Florida Department of Transportation
Research Project Work Order #8, Contract No. BC-355

Final Report

Linking Crash Patterns to ITS-Related Archived Data

By

Mohamed Abdel-Aty, Ph.D., PE
Associate Professor

Anurag Pande
Ph.D. Candidate

Nizam Uddin, Ph.D.
Associate Professor

Haitham Al-Deek, Ph.D., PE
Professor

Essam Radwan, Ph.D., PE
Professor

Department of Civil & Environmental Engineering
University of Central Florida
P.O.Box 162450
Orlando, Florida 32816-2450
Phone : 407 823-5657
Fax : 407 823-3315
Email : mabdel@mail.ucf.edu

November 2004

EXECUTIVE SUMMARY

This report describes the development of real-time crash prediction models for the Interstate-4 corridor in Central Florida area. Crash data for 36.25-mile freeway stretch from the year 1999 through 2002 has been used to link the crash occurrences with real-time traffic patterns observed through loop detector data.

The analysis technique adopted for this phase of the study is with-in stratum matched case-control logistic regression. 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. In the context of this research crash or non-crash is the binary outcome with traffic parameters being the independent variables. The design of the study allows controlling for external factors such as geometric design of the freeway, time of the day, day of the week, etc., and hence they are implicitly accounted for them.

Using this technique two types of models, i.e., simple and multivariate, were developed. Prior to development of the models some of the data related issues such as data cleaning, determination of exact time of the historical crashes, etc. were addressed. Both types of models were evaluated based on their classification performance. It was observed that although the simple models have the advantage of being tolerant in their data requirements their classification accuracy is poorer than that of the final multivariate model. Hence, the simple models were used to deduce spatio-temporal patterns of the variation in crash risk. As a suggested application for these 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 real-time application strategy is also provided in the report.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In recent years the focus of traffic management seems to be shifting from reactive strategies such as incident detection to more proactive ones. Growing concern over traffic safety as well as increased capability to store and process the data has contributed towards this shift. Reliable models that can identify conditions and/or locations of high crash risk based on available real-time sensor data is the most critical part of these proactive strategies.

This report presents the findings of the first phase of an ongoing research effort to link ITS-archived data with crash characteristics/occurrences. The study is currently underway at University of Central Florida in collaboration with Florida Department of Transportation (*FDOT*). As mentioned in the proposal for this research, safety applications of ITS-related archived data have been almost non-existent. There are some related studies which recently concluded or are currently in progress at various parts of the world. However, the advances made by this study go beyond the scope of other studies and put state of Florida *DOT* in the forefront of developing real-time crash prediction models. The models developed as part of this project use real-time data from a series of loop detectors and assess whether or not a crash is likely to occur on instrumented segments of the freeway under consideration. Although further research in this regard is still in progress, the following objectives have been achieved in this phase:

1. A detailed database has been created with relevant characteristics of all the crashes that occurred from January 1999 through December 2002 on the 36.25-mile instrumented corridor of Interstate-4.

2. Identify the location and time of occurrence for all crashes and extract the corresponding loop detector data from the archived ITS database. Inspect the trends in crash patterns over the four year period of analysis.

3. Perform a matched case-control logistic regression analysis in order to identify the spatio-temporal patterns for the variation of crash risk.

4. Based on this statistical methodology, develop a classification model to differentiate between crash prone and normal conditions on the freeway.

5. Demonstrate the online applicability of the methodology for generating the patterns of variation in crash risk at freeway segments along with the classification accuracy.

The report is divided into five chapters in addition to the introduction. The next chapter provides a thorough and critical review of studies aiming at proactive freeway management systems through crash prediction models. The third chapter introduces the study area and describes the data preparation efforts. Note that the data preparation work for this study was carried out with future scope of research in perspective. Subsequent chapter summarizes the modeling technique followed by development of the models. The performance of these models is also evaluated in this chapter. Fifth chapter presents the implementation plan for the models developed here. The summary, conclusions and future scope of this work have been presented in the last chapter.

CHAPTER 2

BACKGROUND

2.1 Safety Applications of ITS-related Archived Data

A stated focus of traffic safety analysis has been on determination of freeway crash patterns. There have been a lot of studies establishing links between traffic flow and crash characteristics. Most of the research analyzing crashes on freeways have been based on the aggregated and static measures of traffic flow variables and have developed frequency type models. Only recently, some research aimed at proactive freeway management through analysis of freeway surveillance data has gained momentum. This chapter presents a summary and critical review of such studies.

2.1.1 Exploratory Studies

Hughes and Council (1999) were one of the first authors to explore the relationship between freeway safety and peak period operations using loop detector data. They concluded that the macroscopic measures, such as *AADT* (Average Annual Daily Traffic) and even hourly volume, in fact, correlate poorly to real time system performance. Their work mostly relied upon the data coming from a single milepost location during the peak periods of the day, on which they tried to overlay the crash time at that particular location to infer about the changes in system performance as it approaches the time of the crash. The changes in the performance were also examined from "snapshots" provided by cameras installed on the freeway.

One of their important conclusions was that as "design inconsistency" has been identified as a factor of crash causation, future research should also consider whether "traffic flow consistency"

as perceived by the driver is an important variable from a human factors standpoint. They also expressed a need for determining the exact time of the crash to avoid "cause and effect" fallacy.

2.1.2 Studies Establishing Statistical Links

Madanat and Liu (1995) came up with an incident likelihood prediction model using loop data as input. The focus of their research was to enhance existing incident detection algorithms with likelihood of incidents. They actually considered two types of incidents *a)* crashes and *b)* overheating vehicles. Binary logit was the methodology used for analysis. They concluded that merging section, visibility and rain are statistically the most significant factors for crash likelihood prediction. However, they acknowledged problems with their data.

Lee et al. (2002) introduced the concept of "crash precursors" and hypothesized that the likelihood of crash occurrence is significantly affected by short-term turbulence of traffic flow. They came up with factors such as speed variation 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 the crash location. Another important factor identified by them was traffic density at the instance of the crash. A crash prediction model was developed using log-linear analysis. According to the authors the log-linear model was chosen for analysis so that the exposure can be easily determined, which would have been difficult, if instead a logit model was used. In order to test the goodness of fit for the model, Pearson chi-square test was performed. The test measured how close the expected frequencies are to the observed frequencies for any combination of crash precursors and control factors. At 95 % confidence level the model yielded a good fit.

In a later study (Lee et al., 2003), they continued their work along the same lines and modified the aforementioned model. They incorporated an algorithm to get a better estimate of time of the crash and the length of time slice (prior to the crash) duration to be examined. They concluded that variation of speed has relatively longer term effect on crash potential rather than density and average speed difference between upstream and downstream ends of roadway sections Also they 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.

Oh et al. (2001) showed that five minutes standard deviation of speed value was the best indicator of "disruptive" traffic flow leading to a crash as opposed to "normal" traffic flow. They used the Bayesian classifier to categorize the two possible traffic flow conditions. Since Bayesian classifier requires probability distribution function for each class, they fitted their crash and no-crash speed standard deviation data to non-parametric distribution functions using Kernel smoothing techniques.

In one of the more detailed analysis of patterns in crash characteristics as a function of real-time traffic flow is done by Golob and Recker (2001, 2002). The methodology used was non-linear (nonparametric) canonical correlation analysis (NLCCA) with three sets of variables. The first set comprised a seven-category segmentation variable defining lighting and weather conditions; the second set was made up of crash characteristics (collision type, location and severity); and the third set consisted of real-time traffic flow variables. Since NLCAA requires reducing collinearity in the data, principal component analysis (PCA) was performed to identify relatively

independent measurements of traffic flow conditions. The results of the PCA are shown in Table 2-1.

Table 2-1 Interpretation of principal components and variable selection (Golob and Recker, 2001)

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

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

While almost all studies have indicated a relationship between crash occurrence and speed variability, a recent study by Kockelman and Ma (2004) found no evidence to the fact that speeds measured as 30-second time series or their variations trigger crashes. The study was conducted for the same area as Golob et al. (2003). Their sample size was limited to 55 severe crashes that occurred during January 1998 and with such a small sample their conclusions remain suspect.

Our group at University of Central Florida has also been actively involved in research linking crash patterns with loop detector data. Various modeling methodologies have been explored e.g., Probabilistic Neural Network (Abdel-Aty and Pande, 2004), matched case-control Logistic Regression (Abdel-Aty et al. 2004), Multi Layer Perceptron and Radial Basis Function neural network architectures (Pande, 2003) and Generalized Estimation Equation (Abdel-Aty and Abdalla, 2004). The data for these studies were collected from 13.2-mile central corridor of Interstate-4 in Orlando. All these studies made significant contributions towards enriching the literature, however, it must be said that there was enough room for improvement.

2.1.3 Critical Review

It is evident that the idea of exploring the loop data in safety research is still in its preliminary stages. Some of the aforementioned studies do promise about their application in future, but they have not fully analyzed the "recipe" of crashes. This is besides the fact that the statistical analysis in some cases isn't really sound from a theoretical point of view.

The research going on in Canada (Lee et al. 2002, 2003) has the advantage over the other research groups in the sense that their loops are placed very near to each other (38 loops on 10 km stretch of the freeway), data is more precise (@ 20 seconds) and moreover they have dual loop detectors. The analysis they have been doing is based on frequency model and not on a model which can be used in real time crash prediction application. Also the way some of the variables are calculated from the loop data (e.g. coefficient of variation of speed) doesn't have good statistical basis. The timeframe that they are using for the crash precursors range from 2 to 8 minutes, which could not be sufficient for processing and intervention.

The Golob and Recker (2001) study has established sound statistical links between environmental factors, traffic flow as obtained from the loop data but their findings 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 classification model developed by Oh et al. (2001) seems to have the most promising online application, also demonstrated in their study, but due to the lack of crash data (only 52 crashes) their model remains far from being implemented in the field. The only factor used for classification is the 5-minute standard deviation of speed, other significant factors such as geometry, weather and other traffic flow variables were not considered. It is also to be understood that if a crash prediction model has to be useful we need to classify the data much ahead of the crash occurrence time and not just 5-minutes prior to the crash so that traffic management authorities have some time for analysis, prediction and dissemination of the information.

