Proceedings of the 8th International
IEEE Conference on Intelligent Transportation Systems
Vienna, Austria, September 13-16, 2005

Traffic Surveillance from a Safety Perspective: An ITS Data Application

M. Abdel-Aty, A. Pande, and N. Uddin

Abstract— **Reactive traffic management strategies such as incident detection are becoming less relevant with the advancement of mobile phone usage. Freeway management in the 21st century needs to shift focus toward proactive strategies that include anticipating incidents such as the crashes. "Predicting" crash occurrences would also be the key to traffic safety. A two-step approach to identify freeway locations with high probability of crashes through real-time traffic surveillance data is presented here. For this study historical crash and corresponding traffic data from loop detectors were gathered from a 58-km (36-mile) corridor of Interstate-4. Following an exploratory analysis two types of logistic regression models, i.e., simple and multivariate, were developed. 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 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 crash occurrence. A demonstration of this two-stage real-time application strategy is also provided in the paper.**

I. Introduction

The emphasis in freeway management has largely been toward analyzing the post-incident traffic surveillance data in order to timely detect traffic incidents. The advancement in cell phone usage is rendering such reactive strategies irrelevant. The focus of freeway management should therefore shift toward anticipating incidents prior to their occurrence and devise countermeasures. Crashes are arguably the most critical and "predictable" type of incidents. However, traditional freeway safety literature does not offer solution to the traffic management problem of anticipating crashes due to their stated focus on crash frequency or crash rate estimation. The traditional approach to traffic safety is not sufficient to "predict" crashes in real-time using traffic flow variables measured from loop detectors. There is a need to estimate models that use dynamic flow variables as input and determine whether or not they potentially precede a crash occurrence.

One such crash prediction model was developed in one of our previous studies [1]. The model achieved satisfactory crash identification and demonstrated the feasibility of predicting crashes in real-time. However, the model was developed using data from a small, dense urban segment of the freeway (Interstate-4 in City of Orlando) with crashes spanning a short period of time (seven months). For this study the crash data was expanded to include crashes that occurred during 4-year period (from 1999 through 2002) on the 58-km (36-mile) instrumented corridor of Interstate-4 in Orlando, FL (USA). A stratified case control dataset consisting of traffic data corresponding to the crash (case) and five matched non-crashes (controls) was created. The purpose of matched case-control analysis is to explore the effects of independent variables of interest on the binary outcome while controlling for other confounding variables through the design of the study. Separate simple (one covariate) as well as multivariate logistic regression models were developed using the matched sample. Based on the results from these models a two stage implementation plan to obtain reliable real-time assessment of potential for crash occurrence is proposed. It is worth mentioning that the approach presented here is data-driven and actual mechanism of crashes is not considered. Detailed vehicle movement data would be needed to establish sound and reliable crash mechanism models; which being unavailable loop detector data have been used as a surrogate measure.

The paper is divided in seven sections. A brief summary of literature is provided in the next section followed by theoretical details of the modeling methodology. Forth section summarizes data collection and preparation. Fifth section deals with preliminary data analysis and details of the multivariate model. It is followed by a two-stage real-time implementation plan and conclusions are provided in the end.

II. Background

Hughes and Council [2] explored the relationship between freeway safety and peak period operations using loop detector data, it was one of the first studies aiming at preemptive traffic management. Lee *et al.* [3] developed a

Manuscript received November 18, 2004. This work was supported in part by the Florida Department of Transportation.

Mohamed Abdel-Aty is with the Department of Civil and Environmental Engineering, University of Central Florida, Orlando, FL 32826 USA (Corresponding author Phone: 407-823-5657; fax: 407-823-3315; e-mail: mabdel@mail.ucf.edu).

Anurag Pande is with the Department of Civil and Environmental Engineering, University of Central Florida, Orlando, FL 32826 USA (e-mail: anurag@mail.ucf.edu).

