A Freeway Safety Strategy for Advanced Proactive Traffic Management

ANURAG PANDE and MOHAMED ABDEL-ATY

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 crashes. A simple approach to identify freeway locations with high probability of crashes through real-time traffic surveillance data is presented here. The crash and loop detector data for the study was collected from 36-mile corridor of Interstate-4 in Orlando, Florida. The analysis is based on simple (one covariate) logistic regression models developed under a matched study design. Individual traffic parameters obtained one at a time from series of loop detectors have been examined as potential covariates to these models. Hazard ratio for each individual covariate is the output from the models. Based on the hazard ratio and its statistical significance it was found that the log of coefficient of temporal variation in speed, standard deviation of volume, and average occupancy expressed as percentage are the parameters that are most critically associated with potential occurrence of multivehicle crashes. The univariate logistic regression models were validated based on their classification performance on an independent set of crash data. Using the relative location of loop detectors measuring these parameters with respect to the crash location contour maps depicting spatial-temporal distribution of crash risk was generated. Using the model outputs, a generic strategy to assess crash risk in real-time is also proposed. With this strategy one can identify the segment of freeway having high potential for crash occurrence within next 15–20 minutes. The crash mitigation and law enforcement set up can be prepared for dispatch to such locations based on the real-time assessment of crash potential.

INTRODUCTION

Incident detection models have been the center for attention in traffic management literature. However, in the recent past the focus seems to be shifting toward more proactive strategies. Diminishing relevance of the incident detection and increased capability to collect, store, and analyze data have all contributed towards this shift. A recent series of studies examining potential crash prediction models underline this new trend. It must be understood however that these studies are often primarily aimed at improving traffic safety since it is reasonable to assume that crashes are the most predictable type of incidents.

The idea of crash prediction essentially involves collection of crash data and establishing statistical links between crash

occurrence and traffic data emanating from detectors surrounding the crash location prior to its occurrence. Due to the dynamic nature of the data involved, the modeling methodologies differ distinctly from traditional crash frequency analyses where static measures of traffic flow parameters are used.

Madanat and Liu (1995) were one of the first researchers to introduce the concept of proactive traffic management. They perceived the importance of incorporating prior probability of incident occurrence into detection models. Two types of incidents, namely crashes and overheating vehicles, were considered. Binary logit was the methodology used for analysis. They concluded that the merging section, visibility, and rain are statistically the most significant factors for crash likelihood prediction.

Hughes and Council (1999) explored the relationship between freeway safety and peak period operations using loop detector data. It was concluded that the “traffic flow consistency”

as perceived by the driver may be an important factor in freeway safety. Lee et al. (2002) hypothesized that the likelihood of a crash is significantly affected by short-term turbulence of traffic flow. Speed variations along the length of the roadway (i.e., difference between the speeds upstream and downstream of the crash location) and also across the three lanes at crash location were identified as significant crash precursors. Another important factor identified by them was the traffic density at the instant of the crash. With these variables, a crash prediction model was developed using log-linear analysis. In a later study Lee et al. (2003) continued their work along similar lines and modified the aforementioned model. They incorporated an algorithm to get a better estimate of the time of the crash and the length of time period (prior to the crash) to be examined. It was found that the average variation of speed difference across adjacent lanes doesn't have direct impact on crashes and hence was eliminated from the model. They also concluded that variation in speed has a relatively longer-term effect on crash potential than do either traffic density or average speed difference between the upstream and downstream ends of roadway sections.

A study by Oh et al. (2001) showed standard deviation of speed in a 5-minute interval to be the best indicator of "disruptive" traffic flow leading to a crash as opposed to "normal" traffic flow. They used the Bayesian classifier to categorize the two possible traffic flow conditions. Probability distribution functions for each class are required for application of Bayesian classifiers. Standard deviation of speed during 5-minute interval belonging to crash and noncrash cases, respectively, were fitted to non-parametric distribution functions using kernel smoothing techniques. Due to a limited sample size of only 52 crashes aforementioned classifiers remain far from being implementable in the field.

Research aiming at freeway crash prediction through loop data was also carried out by Golob and Recker (2001, 2004) in which they established statistical links between crashes, environmental factors and traffic flow as obtained from the loop data. Their findings, however, are limited by the fact that the traffic data is obtained from single loop detectors and speed has to be estimated using a proportional variable (volume/occupancy). The Flow Impacts on Traffic Safety (FITS) tool developed by Golob and Recker (2004) also has its limitations because of a systematic pattern of missing data with in the data used for development of this tool. The geometric characteristics of the freeways are not considered by this tool.

The data used in these studies were obtained from just one station downstream and/or upstream of the crash location. All of these studies have overlooked the "progression" of alarming driving conditions with the flow of traffic. Despite their analytical shortcomings and unresolved issues related to implementation, these studies demonstrated the possibility of determining crash potential at a certain freeway location (or section) in real-time using traffic surveillance data. It is also important to note that if a crash prediction model has to be useful in preventing crashes one needs to identify the crash prone conditions much ahead of the crash occurrence time and not by using data from

0–5 minutes prior to crash occurrence. In which case, the traffic management authorities would not have sufficient time for analysis, prediction and dissemination of information. Two of the previous studies (Abdel-aty and Pande, 2004 and Abdel-Aty et al., 2004) by the authors also utilized loop detector data to develop crash prediction models, through neural network and logistic regression approach, respectively. Both models indicated the significance of the coefficient of variation in speed and that a time period (slice) of 5–10 minutes before the crash occurrence could be used in crash prediction.