There are certain key issues, which either have been overlooked or proper attention has not been given to them. One of them is the determination of exact time of the historical crashes. Except for Lee et al. (2003) all the studies have either relied on the police records or at the most visual inspection of the loop data plots. Even the algorithm developed by Lee et al. (2003) has errors associated with shock-wave progression speed. None of the studies except for those conducted at *UCF* have analyzed data from series of loop detectors in order to examine progression of crash prone conditions on freeways.

Neither of the studies has incorporated driver related characteristics into a crash prediction framework. The critical review shows a sufficient scope of improvement in the field of crash prediction not only in terms of analysis techniques but in data related issues (e.g., time of crash, incorporating driver characteristics, etc.) as well. In this study we have addressed some of these issues such as examination of data from a series of loop detectors, time of the crash estimation etc., while for the others an extension to this effort has already been proposed to *FDOT* and should be addressed in the second phase of the project. The data preparation chapter in this report reflects that the database has been prepared with the intention of overcoming all the limitations of these studies in the next phase of the project.

CHAPTER 3

STUDY AREA AND DATA PREPARATION

3.1 General

The final goal of this research is to develop a predictive system for crash occurrence on Interstate-4 corridor equipped with underground loop detectors. To achieve this objective we need to systematically correlate between the crash characteristics and the loop data (representing ambient traffic flow configuration). Moreover it has to be collated with the geometric design of the freeway at the location of the crash and the environmental conditions at the time of the crash. The system needs to recognize the patterns not leading to crash occurrence as well, hence traffic, environmental and geometric conditions corresponding to selected "non-crash" cases or "normal" freeway operating conditions must also be a part of the database. Drivers belonging to certain groups are known to have high likelihood of being involved in crashes, therefore, a measure for driver characteristics should also be included in the database.

The traffic parameters in this study would be measured in terms of time series of 30-seconds observed from inductive loop detectors in the vicinity of the crash location for a certain period leading up to the crash. It is not difficult to realize the importance of properly fusing the loop detector data with crash data and geometric/environmental/driver related factors that might affect the probability of crash occurrence.

3.2 Introduction to Study Area

The study is being conducted on the Interstate-4 (I-4) corridor in Orlando. The corridor is considered to be an integral part of Central Florida's transportation system. It carries greater

number of people and vehicles than any other transportation facility in the region and serves many of the area's primary activity centers. Though originally designed to serve long distance travelers, the I-4 corridor now has evolved to one serving many shorter trips. No wonder a significant amount of growth in the region is occurring within close proximity to I-4. In recent years, congestion on I-4 has extended well beyond normal peak hours and major crashes have closed the freeway, subsequently resulting in traffic congestion throughout the Orlando metropolitan area. Hence, congestion and delays blended with high crash rates are the major transportation problems facing the freeway.

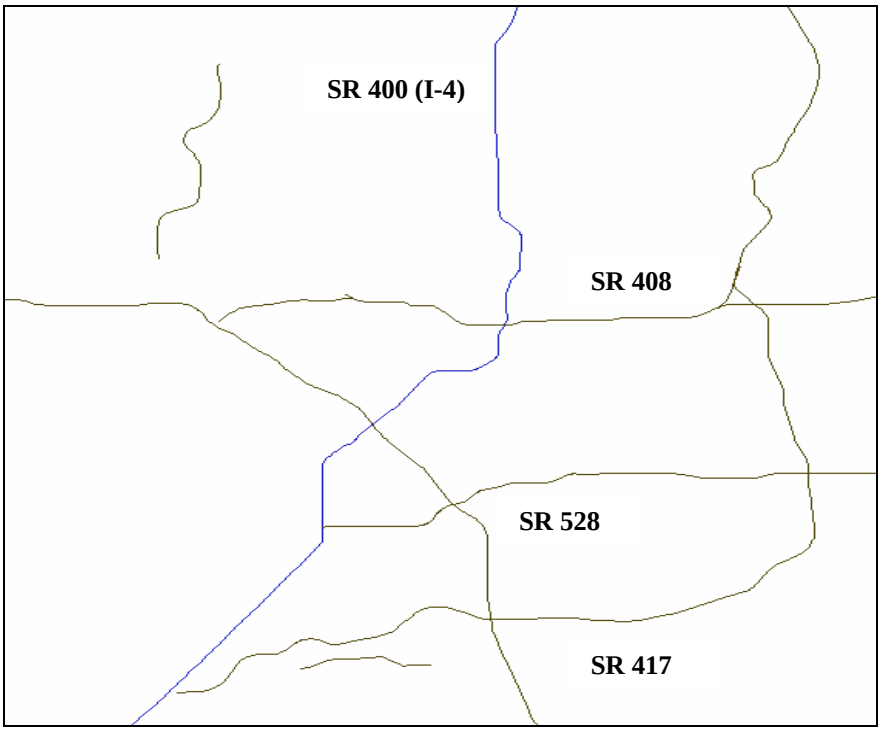

Figure 3-1: I-4 corridor under consideration along with other major roads

Figure 3-1 shows the instrumented Interstate-4 corridor along with the some major roads on the network. The freeway section under consideration is 36.25 miles long and has a total of 69 loop detector stations, spaced out at nearly half a mile. Each of these stations consists of three dual loops in each direction and measures average speed, occupancy and volume over 30 seconds

period on each of the through travel lane. The loop detector data are continuously transmitted to the transportation management center (*TMC*). The source of crash and geometric characteristics data for the freeway is *FDOT* (Florida Department of Transportation) intranet server.

3.3 Crash Data Collection

The first step was to collect crash data for the instrumented freeway corridor over a period of time. Since the loop detectors are known to suffer from intermittent failures it was likely that some of the crashes may not have corresponding loop data available. To ensure that loop data for sufficient number of crashes are available to establish reliable links between crash and traffic characteristics it was decided to be on the conservative side and collect crash data for a period of four years ranging from 1999 through 2002.

There were 3755 crashes reported in all during the four year period (from 1999 through 2002), while we expected some of them to have corresponding loop detector data missing, it was believed that we will be left with a sample large enough for analysis purposes. However, the information extracted for each crash case to create a complete crash database for is shown in Table 3-1.

Table 3-1 The crash characteristics table

Crash Number	Crash report number	Direction (EB or WB)	Mile post	Date of crash	First harmful event	Lane of the crash	Visibility on the roadway	Pavement Condition (Wet, slippery or dry)	Number of fatalities	Number of injuries
1	xx	xx	xx	xx	xx	Xx	xx	xx	xx	xx
2	xx	xx	xx	xx	xx	Xx	xx	xx	xx	xx
|	|	|	|	|	|	|	|	|	|	|
3755	xx	xx	xx	xx	xx	Xx	xx	xx	xx	xx

The table shown above provides sufficient information about each crash; the field "first harmful event" represents *type of the crash.* All other fields are self explanatory. The "milepost" field of the crash characteristics table (Table 3-1) was used to determine the loop detector station nearest to location of each crash and was referred to as the station of the crash. As we will see later, not all the crash characteristics have been analyzed in this phase of the study. None the less, they were made part of the database with future research in perspective.

3.4 Estimation of Time of Historical Crashes

3.4.1 Background

Since the pre-crash loop detector data patterns are being linked with crash characteristics, the time of historical crashes used for analysis becomes very critical. The reason being that if the reported time of the crash is for example, 10 minutes later than the actual time of crash occurrence it would lead to a "cause and effect" fallacy as pointed out by Hughes and Council (1999).

As mentioned in Chapter 2, this issue has not been thoroughly addressed in the literature. The past studies have relied either upon the time obtained from police records or at the most through visual inspection of the loop data plots. Also, there are errors associated with assumptions made by Lee et al. (2003) which happens to be the only study addressing the issue somewhat in detail.

3.4.2 Loop Data used to Estimate Time of the Crash

Since the first objective was to estimate the accurate time of the 3755 crashes, the loop detector data from the station of the crash, 4 upstream stations and 2 downstream stations were collected

for a period of 90 minutes around the reported time (one hour prior and half an hour later) of every crash. The period of 90 minutes was chosen to clearly locate the time when the shockwaves strike the concerned loop detector stations. The two loop detectors downstream of the crash location help to detect the existence of forward recovery shockwave. The most critical part of this methodology is to estimate the speed of the backward forming shockwave. Being on the conservative side and examining data from four upstream stations helps to detect the time of shockwave hit at these stations and ensures that we are able to estimate the speed of the shockwave even for the case when one or two stations are not functioning. Note that for estimating the time of crashes, the loop data in its raw form as time series of 30-seconds were used. Out of these 3755 crashes 1705 crashes did not have any loop detector data available, i.e., none of the seven detectors from which data were sought were functioning on the day of these crashes. The remaining crashes had at least partial data available but there was no assurance that all three lanes from all seven detectors were reporting data.

The loop detectors are known to suffer from intermittent hardware problems that result in unreasonable values of speed, volume and occupancy. Values that include Occupancy>100, speed=0 or >100, flow>25, and flow =0 with speed>0, were removed from the raw 30-second data.

3.4.3 Impact of Crashes on Traffic Flow

Crashes are a specific type of incident and generally have more profound impact on freeway operation. The effects of a crash on traffic flow patterns develop over time both upstream and downstream of the crash. However, the changes in traffic flow characteristics are distinct on loop

detectors located upstream and downstream directions. On the upstream direction, a queue is observed to form, resulting in significant reduction in speed accompanied by an increase in lane-occupancy. On the other hand, decrease in lane flow and occupancy is observed downstream. The critical aspect for determining the time of crash is the time elapsed in the progression of the shockwave from the crash location to the upstream loop detector station. In general this duration (i.e. the shockwave speed) and changes observed in the loop data are affected by the severity of that crash, the roadway geometry, the presence of on- and off-ramps, the distance between loop detector stations, and prevailing traffic flow conditions (Adeli and Karim, 2000).