Nizam Uddin is with the Department of Statistics and Actuarial Sciences, University of Central Florida, Orlando, FL 32826 USA (e-mail: nuddin@mail.ucf.edu).

log-linear model to predict crashes through estimation of crash precursors from loop detector data. In a later study by the same authors [4], the aforementioned model was refined and the coefficient of temporal variation in speed was shown to have a relatively longer-term effect on crash potential than density while the effect of average variation of speed across adjacent lanes was found to be insignificant.

Oh et al. [5] suggested a classification approach for the problem and argued that five minutes standard deviation of speed was the best indicator of "disruptive" traffic flow leading to a crash as opposed to "normal" traffic flow. Abdel-Aty and Pande [6] also used probabilistic neural network (PNN) as the classification algorithm and demonstrated the feasibility of predicting crashes at least 10-minutes prior to their occurrence. In some of the more detailed recent studies Golob and Recker [8, 9] concluded that the collision type is the best-explained characteristic and is related to the median speed and left-lane and interior lane variations in speed. Based on similar results Golob *et al.* [9] used data for more than 1000 crashes over six major freeways in Orange County, California and developed a software tool *FITS* (Flow Impacts on Traffic Safety) to forecast the type 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 *SR 55* (State Road 55) was also demonstrated. Findings from the aforementioned studies point towards potential application of real-time traffic data in the field of traffic safety. However, crashes usually occur due to result of complex interaction between traffic, geometric and environmental factors and it is difficult to explicitly account for the wide range of these factors in any of the modeling frameworks proposed by the studies mentioned above.

The authors in their earlier studies [1, 10, 11] argued that the accuracy of real-time crash prediction model may be increased if the model utilizes information on traffic flow characteristics for both crash and non-crash cases while controlling for other external factors (thereby implicitly accounting for factors such as the geometry and location). It was proposed that this can be achieved using a within-stratum analysis of a binary outcome variable Y (crash or non-crash) as a function of traffic flow variables $X_1, X_2, \ldots X_k$ from matched crash-non-crash cases where a matched set (stratum) can be formed using crash site, time, day of the weak, season, year, etc., so that the variability due to these factors is controlled. The 5-minute average lane occupancy measured upstream and coefficient of variation in speed measured downstream of the crash location were identified to be the most significant crash precursors in the study [1]. However, the study was limited in scope due to insufficient data. A small, dense, and largely urban 21-km (13-mile) section of the freeway corridor was analyzed for just seven months. Due to lack of complete data, issues about the determination of the exact time of historical crashes could not be addressed thoroughly. With largely uniform traffic and crash characteristics on the segment analyzed, the transferability of the model remained suspect. In this study the database has been expanded to include crashes spanning four years on the 58-km (36-mile) corridor. Furthermore, a detailed online application strategy has been proposed in order to identify real-time "black spots" on the freeway corridor.

III. Methodology

To understand the matched case-control logistic regression in the context of the present research problem let's assume that there are N strata with 1 case and m controls in each stratum. The probability of any observation in a stratum being a crash may be modeled using the following linear logistic regression model:

$$logit\{p_j(x_{ij})\} = \alpha_j + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \ldots\ldots\ldots\ldots + \beta_k x_{kij} \qquad (1)$$

where $p_j(x_{ij})$ is the probability that the i^{th} observation in the j^{th} stratum belongs to a crash; $x_{ij} = (x_{1ij}, x_{2ij}, \ldots\ldots x_{kij})$ is the vector of k traffic flow variables $x_1, x_2, \ldots\ldots x_k$; $i = 0, 1, 2, \ldots\ldots m$; and $j = 1, 2, \ldots\ldots N$.

Note that the intercept term in (1) summarizes the effect of control variables (used to form the strata) on the crash probability and would not be identical across strata. In order to account for stratification in the analysis, a conditional likelihood is constructed. This conditional likelihood function is independent of the intercept terms $\alpha_1, \alpha_2, \ldots\ldots \alpha_N$ [12]. So the effects of matching variables cannot be estimated and (1) cannot be used to estimate crash probabilities. However, the values of the β parameters that maximize the conditional likelihood function would also be the estimates of β coefficients in (1). These estimates are log odds ratios and can be used to approximate the relative risk of a crash.

