performance measurement and methodological considerations," presented at the 78th Annual Meeting Transportation Research Board (TRB), Washington, DC, 1999, Paper 0384.
[6] C. Lee, F. Saccomanno, and B. Hellinga, "Analysis of crash precursors on instrumented freeways," *Transp. Res. Rec.*, no. 1784, pp. 1–8, 2002.
[7] ——, "Real-time crash prediction model for the application to crash prevention in freeway traffic," *Transp. Res. Rec.*, no. 1840, pp. 68–77, 2003.
[8] C. Oh, J. Oh, S. Ritchie, and M. Chang, "Real time estimation of freeway accident likelihood," presented at the 80th Annual Meeting Transportation Research Board, Washington, DC, 2001, Paper 01-3445.
[9] M. Abdel-Aty and A. Pande, "Identifying crash propensity using specific traffic speed conditions," *J. Saf. Res.*, vol. 36, no. 1, pp. 97–108, 2005.
[10] T. F. Golob and W. W. Recker, "A method for relating type of crash to traffic flow characteristics on urban freeways," *Transp. Res., Part A Policy Pract.*, vol. 38, no. 1, pp. 53–80, 2004.
[11] T. F. Golob, W. W. Recker, and V. M. Alvarez, "Tool to evaluate the safety effects of changes in freeway traffic flow," *J. Transp. Eng.*, vol. 130, no. 2, pp. 222–230, 2004.
[12] D. Collett, *Modeling Binary Data*. London, U.K.: Chapman & Hall, 1991.
[13] SAS Institute, *SAS/STAT User's Guide*, 1999, Cary, NC: SAS Inst.
[14] M. Abdel-Aty, A. Pande, L. Hsia, and F. Abdalla, "The potential of loop detector data in improving freeway safety," *ITE J.*, vol. 75, no. 12, 2005.
[15] A. Agresti, *Categorical Data Analysis*. New York: Wiley, 2002.
[16] D. W. Hosner and S. Lemeshow, *Applied Logistic Regression*. New York: Wiley, 1989.

**Mohamed Abdel-Aty** is an Associate Professor of civil engineering at the University of Central Florida (UCF), Orlando. His main research interest is in the areas of traffic safety, travel demand analysis, and intelligent transportation systems (ITS). He is the author or coauthor of more than 120 published papers. He is a member of two Transportation Reasearch Board (TRB) committees.

Dr. Abdel-Aty is a member of the Editorial Advisory Board of Accident Analysis and Prevention and the *ITS Journal*. He is a recipient of the 2003 UCF Distinguished Researcher Award. He is a Registered Professional Engineer in Florida.

**Anurag Pande** received the B. Tech. degree in civil engineering from the Indian Institute of Technology (IIT), Bombay, India, in 2002 and the M.S. and Ph.D. degrees in transportation systems engineering from the University of Central Florida, Orlando, in 2003 and 2005, respectively.

He is currently a Research Associate at the University of Central Florida. His research interests include traffic safety analysis, intelligent transportation systems, and statistical and data mining applications in transportation engineering.