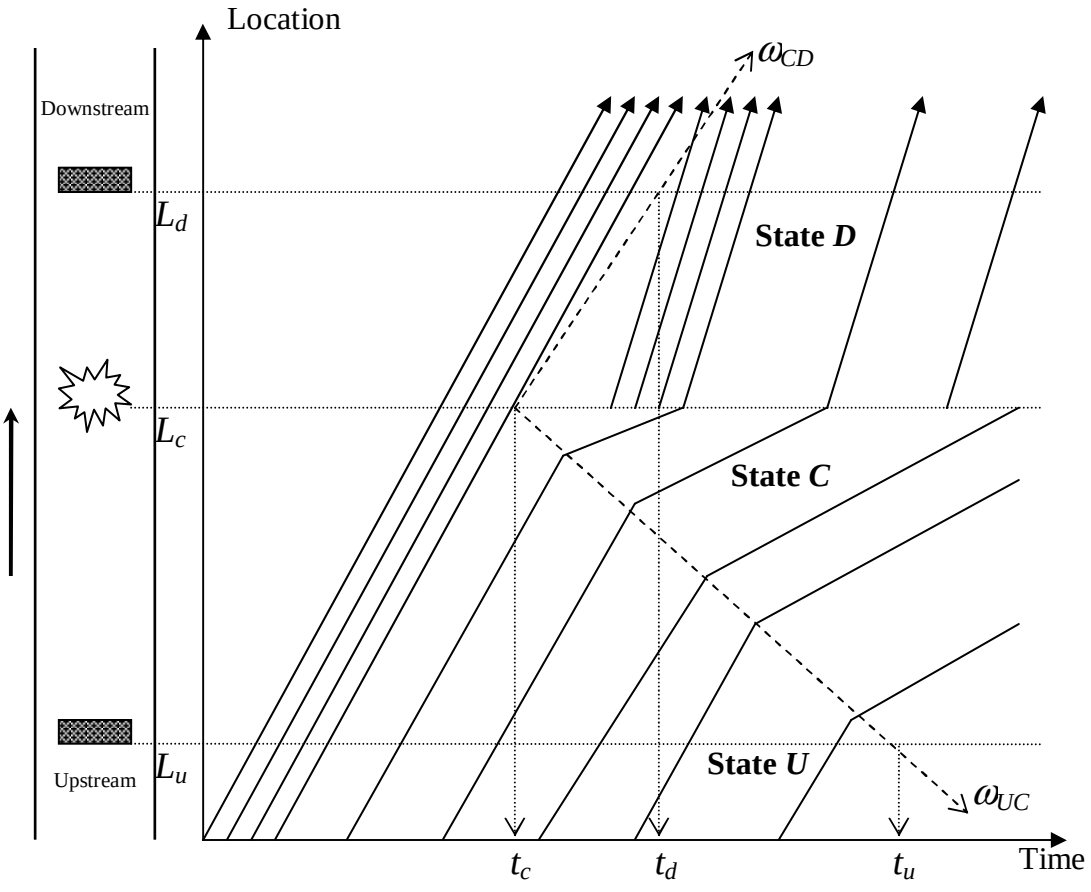

Figure 3-2: Time-space diagram in the presence of a crash
(based on Lee et al. 2002)

The impact of a crash under the assumption of a constant shockwave speed may be shown by a time-space diagram (Figure 3-2). L_d and L_u represent the location of detector stations downstream and upstream of the crash site, respectively. The time t_c, t_d and t_u are time of the crash and time of backward forming shockwave arriving at downstream and upstream stations, respectively. It is clear from the figure that if the speed of backward forming shockwave is known then the time of the crash could be easily estimated. In the Figure 3-2 ω_{UC} and ω_{CD} represent the speeds of backward forming and forward recovery shockwaves, respectively. The times of backward forming shockwave hitting two adjacent upstream stations may be determined by observing when the drops in speed profiles of the two stations occur. The gap between the two arrival times is the time that the shockwave takes to travel from first upstream station to the next upstream station.

3.4.4 Time of the Crash: Estimation and Validation

First step in estimating the time of the crash was to estimate the speed of the backward forming shockwave resulting from the crash. The difference between times of shockwave arrival at the two adjacent stations located immediately upstream of the crash location was used. Since the milepost of all loop detectors on I-4 was known accurately, distance between the two detectors could be used to get the shockwave speed. Once the shockwave speed is known it is not difficult to determine t_c, using the milepost of crash location (also known from the FDOT crash database). The following equation may be used for the estimation:

$$t_u - t_c = \frac{(L_u - L_c)}{\omega_{UC}}$$

All the variables in the above equation have the notation used in Figure 3-2. Due to the underlying assumption made here, that shockwave speed remains constant while it hits the first

and second stations in the upstream direction, it was mandatory to validate the results. The critical issue in the validation was that there is no way to know the actual time of the crash (true value) to compare the shockwave model estimates with. The model was validated using the traffic simulation package *PARAMICS*. A small freeway section on Interstate-4 was simulated and three traffic flow statistics (speed, volume and density) were obtained from locations separated half mile apart on the section just as the loop data is archived for Interstate-4 in real time. Crashes were configured to occur at various locations between a set of two detectors (e.g., very near to upstream or downstream loop, exactly midway between the loops, etc.). The simulation experiment showed that the time of these "artificial" crashes could be accurately estimated using the shockwave method under various scenarios.

3.4.4.1 Aggregation across lanes vs. using lane of the crash

After the methodology was developed and validated as explained above, it could either be applied by aggregating the data across three lanes or by using the data from the specific lane on which the crash had occurred. Lane of the crash was known from the FDOT crash database. The advantage of using the aggregated data was that the time of the crash could be estimated for a large sample of crashes, since the data for at least one of the lanes is obviously available for more crashes than the data for a specific lane. On the other hand since the algorithm relies on the impact of shockwave hitting at successive upstream stations, sometimes the aggregated data (averaged over three lanes) might dampen this impact and the drop in speed or rise in occupancy may not be significant enough to be detected by the algorithm as a shock-wave hit. Hence, it was decided to apply the algorithm for the specific lane of the crash for each case.

3.4.4.2 Crashes at Different Locations

Although results of the above algorithm were validated on the simulation data it was necessary to understand some complexities involved before applying it on the real data, for example for the crashes which occur on the median it is almost impossible to detect any effect on upstream loop detectors. Since even the "rubber neck" effect dies down before being felt at the station immediately preceding the crash location. Hence the algorithm was further examined and validated by looking at speed and occupancy profiles obtained at stations immediately upstream for randomly selected crashes. These crashes were selected from different roadway locations (such as the 3 mainstream lanes, median, shoulder, auxiliary lanes) in order to identify the lanes from which a clear pattern of sudden drop in the loop detector speed data could be observed.

The visual inspection of profiles of several crashes from aforementioned roadway locations led to formulation of following rules:

- For crashes on Left, Center or Right Main Traffic-stream lanes: Estimate time of the crash by applying the existing methodology on the data from the respective lane (i.e., lane of the crash).
- 4^{th} (right most) Traffic lane or Auxiliary Lanes: Use time estimated by applying the existing methodology on the data from right most lane (lane 3).
- Shoulder: No obvious pattern could be observed in the upstream loop data hence it will not be appropriate to modify the reported time.
- Median: No obvious pattern could be observed in the upstream loop data hence it will not be appropriate to modify the reported time.

The logic behind the formulation of the aforementioned rules may be understood through careful inspection of Figure 3-3. It also helps to visualize the trends observed in the speed patterns from the station upstream of three different crash locations. Note that these are the typical speed profiles and most of the other crashes on these roadway locations also depicted similar trend. Crash on center lane (Figure 3-3(a)) represents crashes on mainstream freeway (lanes equipped with loop detectors), while Figure 3-3(b) depicts the speed pattern for crashes on the 4th lane (auxiliary lane) on the freeway. Therefore, as could be seen, the impact of a crash occurring on this lane could be captured by observing the drop of speed on the adjacent lane (rightmost lane equipped with loop detectors). To represent crashes on the shoulder and median, a shoulder crash has been chosen, which of course shows no visible drop pattern in speed on any of the lanes equipped with loop detectors (Figure 3-3(c)). The time series shown in Figure 3-3 has readings obtained from three freeway lanes for a period of 90 minutes (an hour prior and half an hour later to the reported time of each crash). Out of these 180 readings, the 120th is the reported time of the crash.

After applying this methodology for all the crashes having the desirable lane data available, the time of crash was modified accordingly. Due to the unavailability of the specific lane data for the required loop stations and time period for all the crash cases, the time of the crash was modified for 556 crashes from the years 1999 - 2002.

3.4.5 Discussion

Although the poor availability of the loop data did not allow us to modify the time of all historical crashes, the shock-wave and rule based algorithm could be a valuable addition to the

literature. In this study, however, the reported time was used for the remaining crashes, since removing them from the database would have resulted in significant loss of information. This is justified due to an automated system in place in Florida capturing the exact time when a crash is reported. With the wide spread use of the mobile phones the difference between times of occurrence and reporting of a crash is usually minimal. This along with the feedback from the Florida Highway Patrol officials about accurate reporting of the time of the crash gives reason to believe that the reported time is in fact close to actual time of crash occurrence.