Consider two observation vectors $x_{1j} = (x_{11j}, x_{21j}, \ldots\ldots, x_{k1j})$ and $x_{2j} = (x_{12j}, x_{22j}, \ldots\ldots, x_{k2j})$ from the j^{th} strata on the k traffic flow variables. The log odds ratio of crash occurrence due to traffic flow vector x_{1j} relative to vector x_{2j} will have the following form:

$$\log\left\{\frac{p(x_{1j})/[1-p(x_{1j})]}{p(x_{2j})/[1-p(x_{2j})]}\right\} = \sum_{i=1}^{k} \beta_i (x_{i1j} - x_{i2j}) \qquad (2)$$

Note that it is the ratio of the resultants obtained by substituting the two observation vectors in equation 1. The right hand side of (2) depends only on β_j, therefore the estimate for log odds ratio may be obtained using the estimated β coefficients. One may utilize the above relative log odds ratio for predicting crashes by replacing x_{2j} with the vector of values of the traffic flow variables in the j^{th} stratum under normal traffic conditions. Simple average of all non-crash observations within the stratum for each

variable may conveniently be used. If $\bar{x}_{2j} = (\bar{x}_{12j}, \bar{x}_{22j}, \bar{x}_{32j} ..., \bar{x}_{k2j})$ denotes the vector of mean values of the k variables over non-crash cases within the j^{th} stratum, then the log odds of crash relative to non-crash may be approximated by:

$$\log\left\{\frac{p(x_{1j})/[1-p(x_{1j})]}{p(\bar{x}_{2j})/[1-p(\bar{x}_{2j})]}\right\} = \sum_{i=1}^{k} \beta_i (x_{i1j} - \bar{x}_{i2j}) \quad (3)$$

Above log odds ratio can then be used to predict crashes by establishing a threshold value that yields desirable classification accuracy [12].

IV. Data Collection and Preparation

Traffic surveillance 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 seconds: average vehicle counts, average speed, and lane detector occupancy (percent of time the loop is occupied by vehicles). These data are collected from three lanes in each direction through 69 stations spaced at approximately 0.8 km (0.5 mile) for a 58-km (36-mile) stretch. The crash data for the study were collected from the FDOT crash database for the years 1999 through 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 pre-crash loop detector data from stations surrounding the crash location were collected based on the adjusted time of historical crashes estimated through a shockwave and rule-based methodology [10]. Traffic data were extracted for the day of crash and on all corresponding (non-crash) days to the day of every crash. The correspondence here means that, for example, if a crash occurred on April 12, 2002 (Monday) 6:00 PM, I-4 Eastbound and the nearest loop detector was at station 30, data were extracted from station 30, four loops upstream and two loops downstream of station 30 for half an hour period prior to the estimated time of the crash for all Mondays of the same season in the year at the same time. Hence, this crash will have loop data table consisting of the speed, volume and occupancy values for all three lanes from the loop stations 26-32 (on eastbound direction) from 5:30 PM to 6:00 PM for all the Mondays of the same season in the year 2002, with one of them being the day of crash (crash case). More details of this sampling technique, application of this methodology and data cleaning could be found in the earlier study by the authors [1].

The 30-second data have random noise and are difficult to work with in a modeling framework. Therefore, the 30-second raw data was combined into 5-minute level in order to get averages and standard deviations. Thus for 5-minute level aggregation half an hour period was divided into 6 time slices. The stations were named as "*B*" to "*H*", with "*B*" being farthest station upstream and so on. It may be noted that "*F*" is the *station of the crash* with "*G*" and "*H*" being the stations downstream of the crash location. Similarly the 5-minute intervals were given "IDs" from 1 to 6. The interval between time of the crash and 5 minutes prior to the crash was named as slice 1, interval between 5 to 10 minutes prior to the crash as slice 2, interval between 10 to 15 minutes prior to the crash as slice 3 and so on. The arrangement used for stations (B-H) and time slices (1-6) used here is crucial for generating the patterns of crash risk and it's "propagation" in a time-space framework.