This study, by analyzing the data from a series of detectors at different time increments, accounts for the possibility that alarming crash prone conditions on a freeway might actually originate upstream/downstream and "travel" with traffic until they culminate into a crash at a certain time at a downstream/upstream location. Also since the modeling approach presented here uses one covariate at a time and doesn't employ complex models using information from different detectors at different times it is more appealing for real-time application. An implementation plan to assess and update the real-time crash potential on freeways is also demonstrated in the paper.

STUDY AREA AND AVAILABLE DATA

For this study four years of crash data (from 1999 through 2002) were collected from 36-mile of Interstate-4 corridor equipped with loop detectors in the Orlando metropolitan area. The 36-mile segment provides a diverse database with respect to crash, traffic, and geometric characteristics and ensures that the findings of the study are more readily transferable to other freeways. The freeway stretch under consideration has a total of 69 loop detector stations, spaced out at approximately 1/2 mile. Through dual loop detectors (sensors located beneath the pavement) these stations collect and store the following measurements every 1/2 minute for three through lanes in each direction:

- a) Volume (number of vehicles passing each lane in 30 seconds),
- b) Lane-occupancy (percentage of the 30-second interval the loop detector was occupied), and
- c) Average speed (of all vehicles passing over the loop detector in the 30-second interval).

The other component of the data was the crash characteristics for the 3,755 crashes that occurred in the aforementioned corridor during the period of analysis. The location, time, and type of crash were obtained from the Florida Department of Transportation (FDOT) crash database.

MATCHED CASE-CONTROL LOGISTIC REGRESSION

Simple Models: Methodology

The within stratum matched case-control logistic regression is adopted to identify the relationship between traffic parameters

measured through loop detectors and crash occurrences while controlling for external parameters such as the location (i.e., the geometric characteristics), time of the day, day of the week, and season (Abdel-Aty et al., 2004).

For a simple logistic regression model the function of dependent variables yielding a linear function of the independent variables would be the logit transformation.

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

Where $\pi(x) = E(Y|x)$ is the conditional mean of Y (dummy variable representing crash occurrence) given x when the logistic distribution is used. If we assume that the logit is linear in the continuous covariate x , then the equation for the logit would be given by Equation 1. It follows that the slope coefficient, β_1 , is the change in the log odds for an increase of 1 unit in x , i.e. $\beta_1 = g(x + 1) - g(x)$ for any value of x (Agresti, 2002).

The hazard ratio (also known as risk ratio) for an explanatory variable with regression coefficient β is defined as $\exp(\beta)$. 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 larger than 1, an increment in the factor increases the hazard rate. If the hazards ratio is less than 1, an increment in the factor decreases the hazard rate (SAS[®] Institute, 1990).

Data Preparation

Based on the mile-post location for each of the 3,755 crashes (available from FDOT crash database) the loop detector station nearest to its location was determined. This station is referred to as the station of the crash from here on. The next step was to extract precrash loop detector data from the archived loop detector database. Data permitting the time of historical crashes was estimated from a rule based shock-wave methodology, the details of which may be found in Abdel-Aty et al. (2005). The data needs to be prepared as per the requirements of the matched case-control logistic regression technique; therefore if a crash is reported to occur on April 12, 1999 (Monday) 6:00 PM, I-4

Eastbound and the nearest loop detector was at station 30, data were extracted from station 30, five loops upstream and one loop downstream of station 30 for half an hour period prior to the reported time of the crash for all the Mondays of that particular year at the same time. This matched sample design was created in order to control for roadway and geometric factors and driver population on the freeway (e.g., more commuters on weekday peak hours, indicating more young to middle age drivers, etc.). Hence, this crash will have loop data table consisting of the speed, volume and occupancy values for all three lanes from the loop stations 25–31 (on eastbound direction) from 5:30 PM to 6:00 PM for all the Mondays of the year 1999, with one of them being the day of crash. The data were available for only 2046 (out of 3755) crashes. During the time of the remaining crashes some of the loops, from which data were required, were not functioning.

The loop detectors sometime suffer from intermittent hardware problems that result in unreasonable values of speed, volume and occupancy. These values include Occupancy >100 (percent), speed = 0 or >100 (MPH), flow >25 (vehicles per lane), and flow = 0 (vehicles per lane) with speed >0 (MPH) and were removed from raw 30-second data. From the “cleaned” data tables the average and standard deviation of speed were extracted over each lane for six 5-minute intervals recorded prior to the crash on the station nearest to the crash location (referred to as station of the crash), five stations upstream and one station downstream of the station of the crash. It requires creation of 252 fields (7 stations by 6 time slices by 3 lanes by 2 variables [comprising the average and standard deviation of speed]) in the database for each crash. The same 252 fields were extracted for all “corresponding” noncrash days as well.

The nomenclature procedure adopted for defining the station and time slice to which the average and standard deviation belongs is shown in Figure 1. All the stations were named as “B” to “H”, with “B” being farthest station upstream and so on. It should be noted that “F” is the station closest to the crash location and “G” is the first station downstream of the crash location. Similarly the 5-minute intervals were also given “ID” 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, and interval between 10 to

Figure 1 The nomenclature for defining the station and time slice to which any “effect” (average or standard deviation) belongs.

15 minutes prior to the crash as slice 3 and so on. The arrangement of stations and time slices shown in Figure 1 is crucial for generating the patterns of crash risk and its “propagation” in a time-space framework.

If we use average and standard deviation of traffic parameters only from the specific lane of the crash it reduces the size of the dataset to about 30% of the original crash sample due to the fact that data from specific lane of the crash were missing quite often. By aggregating the data on the three lanes in the aforementioned dataset the lane of the crash averages and standard deviations were replaced by values aggregated over three lanes. In this dataset, the averages (and standard deviations) at 5-minute level were based on at most 30 (10×3 lanes) observations.