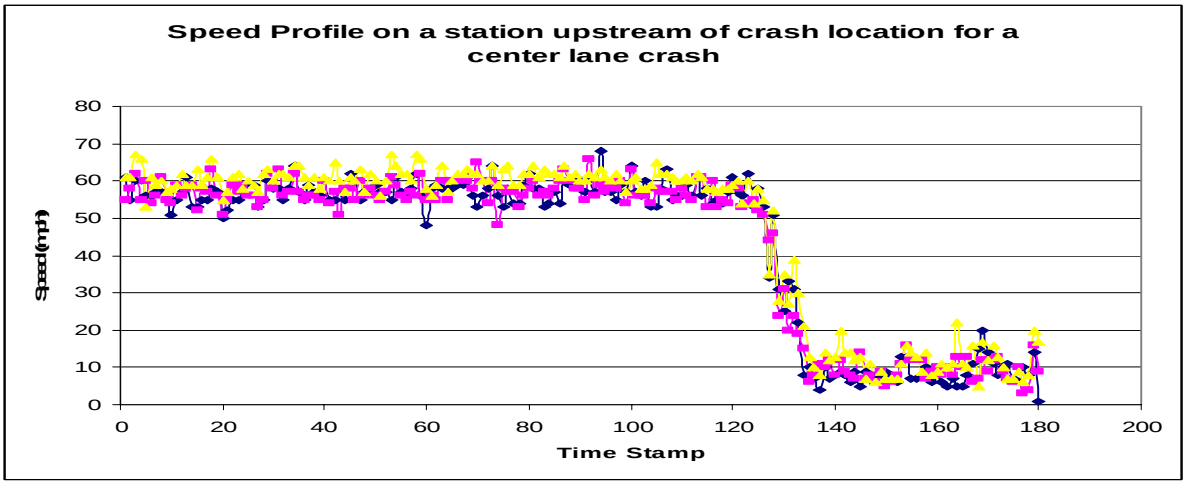

Figure 3-3(a): Typical Speed Profile: Crash on Center Lane

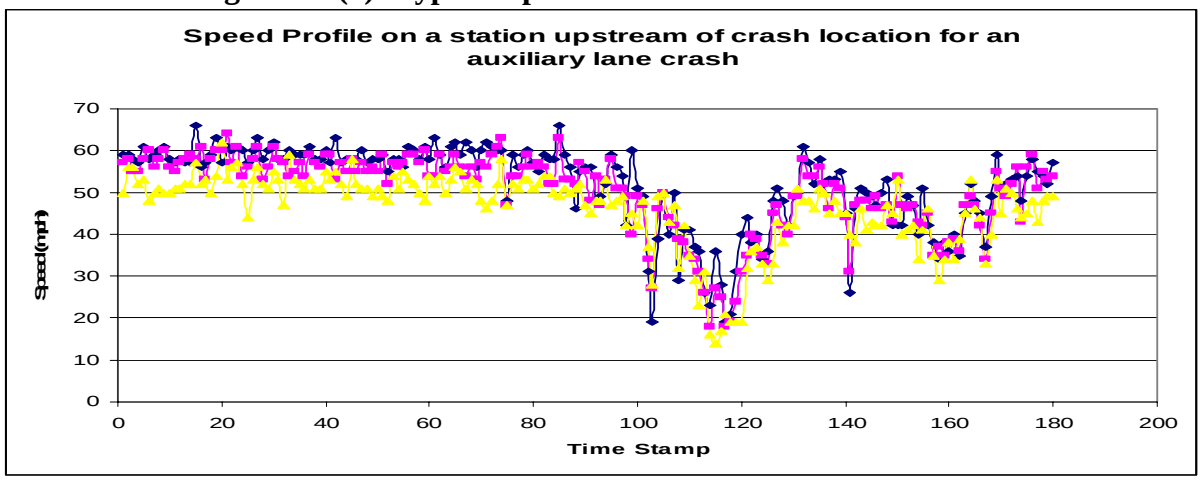

Figure 3-3(b) Typical Speed Profile: Crash on an Auxiliary Lane

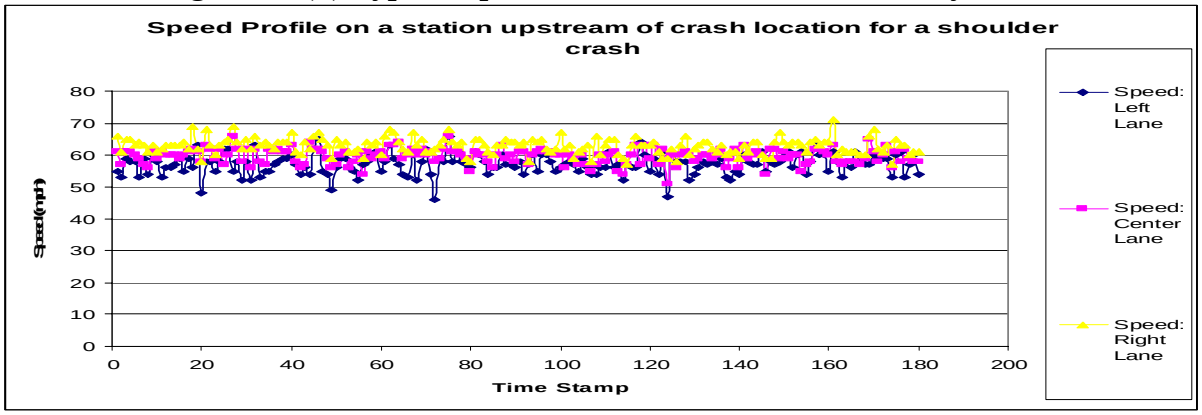

Figure 3-3(c) Typical Speed Profile: Crash on Shoulder

3.5 Loop Data Collection

The most critical part of this study is of course the loop detector data corresponding to crashes. As mentioned in the previous section for the four-year period 1705 crashes had no loop detector data available at all. Hence, the loop data was to be collected for the remaining 2050 crashes. The format of the data collected for analysis largely depends upon the methodology used. Past experience of the research group (e.g., Pande, 2003, Abdel-Aty et al. 2003, Abdel-Aty and Abdalla, 2003) with data from 7-month period of the year 1999 was very beneficial in this regard. Three separate databases consisting of loop detector data have been assembled for this study.

3.5.1 Data for Matched-Case Control Analysis

The matched case-control methodology was identified as an effective tool for modeling the binary outcome: crash or non-crash. To compare traffic characteristics (measured during time prior to crash occurrence from locations surrounding the crash location) that lead to a crash with corresponding normal traffic conditions that did not lead to a crash, traffic data were extracted in a specific matched format.

Loop 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, 1999 (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 the Mondays of the year at the same time. This matched sample design controls for all the factors affecting crash occurrence

such as season, day of week, location on the freeway, etc (thus implicitly accounting for all these factors). Hence, this case 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 year 1999, with one of them being the day of crash (crash case). More details of this sampling technique and application of this methodology may be found in one of the papers by our research group (Abdel-Aty et al., 2004). The format of data tables for this hypothetical crash is shown in Table 3-2.

Table 3-2 Format of the matched data extracted from the I-4 loop detector database for a hypothetical crash case

Day	Station	Y	Time	ELS*	ECS*	ERS*	ELV+	ECV+	ERV+	ELO-	ECO-	ERO-
04/05/99	25	0	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	25	0	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	|	0	|	|	|	|	|	|	|	|	|	|	|
04/05/99	|	0	|	|	|	|	|	|	|	|	|	|	|
04/05/99	31	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	31	0	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	25	1	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	25	1	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	|	1	|	|	|	|	|	|	|	|	|	|	|
04/12/99	|	1	|	|	|	|	|	|	|	|	|	|	|
04/12/99	31	1	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	31	1	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	25	0	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	25	0	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	|	0	|	|	|	|	|	|	|	|	|	|	|
04/19/99	|	0	|	|	|	|	|	|	|	|	|	|	|
04/19/99	31	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	31	0	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
|	|	0	|	|	|	|	|	|	|	|	|	|	|
|	|	0	|	|	|	|	|	|	|	|	|	|
12/27/99	31	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
12/27/99	31	0	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx

ELS* Eastbound Left lane Speed ELV+ Eastbound Left lane Volume ELO- Eastbound Left lane Occupancy
ECS* Eastbound Center lane Speed ELV+ Eastbound Center lane Volume ELO- Eastbound Center lane Occupancy
ELS* Eastbound Right lane Speed ELV+ Eastbound Right lane Volume ELO- Eastbound Right lane Occupancy

The filed *Y* in the table above represents whether the data row corresponds to a crash case or to a matched non-crash case. Such tables were extracted for all 2050 crashes with some loop data available. Note that the number of observations in these tables for different crashes was different due to random failures of the loops. Also, the cleaning mechanism explained above for raw 30-second loop data was again adopted to clean the data.

3.6 Geometric Design Parameters

Although the main purpose of this study is to establish links between real-time traffic characteristics (measured through loop detectors) and crash occurrences, it is extremely important to consider geometric characteristics on the freeway with respect to the crash characteristics. For example, the traffic characteristics leading to a crash on a curved section might be distinct from those leading to crash on a straight section. To obtain the details of the geometric design of I-4 corridor the Roadway Characteristics Editor (*RCI*) database available on *FDOT* Intranet server was used. Geometric design features were extracted for the location of each loop detector station since it was the common link between crash and loop detector database. The structure of this database is shown in Table 3-3. Geometric design of the freeway might differ from one direction to the other, hence the dataset has 138 (*69*2=138*) observations.

Table 3.3 Geometric design of the freeway at loop detector station locations

Loop	Direction	Mile post	Radius (ft)	# of Lanes	Median type and width	Distance to nearest upstream on ramp	Distance to nearest upstream off ramp	Distance to nearest down stream on ramp	Distance to nearest down stream off ramp
2	E	xxx	xxx	xxx xxx	xx xx	xxx	xxx	xxx	xxx
2	W	xxx	xxx	xxx xxx	xx xx	xxx	xxx	xxx	xxx
|		|	|	|	| |	|	|	|	|
71	E	xxx	xxx	xxx xxx	xx xx	xxx	xxx	xxx	xxx
71	W	xxx	xxx	xxx xxx	xx xx	xxx	xxx	xxx	xxx

3.7 Weather Information

The effect of wet weather on crash occurrence is well documented (e.g., Xiao et al. 2000). In Central Florida where snow is not a concern, rain fall is the most important weather related factor affecting visibility as well as the pavement condition. These two parameters are available for historical crashes; however, for the non-crash cases there is no direct way to obtain the weather information at locations from where loop data has been collected. An effort is currently underway by our research group to infer the weather conditions for the non-crash cases using the rainfall information provided by five different rain gauge stations located in the surroundings of the 36-mile corridor. This issue will be further investigated in the second phase of the project.

3.8 Driver Characteristics

While crash involvement of drivers belonging to certain age group or gender etc. has been a major area of research in traffic safety, these factors have not been incorporated into real-time crash prediction models developed so far. This issue will be further investigated in the second phase of the project.

3.9 Concluding Remarks

This Chapter describes the data preparation effort needed for this study and beyond. The data have been prepared keeping in mind the future scope of this project. Significant amount of time and effort is currently being devoted to collection and assembling of database, so that the data issues do not limit the scope of the research. A random non-crash cases, driver characteristics etc. are planned to be used in the next phase of this study. In this phase, however, the matched case-control data base was used. The with-in stratum analysis technique is particularly attractive for modeling purposes since it implicitly accounts for the factors such as freeway geometry. However, in order to understand the mechanism of crashes these factors should be explicitly accounted for in the model. Therefore, the efforts to incorporate weather and driver population composition related factors are currently underway and will be used during modeling stage in the next phase of the project. Determination of the time of historical crashes has been given separate attention and a detailed rule-based algorithm is used to modify the reported time. With 4-years (possibly five years) of crash and non-crash data, the database being developed here would be by far the most comprehensive database created for a real-time crash prediction study. The next chapter explores the with-in stratum logistic regression methodology for crash prediction using the matched dataset of the format shown in Table 3-2.