The parameters were further aggregated across the three lanes and the averages (and standard deviations) for speed, volume and lane-occupancy at 5-minute level were calculated based on 30 (10*3 lanes) observations. Therefore, even if at a location the loop detector from a certain lane was not reporting data, there were observations available to get a measure of traffic flow at that location. Aggregating data across the lanes helps to develop a system for more realistic application scenario since all three lanes at a loop detector stations are less likely to be simultaneously unavailable when the model is used for real-time prediction. Another advantage is that the measures aggregated across lanes not only capture temporal variations (or lack there of) but variations across the three lanes as well.

This dataset consisted of *2046* matched strata included all types of crashes. The type of crash information available in the *FDOT* crash database was utilized to retain only multi-vehicle crashes. Since the ambient traffic characteristics are more likely to affect crashes involving interaction among vehicles rather than the single vehicle crashes that mostly occur during the late night hours. Also, due to intermittent failure of loop detectors the numbers of controls (non-crash cases) available for each case (crash) were not homogeneous. To carry out matched case-control analysis, a symmetric data set was created (i.e., each crash case in the dataset has the same number of non-crash cases as controls) by randomly selecting five non-crash cases for each crash in the dataset. The resulting dataset had *1528* symmetric matched strata available for analysis.

V. Data Analysis

A. Exploratory Analysis and Simple Models

In a logistic regression setting the output of simple (one covariate) models would be the hazard ratio for the parameter used as covariate in the model. The hazard ratio is defined as the exponential of the estimate for model coefficient and represents how much more likely (or unlikely) it is for the crash to occur if the covariate is increased by one unit. Therefore, if the output hazard ratio for a parameter is significantly different from one and, for example, equals two then increasing the value of this variable by one unit would double the risk of a crash around station *F* (station of the crash).

For each of the seven loop detectors (*B* to *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 used one at a time as the risk factor (i.e. independent variable) in the logistic regression model. Exploratory analysis with 5-minute standard deviations and averages of speed showed that the hazard ratios for 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 coefficient of variation in speed was a natural choice as a precursor resulting in hazard ratio values substantially greater than one. Therefore, we combined mean and standard deviation of speed, occupancy and volume into the variables *CVS, CVO, CVV* (coefficients of variation of speed occupancy and volume, respectively, expressed in percentage as (*SS/AS*)**100,* (*SO/AO*)**100,* and *(SV/AV)*100)*. Logarithmic transformation was applied to these coefficients of variation due to skewed nature of their distribution. Further explorations concluded that the variables *LogCVS, AO* and *SV* measured at a range of stations and time-slices had the most significant hazard ratios. To identify time duration(s) and location of loop detector(s) having traffic characteristics significantly associated with the binary outcome (crash vs. non-crash) the hazard ratios were calculated for each of the *126* parameters (7 stations *6 time slices *3 variables i.e., *LogCVS, AO, SV*) through one separate model each. The outcome of each of these models was the hazard ratio corresponding to these variables at various stations and time slices and the *p-value* for the test indicating whether the value is significantly different from unity. It was noticed that the hazard ratio for *LogCVS* increases most significantly as we approach Station F and the time of the crash (Slice 1). The values of hazard ratio for *AO* were low (i.e., only slightly greater than 1.0) yet statistically significant. For *SV* the hazard ratios were found to be significantly *less* than one and tended to decrease as the time and station of crash approached from the downstream direction. It indicated that as *SV* becomes smaller at certain freeway locations the crash risk apparently increases at locations upstream of these sites. It was concluded that in general a higher *LogCVS,* and/or *AO* value and a lower *SV* value would increase the likelihood of crashes.

To understand the patterns of crash risk with respect to time and location of the crash in a time-space framework we generated contour plots of the hazard ratio corresponding to the three parameters (*LogCVS, AO* and *SV)*. One such plot, with hazard ratio for *LogCVS* at various time slice-station combinations as the contour variable, is shown in Fig. 1. These hazard ratios essentially depict the risk for observing a multi-vehicle crash at *Station F*. According to the color scale provided alongside the plot the dark colored regions represent high hazard ratios thereby indicating more risk. It may be observed that region around *Station F* remains fairly dark (i.e., crash prone) for about 20 minute period while upstream and downstream sites (*Station E* and *G*, respectively*)* also show high risk for about 15-20 minute period before recording a crash. These results are significant since they allow leverage in terms of time to predict an impending crash. It is also important to note that the clearest trends in hazard ratio were depicted by the plot corresponding to *LogCVS*, with a stark contrast between locations of crash and other surrounding stations.