We now had two datasets, one with average and standard deviation of traffic parameters calculated over six 5-minute slices with data from lane of the crash only, and the other with data from all three lanes aggregated together. Two more datasets similar to the above were created with the only difference being that the average and standard deviation were computed over ten 3-minute slices as opposed to six 5-minute slices.

In the combined lane datasets (average and standard deviations aggregated over three lanes), even if at a certain station the loop detector from one lane was not reporting data there were observations available to get a measure of traffic at that location. This not only increases the sample size of crashes but also 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 difference between the two datasets is that while the combined lane dataset accounts for the variation (or lack thereof) across the lanes, the individual lane of the crash dataset does not. It may also be noted that average and standard deviations calculated over five minute time slice would be more effective in the crash prediction as it provides more allowance in terms of time to analyze data, estimate and possibly reduce the likelihood of crashes.

Due to the reasons cited above, and exploring all four datasets including 3-minute and 5-minute aggregation with individual lane of the crash/combined lanes, we found that the 5-minute level combined lane dataset is superior and therefore this dataset will be used in the analysis presented in this article. Also, the hazard ratios obtained through this dataset are the ones used to generate spatio-temporal patterns of crash risk.

Type of the crash information available with the FDOT crash database was utilized to prepare the 5-minute level combined lane dataset by only retaining the multivehicle crashes. Since traffic conditions are more likely to impact the crashes involving interaction among vehicles rather than the single vehicle crashes which mostly occur due to errors on the drivers’ part. The final dataset encompassed 1,528 strata with each stratum consisting of one crash and all available corresponding noncrash cases. For these strata complete loop data corresponding to the crash cases were available.

Due to data availability issues, there were different numbers of noncrash cases for each crash. To carry out matched case-

control analysis we created symmetric dataset (i.e., each crash case in the dataset has the same number of noncrash cases as controls) by randomly selecting five noncrash cases for each crash. The results from this symmetric dataset are discussed in the following section. The choice of selecting five as the number of corresponding noncrash cases was based on one of our earlier findings (Abdel-Aty et al., 2004). In that article five separate datasets having crash vs. noncrash ratio as 1: m and m varying from 1 through 5 were analyzed. Results from these five datasets showed no significant differences in their findings and hence it was decided to choose $m = 5$.

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

Analysis and Discussion

For each of the seven loop detectors (B to H) and six time slices (1–6) mentioned above, the values of 5-minute averages (AS , AV , AO) and standard deviations (SS , SV , SO) of speed, volume and occupancy, respectively, are available for all crashes and the corresponding noncrash cases. Exploratory analysis with these original effects showed that the hazard ratio 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 the skewed nature of their distribution. The preliminary analysis also indicated that the variables $LogCVS$, AO and SV had the most significant hazard ratios.

The results of stratified conditional simple (one variable at a time) logistic regression analysis were then analyzed for these three variables ($LogCVS$, AO , SV) at each of the seven loop detectors and six time slices to identify time duration(s) and location of loop detector(s) whose traffic characteristics are significantly correlated with the binary outcome (crash vs. non-crash). This was done by calculating the hazard ratio using proportional hazard regression analysis (*PHREG*) procedure of SAS[®] for each of the 126 ($7 \text{ stations} * 6 \text{ time slices} * 3 \text{ parameters i.e., } LogCVS, AO, SV$) single variable models; one model for each of the three variables $LogCVS$, AO and SV over every station B – H and the duration of time slice 1–6. The outcome of these models was the hazard ratio value for these parameters at various stations and time slices and the p -value for the test indicates whether the value is significantly different from one. The hazard ratio is an

Table 1 Hazard ratios for *LogCVS*, *SV* and *AO* for combined lane parameters measured at 5-minute level during six different time slices and seven stations