CHAPTER 4

LOGISTIC REGRESSION: SIMPLE AND MULTIVARIATE MODELS

4.1 General

Freeway crashes occur as a result of complex interaction between human factors, ambient traffic and environmental conditions, along with the geometric characteristics of the freeway section. This study aims at the identification of traffic characteristics leading to crashes on freeways. Traffic conditions measured as coefficient of variation in speed and lane occupancy have been found to be significant freeway crash precursors (e.g. Lee et al. 2002, 2003). These authors have developed crash prediction models using real-time values of the precursors obtained from underground freeway loop detectors located upstream and/or downstream of crash sites.

However, these models do not take into consideration geometric and environmental factors such as horizontal curve and season of the year. Furthermore, crash precursors are measured from loop detectors in the neighborhood of the crash location at time duration prior to crashes only. 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 (therefore implicitly accounting for factors such as the geometry and location). 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 (henceforth referred to as stratum) can be formed using crash site, time, season, day of the weak, etc., so that the variability due to these factors is controlled. In epidemiological studies, this is known as matched case-control study. Each case refers here to a crash and control to a non-crash. The steps involved can be described as follows:

(i) Select a crash site, and identify loop detector(s) upstream and/or downstream of crash location. Measure traffic flow characteristics from these loop detectors at a time period duration prior to the crash. Use the same loop detectors and time period, measure traffic flow characteristics over *m* other non-crash days (same day of the weak and season). The *m+1* observations (one corresponds to crash and *m* to non-crashes) together form one stratum.

(ii) Repeat step (i) for *N* randomly selected crash locations to form *N* strata.

(iii) Perform within-stratum analysis to identify traffic flow variables that are associated with the binary outcome (crash/no-crash) variable *Y* while controlling variability due to all other external factors that formed the strata.

4.2 Matched Case-Control Logistic Regression: Simple Models

4.2.1 Methodology

The case-control stratum analysis methodology is adopted to identify the relationship between the traffic parameters measured through loop detectors and crash occurrences while controlling for location (i.e., the geometric characteristics), time of the day, day of the week and season (Abdel-Aty et al., 2004).

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

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

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. Under the assumption that the logit is linear in the continuous covariate *x*, the equation for the logit would be $g(x) = \beta_0 + \beta_1 x$. It follows that the slope coefficient, β_1, gives 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*. Hazard ratio is defined as the exponential of this coefficient, in other words it represents how much more likely (or unlikely) it is for the outcome to be present for an increase of "1" unit in *x* (Agresti, 2002). It implies that the hazard ratio significantly different from 1 for a particular parameter is an indicator of strong association of that parameter with crash occurrence. The decision regarding significance is made based on the *p-value*, which represents the probability of drawing the sample being tested if the null hypothesis were actually true. The null hypothesis is formulated as hazard ratio being equal to unity. Therefore, a *p-value* of less than the threshold (selected as 0.05) would indicate the rejection of the null hypothesis and hazard ratio significantly different than unity. It is also noteworthy that a value greater than one signifies that the crash risk increases with an increase in the parameter value while a value less than one indicates an increase in the crash risk as the parameter value goes down.

4.2.2 Data Preparation

Procedure for data preparation was explained in the previous chapter (Section 3.5.1), however, part of it is being repeated here to maintain continuity. As explained in the previous chapter, for matched case-control logistic regression 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, 1999 (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 the Mondays of the same season in that year at the same time. This matched sample design controls for all the factors affecting crash occurrence such the location on the freeway (thus accounting for the geometric factors). Hence, this case 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 year 1999, with one of them being the day of crash (crash case). Details of this sampling technique and application of this methodology may also be found in one of the papers by Abdel-Aty et al. (2004).

Since the 30-second data have random noise and is difficult to work with in a modeling framework, we combined the 30-second data into two separate levels of 3-minute and 5-minute level in order to get averages and standard deviations. Thus for 5-minute 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 should be noted that "*F*" is the station closest to the location of the crash with "*G*" and "*H*" being the stations downstream of the crash location. Similarly the 5-minute intervals were also 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. Similarly for the 3-minute level, the interval between the time of the crash and 3 minutes prior to the crash was named as slice 1, interval between 3 to 6 minutes prior to the crash as slice 2, and interval between 6 to 9 minutes prior to the crash as slice 3 and so on. For 5-minute level

aggregation the arrangement of these time-slices and stations is shown in Figure 4-1. Two effects, namely average and standard deviation were initially calculated for speed, volume and occupancy during each time slice and from each lane at every station. The original data series being at 30-second level, the 3-minute and 5-minute averages (and standard deviations) were based on six and ten observations, respectively. Using information about the specific lane where the crash occurred from the FDOT database, average and standard deviation for only lane of the crash were retained.

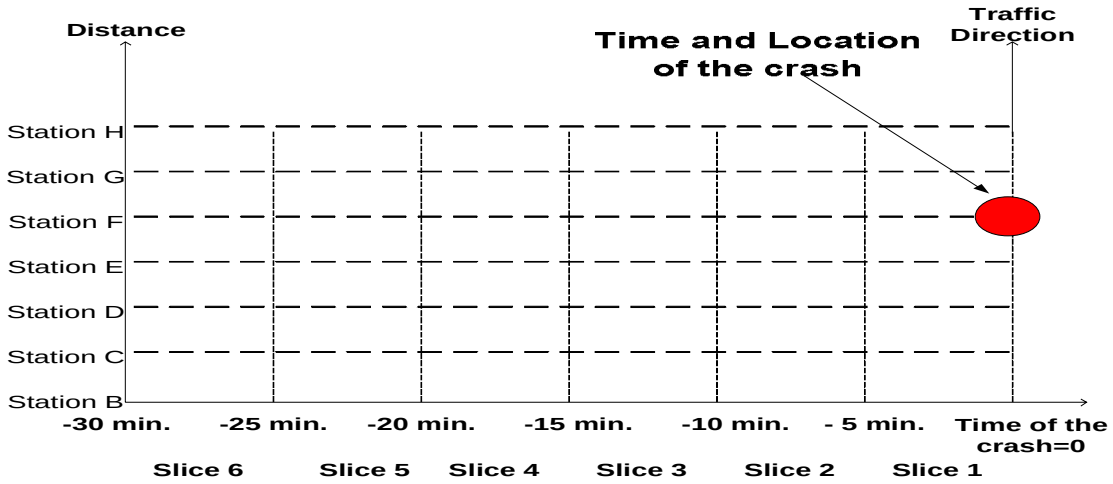

Figure 4-1 Time-space arrangement of all stations and time slices with respect to the crash site and the time of the crash

Using data only from the specific lane of the crash reduced 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. Two more datasets were created, by aggregating the data on the three lanes; hence in the aforementioned three-minute and five-minute datasets the lane of the crash averages and standard deviations were replaced by values aggregated over three lanes. In these datasets, the averages (and standard deviations) at 3-minute and 5-minute level were based on 18 (6*3 lanes) and 30 (10*3 lanes) observations, respectively. Therefore, even if at a certain station loop

detector from one lane was not reporting data there were observations available to get a measure of traffic from that location. This not only increases the sample size of crashes to more than 2000 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.

4.2.3 Analysis

For each of the seven loop detectors (*B* to *H*) and six time slices (1-6) mentioned above, there are values of means (*AS, AV, AO*) and standard deviations (SS, SV, SO) of speed, volume and occupancy, respectively, of all crash and the corresponding non-crash cases. Due to data availability, there were different numbers of non-crash cases for each crash. To carry out matched case-control analysis we created a symmetric data sets (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 all four datasets. The choice of selecting five as the number of corresponding non-crash cases was based on one of our earlier findings (Abdel-Aty et al., 2004) which essentially indicated no differences among the results from five different *1: m* datasets (with *1* crash and *m* corresponding non-crash with *m* varying between one to five). In addition to the aforementioned datasets 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 all the real crash cases were dropped. The results from this dataset were analyzed in order to delineate the differences between real and "pseudo" case control datasets. Exploratory analysis with the original effects (3-minute or 5-minute standard deviations and average of speed) 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 concluded 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* of *SAS*) of each of the *126 (7 stations *6 time slices *3parameters 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 variables 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 estimate of the expected change in the risk 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). Note that the terms such as hazard ratio and the *p-value* were defined in the methodology section (Section 4.2.1, Page 33).

These 126 single variable models were estimated for corresponding hazard ratio values for all five datasets including the four real (3-minute and 5-minute aggregation with individual lane of the crash/combined lanes) and one "pseudo" matched case-control dataset (combined lane at 3-minute aggregation having one non-crash in each strata randomly assigned as crash). The arrangement used for stations and time slices used here is crucial in terms of generating the patterns of crash risk and it's "propagation" in a time-space framework. The results from these datasets are discussed in the following section.

4.2.4 Results and Discussion

First dataset to be analyzed for hazard ratio was the one aggregated to 3-minute level with parameters only from lane of the crash. Table 4-1 shows the results of all the single variable 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., near to 1.0) but it is still 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 risk ratio can be obtained by raising hazard ratio to the power 10. For *SV* the hazard ratios were found to be less than one and appeared to be decreasing as the time and station of crash approached in the downstream direction. Since it is the value of hazard ratio that is significantly different from one (and not necessarily a high 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. This analysis was based on a small

sample size due to missing data from individual lane on which the crash occurred and also the determination of these risk ratio values would require the data from each individual lane to be available, therefore we next conducted our analysis on 3-minute level data combined over three lanes. In the combined lane data, the same trends in hazard ratio are essentially observed in a time-space framework.