B. Multivariate Models

The results from exploratory analysis had shown that three parameters, namely, the *LogCVS, SV* and *AO* are most significantly associated with crash occurrence. These three parameters correspond to 126 variables (three parameters measured from 7 stations during 6 time slices) as potential independent variables for the final multivariate model. Based on the results from the previous section we could discard *Station B, C* and D from consideration in the final model. Even though hazard ratio from these stations were significantly different from unity they were less significant than their counterparts belonging to *Station E, F, G* and *H.*

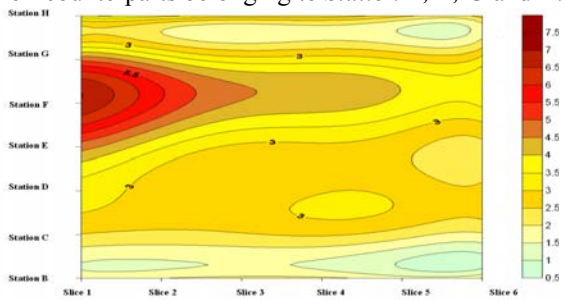

Fig. 1. Spatio-temporal pattern of the hazard ratio for *LogCVS* obtained from 5-minute combined lane dataset for multi-vehicle crashes

Also, even though time slice 1 (0-5 minutes prior to time of the crash) exhibited significant hazard ratios; being too close to the actual time of the crash it was not useful in practice for crash prediction models. This time slice was, therefore, ignored from further considerations. For each of the remaining five time slices (with first slice being ignored), we have $p = 12$ traffic flow variables; *LogCVS, SV,* and *AO* at each of the four loop detectors *E, F, G* and *H*. To identify most significant variables during each time slice among the set of 12 potential variables (three parameters measured at four stations), the binary outcome variable y was modeled using stratified conditional logistic regression method described above in the previous section. The *SAS®* procedure for proportional hazard regression analysis (*PHREG*) allows one to identify significant variables using *stepwise* automatic search procedure. The procedure resulted in three significant variables for time slice 2 (5-10 minutes before crash occurrence): *LogCVS F2* $=log_{10}(CVS)$ from station *F* (the station of the crash) and *AOG2 = AO* at station G (the downstream station) and *SVG2 = SV* at station G (the downstream station). All other variables are 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 decision regarding the selection of the time slice was made based on the classification accuracy achieved from each model. The model developed from slice 2 described above was found to be the best in this regard. Thus, the final model includes variables *LogCVSF2* and *AOG2* and *SVG2*. The details of the final predictive model are provided in Table 1. First two variables have positive beta coefficients (and a hazard ratio greater than 1), which mean that the odds of observing a crash at *Station F* increase as these variables increase while *SVG2* had a negative beta coefficient implying increasing odds of a crash as this parameter decreases.

TABLE 1: FINAL MODEL DESCRIPTION

Variable	Parameter Estimate	p-value	Hazard Ratio
LogCVSF2	1.2140	<.0001	3.367
AOG2	0.0246	<.0001	1.025
SVG2	-0.1912	<.0001	0.826

As previously explained in the modeling methodology section, the odd ratio given by (4) may be used to classify crash and non-crash cases. Following the classification procedure the model provided more than 62% of crash identification on the matched case-control dataset using the threshold of unity for the odd ratio. Note that this threshold (chosen to be equal to one here) may be further varied in order to achieve desirable classification given the tradeoff between overall classification accuracy (crash and non-crash) and crash identification. The threshold of unity provided reasonable balance between the two conflicting attributes (i.e., overall classification and crash identification) and hence is recommended as the cut-off value. The simple models have the advantage due to their data requirement; the decision regarding selection of models must be made based on their classification accuracy. Of all simple models, the one with *LogCVSF2* as the independent covariate happens to be the single most significant model. The crash identification was only *59%* when the single covariate model with *LogCVSF2* was used for classification. It is less than *62.5%* achieved by the multivariate model (with odd ratio cutoff set at 1.0). The multivariate model, therefore, is recommended for a reliable classification of the patterns.