Station	Time slice	<i>LogCVS</i>			<i>SV</i>			<i>AO</i>		
		Hazard ratio	chi-sq.	<i>p</i> -value	Hazard ratio	chi-sq.	<i>p</i> -value	Hazard ratio	chi-sq.	<i>p</i> -value
<i>B</i>	1	1.902	22.9338	<.0001	0.918	6.8264	0.009	1.025	23.2826	<.0001
<i>B</i>	2	1.924	23.6444	<.0001	0.924	5.7255	0.0167	1.027	26.7403	<.0001
<i>B</i>	3	2.456	44.2759	<.0001	0.929	5.0971	0.024	1.03	31.7949	<.0001
<i>B</i>	4	1.852	21.2755	<.0001	0.944	3.0388	0.0813	1.024	21.4939	<.0001
<i>B</i>	5	1.873	21.738	<.0001	0.913	7.7708	0.0053	1.027	28.8965	<.0001
<i>B</i>	6	2.024	26.6929	<.0001	0.939	3.5473	0.0596	1.025	22.6704	<.0001
<i>C</i>	1	2.516	49.1638	<.0001	0.94	3.8469	0.0498	1.03	37.2577	<.0001
<i>C</i>	2	2.258	39.2117	<.0001	0.95	2.5053	0.1135	1.032	41.4822	<.0001
<i>C</i>	3	2.279	39.1038	<.0001	0.907	8.8419	0.0029	1.029	34.4992	<.0001
<i>C</i>	4	2.709	58.845	<.0001	0.969	0.9697	0.3248	1.029	35.7133	<.0001
<i>C</i>	5	2.215	35.9575	<.0001	0.924	5.8427	0.0156	1.032	40.4467	<.0001
<i>C</i>	6	2.002	28.3212	<.0001	0.956	2.099	0.1474	1.028	32.5375	<.0001
<i>D</i>	1	3.331	73.4693	<.0001	0.887	13.1378	0.0003	1.035	49.7569	<.0001
<i>D</i>	2	3.132	65.5422	<.0001	0.886	13.1844	0.0003	1.035	44.2893	<.0001
<i>D</i>	3	2.43	40.1837	<.0001	0.933	4.4608	0.0347	1.036	47.3307	<.0001
<i>D</i>	4	3.074	63.8774	<.0001	0.879	14.8615	0.0001	1.037	48.185	<.0001
<i>D</i>	5	2.735	51.8499	<.0001	0.894	11.6207	0.0007	1.036	46.9616	<.0001
<i>D</i>	6	2.499	42.8092	<.0001	0.877	14.9141	0.0001	1.036	43.6042	<.0001
<i>E</i>	1	4.436	123.4947	<.0001	0.879	16.8999	<.0001	1.051	103.5387	<.0001
<i>E</i>	2	3.335	82.6848	<.0001	0.88	16.0371	<.0001	1.05	97.1036	<.0001
<i>E</i>	3	3.025	72.4558	<.0001	0.865	20.6933	<.0001	1.046	84.5517	<.0001
<i>E</i>	4	3.257	79.46	<.0001	0.856	23.9313	<.0001	1.044	77.2271	<.0001
<i>E</i>	5	2.664	55.4917	<.0001	0.846	27.8307	<.0001	1.042	72.1956	<.0001
<i>E</i>	6	2.426	45.6673	<.0001	0.862	21.5363	<.0001	1.044	75.3528	<.0001
<i>F</i>	1	7.237	246.3844	<.0001	0.822	42.8032	<.0001	1.052	120.8515	<.0001
<i>F</i>	2	5.58	182.5256	<.0001	0.844	30.9312	<.0001	1.055	128.7943	<.0001
<i>F</i>	3	4.485	151.4043	<.0001	0.842	31.7211	<.0001	1.054	133.1569	<.0001
<i>F</i>	4	3.801	120.7948	<.0001	0.856	25.9338	<.0001	1.048	106.0348	<.0001
<i>F</i>	5	3.654	111.6472	<.0001	0.867	22.3968	<.0001	1.049	100.3394	<.0001
<i>F</i>	6	3.809	116.4946	<.0001	0.865	22.4915	<.0001	1.046	96.1598	<.0001
<i>G</i>	1	4.705	153.769	<.0001	0.806	47.8694	<.0001	1.054	115.1149	<.0001
<i>G</i>	2	3.899	112.8892	<.0001	0.851	27.7995	<.0001	1.049	95.1991	<.0001
<i>G</i>	3	3.037	80.7431	<.0001	0.791	54.8999	<.0001	1.049	101.6789	<.0001
<i>G</i>	4	3.519	101.9139	<.0001	0.853	26.8096	<.0001	1.053	110.2725	<.0001
<i>G</i>	5	3.209	82.9753	<.0001	0.825	37.4876	<.0001	1.045	84.6264	<.0001
<i>G</i>	6	2.964	72.1287	<.0001	0.862	22.2225	<.0001	1.044	80.3909	<.0001
<i>H</i>	1	3.976	105.4923	<.0001	0.846	26.805	<.0001	1.054	101.7469	<.0001
<i>H</i>	2	3.635	87.8667	<.0001	0.809	41.2345	<.0001	1.052	91.6664	<.0001
<i>H</i>	3	3.476	81.9115	<.0001	0.825	35.9595	<.0001	1.055	100.6115	<.0001
<i>H</i>	4	3.139	69.9456	<.0001	0.823	36.9363	<.0001	1.053	96.651	<.0001
<i>H</i>	5	2.623	50.2791	<.0001	0.873	18.6939	<.0001	1.051	88.7727	<.0001
<i>H</i>	6	2.871	59.8772	<.0001	0.84	30.0702	<.0001	1.05	84.8237	<.0001

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 (e.g., 2) then increasing the value of this variable by one unit would double the risk of a crash at station *F* (station of the crash). Table 1 shows the results of all single covariate models for *LogCVS*, *SV*, and *AO*. The table shows how the hazard ratio for *LogCVS* and *AO* increases as we approach the Station of the crash (Station *F*) and time of the crash (Slice 1). Although the values of hazard ratio for *AO* is low (i.e., closer to 1.0) but still is significant (Note the chi sq. statistic and *p*-value). The reason for the low value is that occupancy usually changes by 1% quite frequently on freeways and it is more meaningful to

represent the increased risk of observing a crash resulting from 10% increase in occupancy. This modified hazard ratio can be obtained by raising hazard ratio to the power 10. For *SV* the hazard ratios are less than one and tend to be decreasing as the time and station of crash approached in the downstream direction. Since it is the value of hazard ratio significantly different from one (and not necessarily a higher value) that makes the variable a better crash precursor, ratio for *SV* indicates that as this parameter becomes smaller at certain freeway locations the crash risk apparently increases at locations upstream of these sites.

Based on these results, it can be argued that a higher *LogCVS*, *AO* value and lower *SV* value increases the likelihood of crashes.

While for *LogCVS* this trend is observed starting at about 1 to 1.5 miles (from Station D) upstream of the crash location, it is considerably clear at about 1/2 mile upstream and also downstream. It is also clear, based on the rise observed in hazard ratios that the “ingredients” for a crash starts at about 15 minutes before the crash. The *LogCVS* factor represents high variation in speed relative to the average speed, and the *SV* factor represents low variation in volume. Lower speed associated with high variance (leading to a high value of coefficient of variation) depicts turbulence in traffic that could be explained by frequent formation of queues followed by their quick dissipation. Significant hazard ratios for high occupancy downstream of the crash site indicate initiation of a queue formation causing a backward forming shockwave leading to unstable traffic and high potential of crash occurrence near the station of the crash.