Table 4-1: Hazard ratios for LogCVS, SV and AO for individual lane parameters measured at 3-minute level during six different time slices and seven stations

		LogCVS			*SV*			*AO*		
Station	**Time slice**	**Hazard Ratio**	**chi-sq.**	**p-value**	**Hazard Ratio**	**chi-sq.**	**p-value**	**Hazard Ratio**	**chi-sq.**	**p-value**
B	1	1.716	19.1895	<.0001	0.956	1.6018	0.2057	1.024	17.969	<.0001
B	2	1.512	10.8127	0.001	0.996	0.0117	0.9139	1.02	14.1926	0.0002
B	3	1.924	26.1665	<.0001	0.963	1.2512	0.2633	1.02	13.8534	0.0002
B	4	2.155	36.0467	<.0001	0.957	1.514	0.2185	1.022	16.6653	<.0001
B	5	1.786	22.5242	<.0001	0.932	4.1029	0.0428	1.023	17.2121	<.0001
B	6	1.749	18.8261	<.0001	0.92	5.9009	0.0151	1.019	12.3945	0.0004
C	1	2.41	50.4977	<.0001	0.956	1.6973	0.1926	1.021	17.5138	<.0001
C	2	2.235	43.462	<.0001	0.962	1.3057	0.2532	1.021	18.4326	<.0001
C	3	2.103	34.5474	<.0001	0.963	1.1919	0.2749	1.021	18.9807	<.0001
C	4	2.149	38.0753	<.0001	0.958	1.6183	0.2033	1.023	23.1371	<.0001
C	5	1.859	25.388	<.0001	0.971	0.7426	0.3888	1.022	18.2844	<.0001
C	6	2.694	64.2967	<.0001	0.985	0.1984	0.656	1.027	29.0918	<.0001
D	1	2.489	53.2179	<.0001	0.928	4.5202	0.0335	1.037	48.8821	<.0001
D	2	2.179	37.8872	<.0001	0.935	3.7976	0.0513	1.034	41.6895	<.0001
D	3	2.333	43.5329	<.0001	0.958	1.5171	0.2181	1.03	31.9471	<.0001
D	4	2.199	36.5207	<.0001	0.984	0.2075	0.6487	1.032	34.4802	<.0001
D	5	1.802	21.2986	<.0001	0.926	5.1727	0.0229	1.027	25.7885	<.0001
D	6	2.318	42.453	<.0001	0.95	2.1475	0.1428	1.034	39.3768	<.0001
E	1	2.684	70.8647	<.0001	0.95	2.1628	0.1414	1.035	53.2115	<.0001
E	2	2.633	62.6352	<.0001	0.89	11.2021	0.0008	1.04	61.6082	<.0001
E	3	2.627	60.7399	<.0001	0.936	3.7193	0.0538	1.044	68.0471	<.0001
E	4	2.633	62.146	<.0001	0.898	9.6234	0.0019	1.038	56.3305	<.0001
E	5	2.141	39.3845	<.0001	0.867	15.7183	<.0001	1.036	49.9789	<.0001
E	6	1.994	31.6068	<.0001	0.88	12.9029	0.0003	1.036	52.5798	<.0001
F	1	4.532	177.723	<.0001	0.816	36.0686	<.0001	1.039	66.4986	<.0001
F	2	3.145	106.751	<.0001	0.874	16.3643	<.0001	1.041	67.4152	<.0001
F	3	3.989	137.278	<.0001	0.867	17.8128	<.0001	1.042	67.5865	<.0001
F	4	3.454	118.519	<.0001	0.831	27.6133	<.0001	1.048	86.1573	<.0001
F	5	3.293	104.619	<.0001	0.882	13.7532	0.0002	1.042	71.0116	<.0001
F	6	2.819	82.7555	<.0001	0.863	19.648	<.0001	1.042	68.8663	<.0001
G	1	3.505	114.199	<.0001	0.819	30.8908	<.0001	1.036	52.8328	<.0001
G	2	3.035	88.463	<.0001	0.81	34.9073	<.0001	1.038	59.2016	<.0001
G	3	2.972	81.9169	<.0001	0.842	23.3896	<.0001	1.036	51.051	<.0001
G	4	2.793	74.5892	<.0001	0.885	12.8455	0.0003	1.038	59.6686	<.0001
G	5	2.572	65.2376	<.0001	0.848	21.3729	<.0001	1.039	58.2009	<.0001
G	6	2.378	56.5465	<.0001	0.853	20.6301	<.0001	1.041	65.3361	<.0001
H	1	3.245	88.9293	<.0001	0.875	12.5506	0.0004	1.054	76.2902	<.0001
H	2	2.719	64.2434	<.0001	0.826	24.9019	<.0001	1.047	64.7683	<.0001
H	3	2.334	46.6994	<.0001	0.839	22.2001	<.0001	1.05	67.4602	<.0001
H	4	2.677	59.1609	<.0001	0.83	24.2464	<.0001	1.051	67.8784	<.0001
H	5	2.475	51.3797	<.0001	0.884	11.215	0.0008	1.053	75.5672	<.0001
H	6	2.897	66.1671	<.0001	0.842	21.6698	<.0001	1.056	77.1597	<.0001

To assess the fact that these results are really depicting an association between traffic flow variables and crash occurrence we next analyzed hazard ratios from the "pseudo" crash matched dataset. As expected the trends were either non-existent (as was the case with *LogCVS* and *SV* with values significantly close to one) or they were exactly reversed (as was the case with *AO* with hazard ratio significantly less than one). Table 4-2 contrasts the differences between 3-minute combined lanes matched and "pseudo" matched dataset results and further indicates strong association of these variables with crash occurrences.

With the five minute aggregated datasets again similar trends were observed for hazard ratios corresponding to *SV* and *AO* while in the case of *LogCVS,* the hazard ratio and corresponding chi-square statistic were magnified depicting stronger association of 5-minute coefficients of variation in speed with crash occurrence. In data aggregated to 5-minute level hazard ratios for parameters *LogCVS* and *SV* corresponding to combined lane data were higher and lower, respectively, than their individual-lane counter parts. Table 4-3 shows results from these two datasets. The essential difference between the two datasets is that while the combined lane dataset accounts for the variation across the lanes wherever possible, the individual lane of the crash dataset does not. The magnified difference between unity and both hazard ratios (corresponding to *LogCVS* and *SV*) in the combined lane data indicates that similar volumes with varying speeds across lanes might be a contributing factor for freeway crashes. Also, note that the sample size in case of combined lanes is about four times larger than in the case of individual lane. Hence it was decided to use the combined lane data for hazard ratio calculation as well.

Table 4-2: Comparison between hazard ratios for two separate with-in stratum analyses for combined lane parameters measured at 3-minute level, one with strata of 6 with one real crash and other with strata of 6 with one non-crash randomly assigned as crash

Station	Time slice	*Hazard Ratios Corresponding to*					
		LogCVS		***SV***		***AO***	
		Strata with Real crash	***Strata with Pseudo crash***	***Strata with Real crash***	***Strata with Pseudo crash***	***Strata with Real crash***	***Strata with Pseudo crash***
B	1	1.961	0.743	0.918	1.027	1.027	0.961
B	2	1.875	0.861	0.924	1.001	1.025	0.96
B	3	1.826	0.744	0.929	1.069	1.025	0.959
B	4	2.287	0.771	0.944	1.066	1.029	0.965
B	5	2.04	0.755	0.913	1.074	1.027	0.961
B	6	1.766	0.682	0.939	1.022	1.025	0.959
C	1	2.66	0.827	0.94	1.045	1.028	0.974
C	2	2.164	0.845	0.95	1.024	1.028	0.979
C	3	2.171	0.814	0.907	1.034	1.028	0.976
C	4	2.208	0.877	0.969	1.058	1.027	0.977
C	5	1.732	0.723	0.924	1.041	1.02	0.975
C	6	2.16	0.715	0.956	1.003	1.024	0.973
D	1	2.605	0.763	0.887	0.965	1.037	0.966
D	2	2.365	0.784	0.886	0.94	1.036	0.963
D	3	2.387	0.796	0.933	0.944	1.032	0.962
D	4	2.276	0.71	0.879	0.959	1.032	0.961
D	5	1.984	0.655	0.894	0.949	1.031	0.963
D	6	2.195	0.717	0.877	0.949	1.035	0.961
E	1	3.285	0.857	0.879	0.975	1.044	0.969
E	2	2.795	0.977	0.88	0.977	1.043	0.97
E	3	2.461	1.048	0.865	0.991	1.043	0.97
E	4	2.612	0.954	0.856	0.939	1.039	0.97
E	5	2.073	0.829	0.846	0.937	1.038	0.961
E	6	2.31	0.896	0.862	0.9	1.033	0.973
F	1	4.529	0.761	0.822	0.995	1.046	0.959
F	2	3.739	0.86	0.844	0.988	1.045	0.963
F	3	3.755	0.695	0.842	0.996	1.049	0.957
F	4	4.037	0.845	0.856	1.022	1.053	0.96
F	5	3.361	0.715	0.867	0.992	1.048	0.964
F	6	2.764	0.81	0.865	1.001	1.046	0.962
G	1	3.344	0.795	0.806	1.052	1.041	0.961
G	2	3.18	0.815	0.851	1.06	1.042	0.964
G	3	2.647	0.794	0.791	1.02	1.037	0.967
G	4	2.503	0.872	0.853	1.059	1.043	0.967
G	5	2.326	0.785	0.825	1.004	1.043	0.966
G	6	2.507	0.892	0.862	0.976	1.045	0.963
H	1	2.675	0.815	0.846	1.03	1.044	0.959
H	2	2.721	0.663	0.809	1.012	1.046	0.958
H	3	2.622	0.711	0.825	1.026	1.044	0.958
H	4	2.428	0.782	0.823	1.042	1.047	0.959
H	5	2.598	0.761	0.873	1.022	1.046	0.96
H	6	2.545	0.75	0.84	1.015	1.048	0.959

Table 4-3: Comparison between hazard ratios for two separate with-in stratum analyses, one with combined lane parameters, and other with parameters measured from lane of the crash both at 5-minute level