VI. REAL-TIME APPLICATION

A. Phase 1-Simple Model Application

The basic idea for the two-step implementation plan proposed here is to first estimate the measure of crash risk for next 10-15 minutes using the simple models. If there is an indication for a crash then subject the data to the final multivariate model for classification which would assess the crash risk for next 5-10 minutes since parameters in the final model belong to time slice 2 (refer Table 1).

For a real-time application, the instrumented freeway corridor can be divided into 69 (which is the total number of loop detector stations) segments in each direction such that each loop detector remains at the center of each 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). These sets of five stations would correspond to *station D* through *station H* used in the modeling procedure.

The measure for crash risk may be estimated by multiplying the observed *LogCVS* value at these stations with an appropriate time slice 3 hazard ratio which by definition would provide the measure of crash risk relative to the situation if the value for the covariate *(LogCVS)* were zero. In other words, time slice 3 hazard ratio corresponding to s*tation D* would be chosen if the station is most upstream of the set of five, *station H* if it is the most downstream, and, s*tation F* if it is the station belonging to that particular segment and so on. This measure for crash risk may be updated in real-time on a continuous basis as soon as new observations come in. For example, we first calculate the 5-minute level *LogCVS* based on the available ten most recent observations and then after 30-seconds as the latest observation (since loop data is collected every 30 seconds) come in they are included in the calculation of *LogCVS* replacing the far most observation. The measure of crash risk may also be plotted as a contour variable in a time space framework similar to the plots for raw hazard ratios shown in Fig. 1. Based on the changing patterns depicted by the continuously updated plots, freeway locations with high crash risk may be identified in real-time. Since the objective of the paper is to propose a generic plan for traffic surveillance from a safety perspective the authors are not proposing any threshold on the measure of crash risk to determine exactly what value constitutes a high enough risk and would trigger the application of the multivariate model. Such decisions are to be made after exhaustive location specific field testing which is beyond the scope of this generic implementation plan.

B. Phase 2-Multivariate Model

Following the detection of hazardous patterns through the measure of crash risk obtained from simple models the multivariate model may be applied for classification of patterns into leading or not leading to a crash. As explained in one of the previous sections, the log odds can be calculated using (4) to classify the patterns into crash and non-crash cases. For this purpose, we first calculate the mean for the three covariates included in the final model: *LogCVSF2, AOG2* and *SVG2* on all five non-crashes within each matched stratum of the *1:5* matched dataset. For j^{th} matched set, the vector $\bar{x}_{k2j}$ in (4) may be replaced by the

vector of these non-crash means and the most current five-minute data on the three variables for x_{k1j} can be used to calculate the odds ratio for the purpose of identifying a crash. The *RHS* of (4) with estimated values of the parameters from Table 1 can be written as:

$$\exp(1.214(LogCVSF2 - .951) + .024(AOG2 - 13.260) - .191(SVG2 - 2.564)) \quad (5)$$

Note that the average vectors ($\bar{x}_{k2j}$) on the RHS of (4) have been replaced with the respective means of these covariates over non-crash cases in the matched dataset. The values for the three parameters (*LogCVSF2, AOG2* and *SVG2*) obtained from the loop detectors in real-time would be used as independent variables in this expression above to obtain the ratio of odds for having a crash vs. not having a crash. If the resultant odd ratio exceeds unity then the patterns would be classified as a crash. However, note that this threshold would also have to be calibrated through location specific field testing. Data from *station F* and *G* (*LogCVS* from the station of the crash and the AO and SV from the station one immediately following it in the downstream direction) may be collected and updated continuously. To obtain an updated odds ratios every 30-seconds the last set of observations in the 5-minute period may be replaced by the data most recently recorded. In other words the values for *LogCVS, SV* and *AO* are updated on a continuous basis by calculating means and standard deviations of the parameters as moving averages.