Hazard ratios significantly less than one for *SV* parameters indicate that low variability in volumes is positively correlated with crash occurrences on freeways. A possible interpretation of this criterion might be that in case of high variability in volume, the density changes and consequently the gaps between vehicles change which alert the drivers. On the other hand, in case of low variability in volume, the density and the gap remain almost fixed in the traffic stream which causes the drivers to relax thus slowing their reaction time. It could also be that low variability of volume might sometimes be associated with queues (although low variability can also occur in better level of service with no queues). Also, low standard deviation of volume, with all three lanes combined, not only represents very low temporal variance in volume but almost same number of vehicles on three lanes as well. This coupled with high variation in speed at these locations, might cause drivers to make lane changing maneuvers near to the station of the crash in order to maintain their speeds. This will result in increased likelihood of conflict between vehicles. In general, however, queue formation and shockwaves are a common cause of rear-end crashes on freeways. It must be pointed

out that available loop detector/crash data does not provide us with enough resolution to estimate the number of lane changing maneuvers or to determine exact mechanism of crashes in order to verify some of the aforementioned postulations. The analysis merely shows *association* of low *SV* values downstream of the crash site with crash occurrence and no conclusions regarding the *causality* of crashes should be drawn from this discussion.

As an obvious extension, the analysis was followed up with multivariate model building. The stepwise procedures resulted in a model with three significant variables for time slice 2 (5–10 minutes before crash occurrence): $LogCVSF2 = \log_{10}(CVS)$ 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). We then examined interactions among these three parameters. As it turned out the no interaction terms were significant other than the interaction between *LogCVSF2* and *AOG2*. The results from the multivariate models are consistent with the findings of the simple models.

Spatio-Temporal Variation of Crash Risk

The crash risk for the multi-vehicle crashes corresponding to the observed values of 5-minute combined lane *LogCVS*, *AO* and *SV* is shown in Figure 2(a), 3, and 4, respectively. Note that in Figure 2(a) and 3 the dark colored region represents high hazard ratios thereby identifying more risk while in Figure 4 the dark regions of the plot represent low hazard ratios (the values corresponding to *SV* are less than 1), but still signify more risk (of having a crash around Station *F*) associated with corresponding time slice and location.

As we can see in all three plots (2[a], 3 and 4) region around Station *F* remains fairly dark (i.e., crash prone) for about a 20-minute period while upstream and downstream sites (Station *E* and *G*, respectively) also show high risk for about

Figure 2(a) Spatio-temporal pattern of the hazard ratio for *LogCVS* obtained from 5-minute level combined lane dataset for multivehicle crashes.

Figure 2(b) Spatio-temporal pattern of the hazard ratio for *LogCVS* obtained from “pseudo” crash case dataset.

15–20 minute period before recording a crash. These results are significant since they allow leverage in terms of time to be able to anticipate an impending crash. It is however important to note that the most clear trend is depicted by the plot corresponding to *LogCVS*, since a stark contrast may be seen between location of crash and surrounding locations. Plot corresponding to *SV* (Figure 3) appears dark for locations downstream of the crash location which indicates that low variance in volume coupled with high variation in speed at freeway locations (say *Station G*) increases odds of having a crash upstream (*Station F*) of that site. However, the trends aren’t as clear about location of the crash as they were in the case of *LogCVS*. It is also to be seen in the context that the hazard ratios for *LogCVS* were more significant than those of *SV*.

To assess the fact that these results are really depicting an association between traffic flow variables and crash occurrence

and not some random patterns in the data we also estimated hazard ratios for 126 covariates (7 stations by 6 time slices by 3 parameters [comprising the *LogCVS*, *AO*, and *SV*]) using the “pseudo” crash matched dataset. As expected the trends were either nonexistent (as was the case with *LogCVS* and *SV* with hazard ratios not significantly different than unity) or they were reversed (as was the case with *AO* with hazard ratio significantly less than one). The plot with hazard ratios corresponding to *LogCVS* obtained from “pseudo” dataset give an idea about “normal” conditions on freeways (see Figure 2[b]). The plot is in perfect contrast with its counterpart in Figure 2(a) that shows hazard ratio for *LogCVS* from the real matched case control dataset. It provides visual evidence for the contribution of traffic factors toward crash occurrence.

The results show that even if the first time slice (0–5 minutes prior to a crash) is excluded due to practical considerations of

Figure 3 Spatio-temporal pattern of the modified hazard ratio (increase in crash risk when there is ten unit increase in occupancy rather than one) for *AO* obtained from 5-minute level combined lane dataset for multivehicle crashes.

Figure 4 Spatio-temporal pattern of the hazard ratio for SV obtained from 5-minute level combined lane dataset for multivehicle crashes.

the time required to act on the information, it was shown that the crash prone conditions in terms of high coefficient of variation in speed, low variation in volume and high occupancy are not ephemeral on freeway sections. Based on these findings the models developed with 5-minute level combined lane dataset excluding single-vehicle crashes were selected to demonstrate the implementation plan for field application in the following section.

Validation of Model Performance

The results from the pseudo case-control dataset indicated the differences in freeway traffic patterns under crash prone and noncrash conditions but they do not automatically guarantee nonoccurrence of false-alarms. The reason is that crashes, however frequent on the freeway segment under consideration here, are still rare events. Hence, it is a good idea to validate the univariate logistic regression models based on their classification performance on independent set of data.

To evaluate the classification performance of these models we extended our database and added 381 new crashes and corresponding loop detector data from first six months of the year 2003. The matched *noncrash* loop detector data corresponding to these crashes were also added to the database. These crash and noncrash data points were used as the validation set for the models. Single variable logistic regression models developed in this study were applied on the validation dataset based on the classification methodology developed in one of our earlier papers (Abdel-Aty et al., 2005b). In this section the details of classification performance of one such model involving *LogCVSF2* as the covariate have been discussed. *LogCVSF2* has been selected as the covariate to demonstrate the classification performance since it is the most significant univariate model among all.