Station	Time slice	*Hazard Ratios Corresponding to*					
		LogCVS		***SV***		***AO***	
		Parameters from all lanes combined	***Parameters from Lane of the crash***	***Parameters from all lanes combined***	***Parameters from Lane of the crash***	***Parameters from all lanes combined***	***Parameters from Lane of the crash***
B	1	1.776	1.947	0.902	0.961	1.024	1.031
B	2	1.805	1.786	0.944	1.017	1.026	1.029
B	3	2.347	2.519	0.937	1.073	1.027	1.03
B	4	1.645	1.987	0.907	0.991	1.022	1.023
B	5	1.688	2.008	0.849	1.044	1.025	1.033
B	6	1.803	2.366	0.907	1.051	1.021	1.029
C	1	2.301	2.106	0.933	0.921	1.029	1.017
C	2	2.107	2.218	0.905	1.01	1.03	1.033
C	3	1.973	2.185	0.927	1.01	1.026	1.026
C	4	2.369	2.696	0.902	1.034	1.026	1.028
C	5	2.104	2.389	0.923	0.97	1.029	1.025
C	6	1.889	2.14	0.899	1.004	1.025	1.027
D	1	2.92	2.14	0.849	0.989	1.035	1.037
D	2	2.525	1.878	0.944	0.99	1.032	1.03
D	3	1.973	1.965	0.868	1.039	1.03	1.038
D	4	2.494	2.376	0.88	0.998	1.033	1.047
D	5	2.301	1.888	0.856	0.952	1.032	1.036
D	6	1.972	2.096	0.857	0.982	1.03	1.044
E	1	4.096	3.419	0.898	0.936	1.049	1.042
E	2	2.909	2.81	0.838	0.851	1.045	1.04
E	3	2.533	2.624	0.862	1.017	1.04	1.047
E	4	2.653	3.004	0.88	0.925	1.038	1.046
E	5	2.273	2.405	0.862	0.929	1.038	1.044
E	6	2.017	2.691	0.836	0.93	1.037	1.051
F	1	6.216	4.217	0.808	0.849	1.047	1.041
F	2	4.818	3.94	0.823	0.863	1.046	1.036
F	3	4.033	4.135	0.871	0.858	1.047	1.036
F	4	3.185	4.137	0.839	0.897	1.042	1.038
F	5	3.322	3.699	0.9	0.894	1.043	1.048
F	6	3.361	3.565	0.86	0.842	1.041	1.032
G	1	4.3	4.38	0.791	0.845	1.048	1.046
G	2	3.735	3.551	0.779	0.791	1.043	1.048
G	3	2.783	3.727	0.826	0.859	1.042	1.047
G	4	2.964	3.646	0.818	0.82	1.044	1.051
G	5	2.736	3.507	0.816	0.878	1.039	1.041
G	6	2.434	3.431	0.799	0.846	1.038	1.043
H	1	3.122	3.628	0.803	0.879	1.046	1.061
H	2	3.244	2.46	0.805	0.833	1.047	1.046
H	3	2.791	2.288	0.831	0.854	1.049	1.057
H	4	2.642	3.396	0.807	0.949	1.046	1.058
H	5	2.152	2.828	0.826	0.854	1.045	1.063
H	6	2.437	2.862	0.868	0.937	1.043	1.049

In short, 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 1.5 miles (from Station C) upstream of the crash location, it is considerably clear at about ½ 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. The other factor, low value of *SV,* indicates 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 stable volume in terms of time 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.

Beside these overall trends the results outline the differences between coefficients of variation/average measured at varying length of time slices (three and five minutes) as well. The five minute time slice would be more effective in the crash prediction as it not only has higher and more significant hazard ratio for *LogCVS* but it also provides more allowance in terms of time to analyze data, estimate and possibly reduce the likelihood of crashes. From here on we will focus our attention on 5-minute aggregate data with all lanes combined together rather than individual lane and/or data aggregated to the 3-minute level.

4.2.5 Spatio-temporal Variation of Crash Risk

As argued earlier, the analysis from here on is based on the 5-minute averages, standard deviations and coefficient of variation. To depict the patterns in the hazard ratio we show the contour plots of the ratio for all three variables found significant in a time-space framework. But first the type of the crash information available with the FDOT crash database was utilized in order to "clean" the 5-minute combined lane dataset by only retaining multi-vehicle crashes. Since the traffic conditions are more likely to impact the crashes involving interaction among vehicles rather than the single vehicle crashes mostly occurring due to error on the drivers' part. Once this cleaned database was used for analysis it was found that the hazard ratio values were further boosted for *LogCVS* and *AO* while they further dropped in the case of *SV* as expected. 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 4-2(a), 4-3(a) and 4-4(a), respectively. Note that in Figure 4-2(a), and 4-3(a) the dark colored region represents high hazard ratios thereby identifying more risk while in Figure 4-4(a) 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. The contour plots for hazard ratios obtained from "pseudo" dataset give an idea about "normal" conditions on freeways (See Figures 4-2(b), 4-3(b) and 4-4(b)). These figures are in perfect contrast with their counterparts showing hazard ratio for a real matched case control dataset. It provides visual evidence for the contribution of traffic factors toward crash occurrence.

As we can see in all three plots (4-2(a), 4-3(a) and 4-4(a)) 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 be able to predict and avoid 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 (Figure 4-3(a)) corresponding to *SV* appears dark for locations downstream of the crash location which indicates that very stable flow 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*.

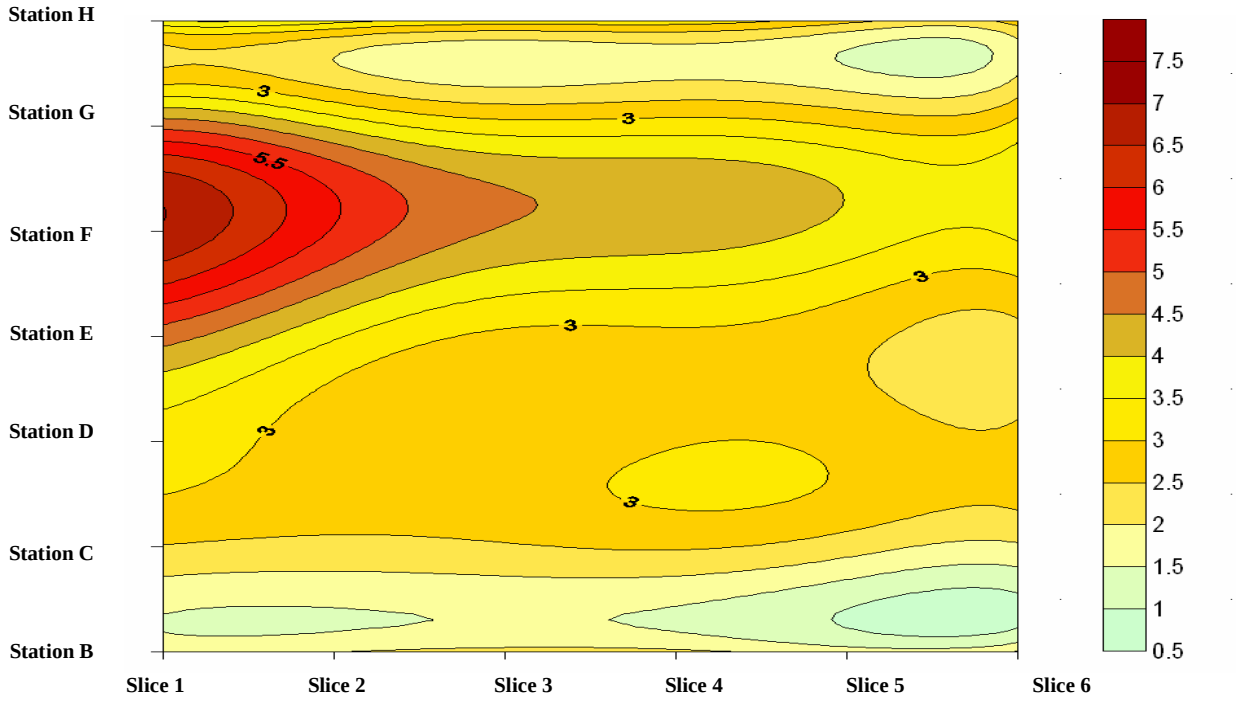

Figure 4-2(a): Spatio-temporal pattern of the hazard ratio for LogCVS obtained from 5-minute combined lane dataset for multi-vehicle crashes

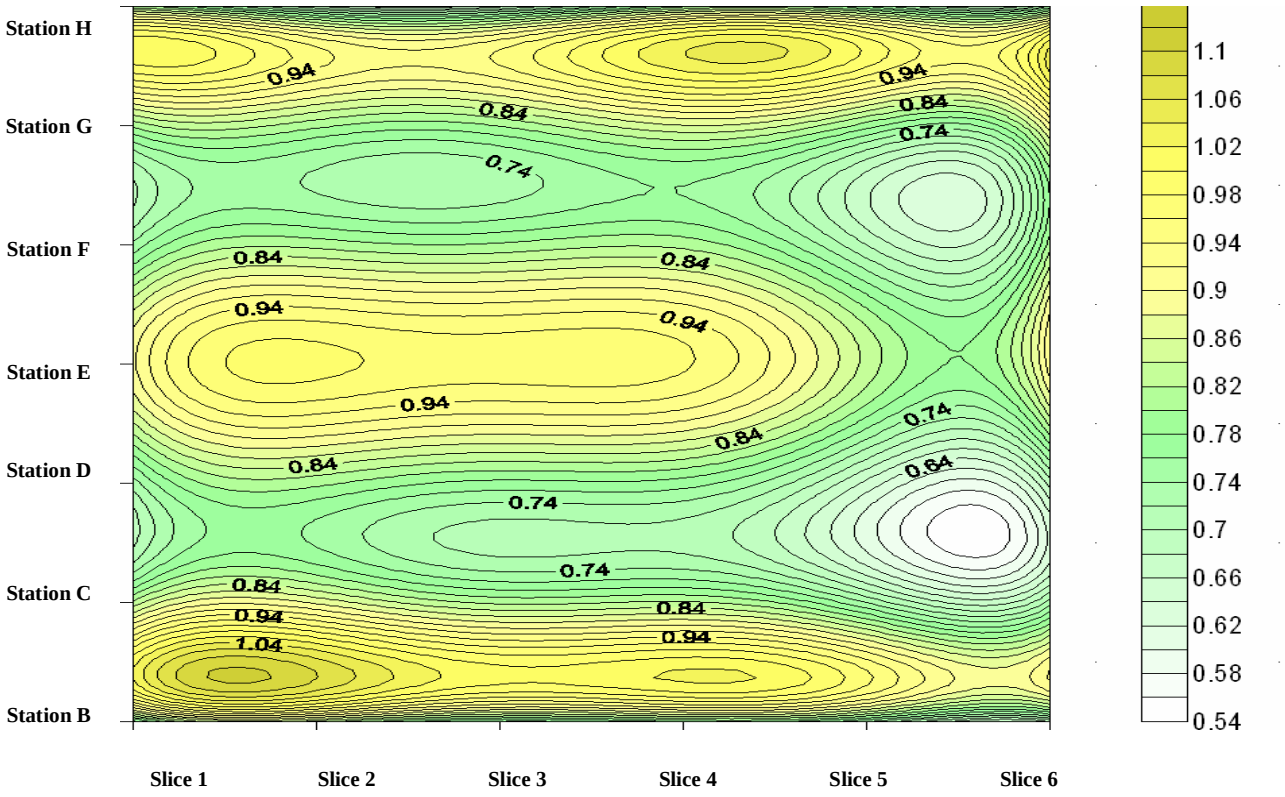

Figure 4-2(b): Spatio-temporal pattern of the hazard ratio for LogCVS obtained from "pseudo" crash case dataset