VII. Conclusion

A statistical link between turbulent traffic conditions and crash occurrences was established through a detailed analysis of loop detector data corresponding to the multi-vehicle crashes that occurred on the instrumented corridor of Interstate-4 during 1999 through 2002. Following an exploratory analysis a series of simple (involving one covariate) logistic regression models were estimated to deduce the spatio-temporal variation of crash risk. Based on the results from the simple models a multivariate logistic regression model was estimated through a step-wise procedure. For the final model, *average occupancy* and *standard deviation of volume* observed at the downstream station (*Station G*), during the slice of 5-10 minutes prior to the crash (*time slice 2*) along with the *coefficient of variation in speed* at the station closest to the location of the crash (*Station F*) during the same time slice were found to affect the crash occurrence most significantly. It was shown that using 1.0 as the threshold for the log odds ratio, over 62% crash identification was achieved from the final model on the matched case-control dataset. This modeling approach may be extended to any freeway similarly equipped with loop detectors although model parameters would need to be calibrated using field data from that freeway.

A real-time application plan for these models was demonstrated in the paper. Essentially the proposed plan states that a preliminary assessment of the freeway conditions may be made using the measure of crash risk assessed using simple models and if this measure indicate high risk of crash occurrence for next 10-15 minutes; the data may be further subjected to the multivariate model for classification. If the classification model identifies patterns from the detectors as crash prone then the traffic management authorities can keep the incident mitigation squads on alert in anticipation of a crash so that the impact of crash occurrence on freeway operation may be minimized. At this point the traffic safety application of the plan proposed here is limited and more aggressive strategies such as variable speed limits, warning drivers through variable message signs etc., need to be explored. These techniques would allow more proactive intervention and help reduce the crash potential. Another point of consideration while devising these strategies would be that although all multi-vehicle crashes were included in the modeling procedure, the methodology presented here might result in better identification of rear-end crashes, which are the most common type of crashes on freeways.

References

[1] Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., and Hsia, L., Predicting freeway crashes based on loop detector data using matched case-control logistic regression. Forthcoming in the *Transportation Research Record,* 2004.

[2] Hughes, R., and Council, F., 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., 1999.

[3] Lee, C., Saccomanno, F., and Hellinga, B., Analysis of crash precursors on instrumented freeways. *Transportation Research Record 1784*, 2002, pp. 1-8.

[4] Lee, C., Hellinga, B., and Saccomanno, F., Real-time crash prediction model for the application to crash prevention in freeway traffic. *Transportation Research Record 1840*, 2003, pp. 67-78.

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

[6] Abdel-Aty, M., and Pande, A., Classification of real-time traffic speed patterns to predict crashes on the freeways. Presented at the *83rd Annual Meeting of the Transportation Research Board* (TRB), Washington D.C., 2004.

[7] Golob, T. and Recker, W., Alvarez, V., Freeway safety as a function of traffic flow**.** *Accident Analysis & Prevention, Volume 36, Issue 6,* November 2004, pp. 933-946.

[8] Golob, T. , Recker, W. and, Alvarez, V., A tool to evaluate the safety effects of changes in freeway traffic flow. *Journal of Transportation Engineering – ACSE, 130,* 2004, pp. 222-230.

[9] Golob, T. and Recker, W., A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A: Policy and Practice, Volume 38, Issue 1,* January 2004, pp. 53-80.

[10] Abdel-Aty,M., A., Pande, A., Hsia L., and Abdalla, F., The potential of loop detector data for improving safety. Forthcoming in the *ITE Journal*, 2005.

[11] Pande, A., Abdel-Aty, M. and Hsia L., Spatio-temporal variation of risk preceding crash occurrence on freeways. Accepted for presentation at the *84th Annual Meeting of Transportation Research Board*, Washington, D.C., 2005.

[12] Collett, D., Modelling binary data*. Chapman and Hall*, 1991.