Based on the value of parameter *LogCVSF2* and the β coefficient a measure of crash risk was determined for each ob-

servation in the validation dataset. This measure of risk was the parameter used for classification. The observations in the validation dataset could be classified using a threshold value on this measure of crash risk. However, one needs to examine the classification accuracy of the model by evaluating the performance of the model at various threshold values for this measure. The classification accuracy is of course sensitive to the threshold used and therefore an arbitrary selection of the threshold is not preferable. To assess the classification accuracy of the model at various threshold values, cumulative proportions of crashes above and non-crashes equal or below a range of this *measure of risk* values were determined and were plotted against the measure of crash risk in Figure 5. For convenience, measure of risk threshold less than or equal to 5 are shown on the horizontal axis.

In Figure 5, the grey (lighter) curve indicates the cumulative proportion of crash cases that have measure of risk greater than the corresponding measure of crash risk on the horizontal axis and the black (darker) curve indicates the cumulative proportion of noncrash cases with measure of risk less than or equal to the corresponding risk value on the horizontal axis. One may choose a threshold value along the horizontal axis and determine the proportions of crashes and noncrash cases that would be correctly classified by the model under consideration. For example, if measure of risk equal to one is chosen as the cut off point, then little more than 59% of crashes and noncrash cases in the dataset would be correctly classified by the single covariate model with *LogCVSF2* as the independent variable.

The plot shown in Figure 5 could be useful in selecting a measure of risk value that would satisfy the requirement of a desired accuracy. Note, however, that both (crash and noncrash) classification accuracies cannot be increased simultaneously and there is a trade-off involved. Decision for the threshold needs to be made carefully by keeping the real-time application in perspective. For example, during a free-flow operation period a lower value may be set as threshold so that most of the crashes

Figure 5 Classification performance of the simple model with *LogCVSF2* as covariate: Cumulative proportion of crashes above and noncrash cases below a range of measure of risk threshold values.

are identified even if that increases the number of “false-alarms” because speed is known to be positively associated with the severity of crashes.

If the classification approach and cutoff values are to be used for real-time crash “prediction” they must first be calibrated carefully. The cutoff would vary from a freeway to another and by location, time of day, day of week on the same freeway. The calibration of the threshold values should be done by the specific agency (e.g., a Regional Transportation Management Center) that wishes to apply the findings of this study. At present we are more interested in presenting a generic way to analyze real-time freeway traffic data so that the relative level of short-term crash risk on the freeway may be assessed. An implementation plan for such a strategy is provided in the following section. However, the description of classification performance of the logistic regression model involving *LogCVSF2* as covariate is sufficient to validate the promise that this methodology has to offer.

IMPLEMENTATION PLAN

Procedure and Data Requirement

The single covariate models need information from one loop detector station at a time. It makes these models particularly attractive given that few loops often tend to malfunction in practice. The output for each of the simple models developed in one of the previous sections 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 crash risk relative to the situation if the value

of the covariate were zero. This parameter, defined in the following equation, has been chosen as the measure of crash risk to identify the high risk location in real-time in the implementation plan.

$$\text{Measure of crash risk} = \text{Parameter value} \times \text{hazard ratio corresponding to the parameter}$$

For example, according to the above definition the *measure of risk* for having a crash in the vicinity of any station (*station F*) within next 5–10 minutes due to *LogCVS* value of 1.2 at a station located two station upstream (*station D*) would be $1.2 \times 3.132 = 3.7584$ (3.132 is the hazard ratio corresponding to the parameter *LogCVSD2*).

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),

Table 2 Hazard ratios from single covariate models consisting of *LogCVS* from five stations and six time slices

Hazard ratio corresponding to station	Hazard ratio to assess 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)
<i>D</i>	3.331	3.132	2.430	3.074	2.735	2.499
<i>E</i>	4.436	3.335	3.025	3.257	2.664	2.426
<i>F</i>	7.237	5.580	4.485	3.801	3.654	3.809
<i>G</i>	4.705	3.899	3.037	3.519	3.209	2.964
<i>H</i>	3.976	3.635	3.476	3.139	2.623	2.871

Table 3 Snap shot of 5-minute *LogCVS* (values updated every 30-seconds) calculated as a moving average starting 15 minutes prior to crash occurrence

Date-time	Station	Station of crash	<i>LogCVS</i>
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

as defined earlier in the article. 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 *Station D–H* (two upstream stations, station closest to the crash location and two stations downstream, respectively). Note that the hazard ratios from station *B* and station *C*, the two stations located farthest upstream of the station of the crash, were not as critically associated with crash occurrence as those from sta-

tion *D* to station *H*. Therefore, the set of loop detectors for the implementation plan consists of only five stations as opposed to seven used for the analysis. The hazard ratios corresponding to *LogCVS* measured at these five stations (*D–H*) at all six time slices are shown in Table 2. Among the three possible parameters *LogCVS* was chosen because the plot depicting the spatio-temporal variation of crash risk (Figure 2) showed stark contrast between the station of the crash and other locations. Note that 5-minute level combined lane database with only multi-vehicle crashes was used to estimate these hazard ratios.

With the hazard ratio for *LogCVS* from station *D* to station *H* (Table 2) one can observe the change in 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 come in. For example, we first calculate the *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 foremost observation. The *LogCVS* measured at different stations may be used to calculate the measure of crash risk for a period up to next thirty minutes by multiplying the corresponding hazard ratio with the *LogCVS* value. In other words, hazard ratio corresponding to Station *D* would be chosen if the station is most upstream of the set of five, *Station G* if it is the most downstream, and, *Station F* if it is the station belonging to that particular segment and so on. Decision about the time slice to be chosen for the hazard ratio value depends upon how much time ahead we need the information, that is, to obtain the crash risk within the next 10–15 minute hazard ratio belonging to *slice 3* should be chosen while for next 5–10 minutes *slice 2* hazard ratio will be used. The measure of crash risk may then be plotted as a contour variable in a time-space framework similar to the plots for hazard ratio shown in the previous section. Based on the changing patterns depicted by the continuously updated plots, freeway locations with high crash risk may be identified in real-time.

In this section we illustrate how the patterns in the crash risk may be observed through the contour plots with historical loop detector data belonging to a crash and a noncrash case. Table 3 shows a sample of *LogCVS* calculated as a moving average from real-life historical traffic speed data from a set of five detectors,

Table 4(a) The measure for risk of observing a crash in the segment belonging to *station F* with in next 30 minutes at time 4:19:30 PM

Measure of risk according to <i>LogCVS</i> 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

Table 4(b) The measure for risk of observing a crash in the segment belonging to station F with in next 30 minutes at time 4:20:00 PM

Measure of risk according to <i>LogCVS</i> 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

starting 15 minutes prior to time of the crash. Note that data was collected prior to a real crash that occurred on April 6, 1999 near station 34 at 4:35 PM on Eastbound Interstate-4. Note that the formulation for *LogCVS* remains the same as in the modeling phase, the details of which may be found in the previous sections.

Table 2 depicted the hazard ratios corresponding to station D–H at all six time slices. In Table 4(a–c) the process for calculating the values for the contour variables (measure of crash risk obtained by multiplying *LogCVS* values with corresponding hazard ratio) is shown. In the first row in Table 4(a), 1.42, which is the *LogCVS* value obtained at station 32 (corresponds to Station D of the analysis) during five minute period of 4:14:30 to 4:19:30 PM, is multiplied by the hazard ratio for station D at each time slice (1–6) to obtain the measure of crash risks up to next half hour. 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 five-minute period. Third, fourth, and fifth row of the table are made up by the hazard ratio corresponding to stations F, G, and H multiplied by the value of *LogCVS* at these stations. To assess the risk for various future time periods the same value of *LogCVS* is used, however the value of hazard ratio is based on the time slice, for example, for next 10 minutes hazard ratio corresponding to slice 2, for next 15 minutes hazard ratio corresponding to slice 3 and so on.

Table 4(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 4(c) the

value of the independent covariate (*LogCVS*) are further updated with the most recent speed observations. It may be noted that in Table 4(a) the values of *LogCVS* are highlighted in yellow (light color) to associate them with the observations in Table 3 from the same point in time (4:19:30 PM). Similarly, in Table 4(b) and 4(c), the updated values for *LogCVS* are highlighted red (dark) and green (medium) to associate them with their respective times of observation (4:20:00 and 4:20:30 PM, respectively) in Table 3.

It is important to note that the *measure of risk* assesses the risk of observing a crash in the vicinity of the middle station of the set of five stations based on the *LogCVS* values at all five stations. The *LogCVS* at station F is expected to impact the risk of having crash at this location more than the *LogCVS* observed at any other station. Since the hazard ratios for *LogCVS* from station F are higher than those from station E, the *measure of risk*, which is the multiplication of *LogCVS* value and the corresponding hazard ratio, can sometimes be higher even though *LogCVS* itself at station F is lower than *LogCVS* at station E.

In the example chosen for demonstration, although the *LogCVS* at station 33 (station E) is higher than those at station 34 (station F) the *measure of risk* is higher for station 34 since we are trying to assess the risk of a crash near station 34 for it being the middle of the set of five (32–36) chosen for demonstration. The high risk of crash at station 33 (indicated by high *LogCVS* value) would be demonstrated if we choose to analyze the preceding set of five stations consisting of stations 31–35. In that case (not discussed in the article) we would be assessing the risk of crash at station 33 (middle station of the set of five)

Table 4(c) The measure for risk of observing a crash in the segment belonging to station F with in next 30 minutes at time 4:20:30 PM

Measure of risk according to <i>LogCVS</i> 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

Figure 6(a-c) Illustrative pattern of variation in measure for risk of observing a crash in the segment belonging to station F updated every 30 seconds 15 minutes prior to a crash.

and LogCVS from station 33 would be multiplied by the hazard ratio corresponding to station F.

Three contour plots depicting the variation in crash risk generated from this data are shown in Figure 6(a-c). It can clearly be seen that the region about station F remains dark indicating high risk for a crash occurrence. It may be noted that the values for contour variable in Figure 6(a) comes from the corresponding cells of Table 4(a) and the plot is updated into Figure 6(b) as soon as the new set of readings are recorded after 30 seconds. The values for contour variable in the updated plot, Figure 6(b), are given by Table 4(b) which eventually turns into Figure 6(c) after 30-seconds taking input from Table 4(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 and only three observations out of thirty are updated after 30-seconds. These figures may be contrasted with similar patterns generated for the same time of the day prior to a corre-

sponding matched noncrash case (On April 27, 1999 from the same set of stations) shown in Figure 7(a-c). In a real-time application of the models these measures of risk may be calculated continuously and the corresponding plots can be generated using the color scheme depicted on the side of each contour plot. According to the color scale the dark (red) colors represent the regions of the contours where the measure of crash risk exceeds 6.0. There is no such region in Figure 7(a-c). It should be noted that the difference between the crash and noncrash case is highlighted here to illustrate the application, in some other cases; however, the difference may not be as clear. If there is a consistent pattern of high risk (depicted by the red [dark] colors) then the authorities should consider it as a warning. Note that a generic color scheme is proposed here and no exact threshold has been recommended because these thresholds would have to be calibrated for every station based on factors such as the time of the day.

Figure 7(a-c) Illustrative pattern of variation in measure for risk of observing a crash in the segment belonging to station F updated every 30 seconds for a noncrash scenario.

SUMMARY AND CONCLUSIONS

The objective of this research endeavor was to develop a simple implementation strategy for proactive traffic management involving real-time assessment of crash risk on freeways. A detailed crash database was assembled for all multi-vehicle crashes that occurred on the instrumented corridor of Interstate-4 in the years 1999–2002. It was shown statistically that turbulence in traffic conditions before a crash (both time and space) is associated with crash occurrence. This means that we can flag “dynamic black spots” in real-time if such turbulence is observed in the future.

Case-control logistic regression with a matched study design was used as the analysis technique. The matched design of the study accounts for external factors such as the freeway geometry, time of the day and day of the week. Following the exploratory analysis, a series of simple (involving one covariate) logistic regression models were estimated for multi-vehicle crashes based on the statistical link between crash occurrence and the turbu-

lence in the traffic flow observed through the loop detectors. The models were validated based on the classification performance on an independent set of crash and non-crash data belonging to first six months of the year 2003.

It was shown that the results from these models could be used to obtain spatio-temporal variation of the crash risk. A generic real-time application plan for these models was also demonstrated. The plan proposed here essentially assesses the freeway conditions with respect to probable crash occurrence based on the spatio-temporal distribution of the hazard ratios for the parameter *LogCVS*. In this study no specific threshold for the measure of crash risk was proposed to separate crashes from noncrash cases since it would vary based on several factors such as the location, geometry, time of the day, day of the week, and would have to be calibrated based on the exact application scenario.

If the conditions are identified as hazardous, traffic management authorities can keep their crash mitigation squad close to such locations on alert so that the impacts of crashes may be

minimized. It would enhance incident detection capabilities of the traffic management centers.

Although the results from this study are more likely to be used for operational purposes; potential planning applications exist as well. For example, if there are some freeway segments where the plots continuously output crash prone patterns, these segments may be closely watched through freeway cameras at certain times of the day. This would help recognize any problems associated with these locations such as weaving sections, ramps, etc. that lead to hazardous traffic conditions. It might be difficult to identify these problems through traditional frequency analysis.

This strategy may potentially be used in order to develop and undertake short-term preventive measures to avoid crashes in real-time. Aggressive intervention strategies would be required to calm down the prevailing hazardous conditions on the freeway. The authors believe that the problem of intervening with measures such as variable speed limits or flashing warning on the Variable Message Sign and thereby reducing the freeway crash potential is a non-trivial one and demands separate attention. The potential impacts of such techniques on the driver behavior will also need to be assessed. The authors maintain that it is an issue that demands separate attention and is beyond the scope of this article.

ACKNOWLEDGEMENT

The authors wish to thank the Florida Department of Transportation for funding this research. The UCF data warehouse provided the data that was used in the analysis. All opinions and results are those of the authors and do not reflect the opinion of any other party.

REFERENCES

- Abdel-Aty, M. and Pande, A. (2004). Classification of real-time traffic speed patterns to predict crashes on freeways. Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington D.C.
- Abdel-Aty, M., Uddin, N., Abdalla, F., Pande, A., and Hsia, L. (2004). Predicting freeway crashes based on loop detector data using matched case-control logistic regression. *Transportation Research Record, No. 1897, Transportation Research Board*, National Research Council, Washington, D.C., 88–95.
- Abdel-Aty, M., Pande, A., Hsia, L., and Abdalla, F. (2005a). The potential of loop detector data in improving freeway safety. Forthcoming in the *ITE Journal*.
- Abdel-Aty, M., Uddin, N., and Pande, A. (2005b). Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. Forthcoming in the *Transportation Research Record*.
- Agresti, A. (2002). *Categorical data analysis, 2nd Ed.* John Wiley and Sons, Inc., New York, NY.
- Golob, T. and Recker, W. (2001). Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *California PATH Working Paper UCB-ITS-PWP-2001-19, Institute of Transportation Studies*. Berkeley, University of California.
- Golob, T. and Recker, W. (2004). A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A: Policy and Practice*, **38**, 53–80.
- Hughes, R. and Council, F. (1999). On establishing relationship(s) between freeway safety and peak period operations: Performance measurement and methodological considerations. Presented at the 78th Annual Meeting of Transportation Research Board, Washington, D.C.
- Lee, C., Saccomanno, F., and Hellinga, B. (2002). Analysis of crash precursors on instrumented freeways. *Transportation Research Record, No. 1784, Transportation Research Board*, 1–8. National Research Council, Washington, D.C.
- Lee, C., Saccomanno, F., and Hellinga, B. (2003). Real-time crash prediction model for the application to crash prevention in freeway traffic. *Transportation Research Record, No. 1840, Transportation Research Board*. National Research Council, Washington, D.C. 67–78.
- Madanat, S. and Liu, P. (1995). A prototype system for real-time incident likelihood prediction. *IDEA Project Final Report (ITS-2)*, *Transportation Research Board*. National Research Council, Washington, D.C.
- Oh, C., Oh, J., Ritchie, S., and Chang, M. (2001). Real time estimation of freeway accident likelihood. Presented at the 80th Annual Meeting of Transportation Research Board, Washington, D.C.
- SAS[®] Institute Inc. (1990). *SAS/STAT[®] User's Guide. Version 6, Fourth Ed.* Cary, NC.