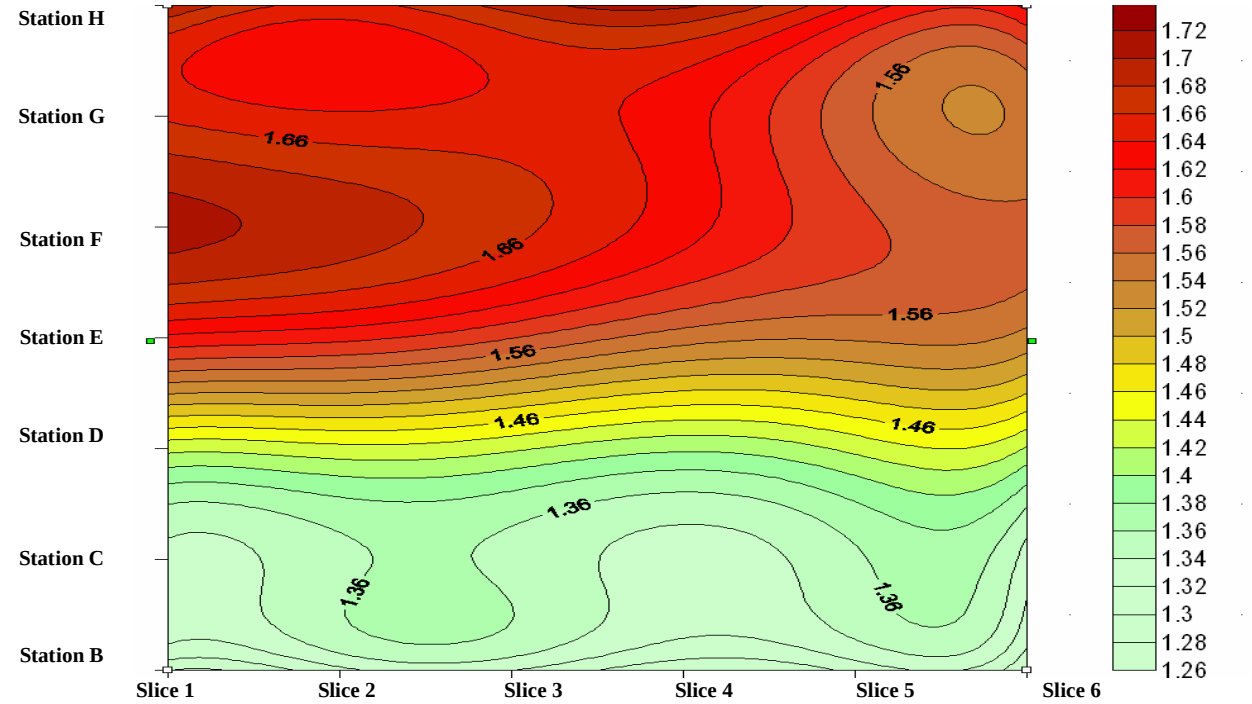

Figure 4-3(a): Spatio-temporal pattern of the modified hazard ratio (increase in crash risk when there is ten unit increase in occupancy rather than one) for AO obtained from 5-minute combined lane dataset for multi-vehicle crashes

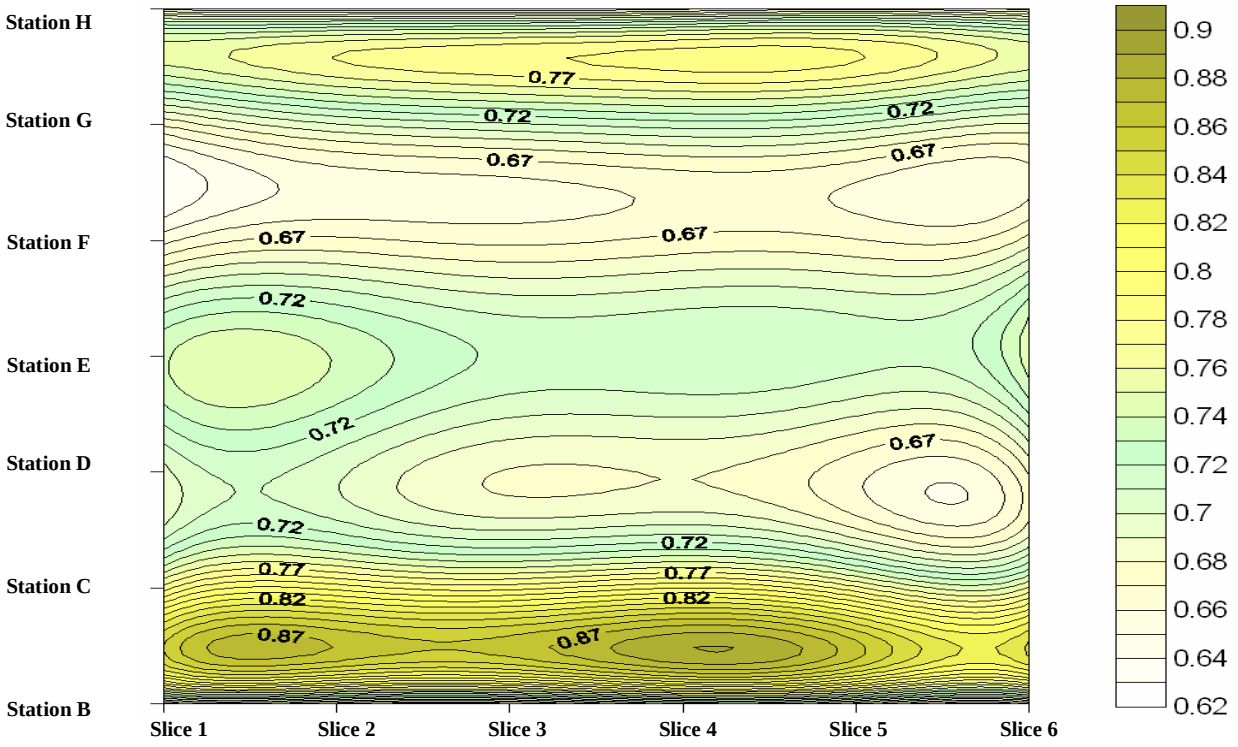

Figure 4-3(b): Spatio-temporal pattern of the modified hazard ratio (increase in crash risk when there is ten unit increase in occupancy rather than one) for AO obtained from "pseudo" crash dataset

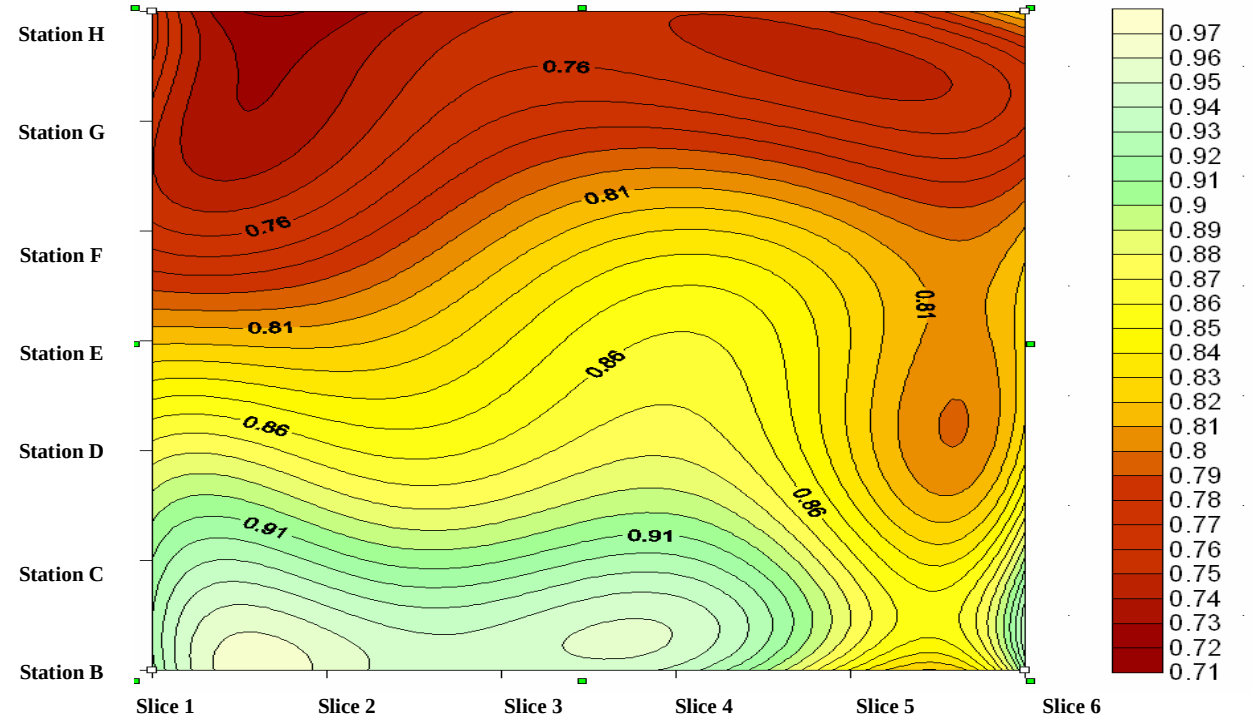

Figure 4-4(a) Spatio-temporal pattern of the hazard ratio for SV obtained from 5-minute combined lane dataset for multi-vehicle crashes

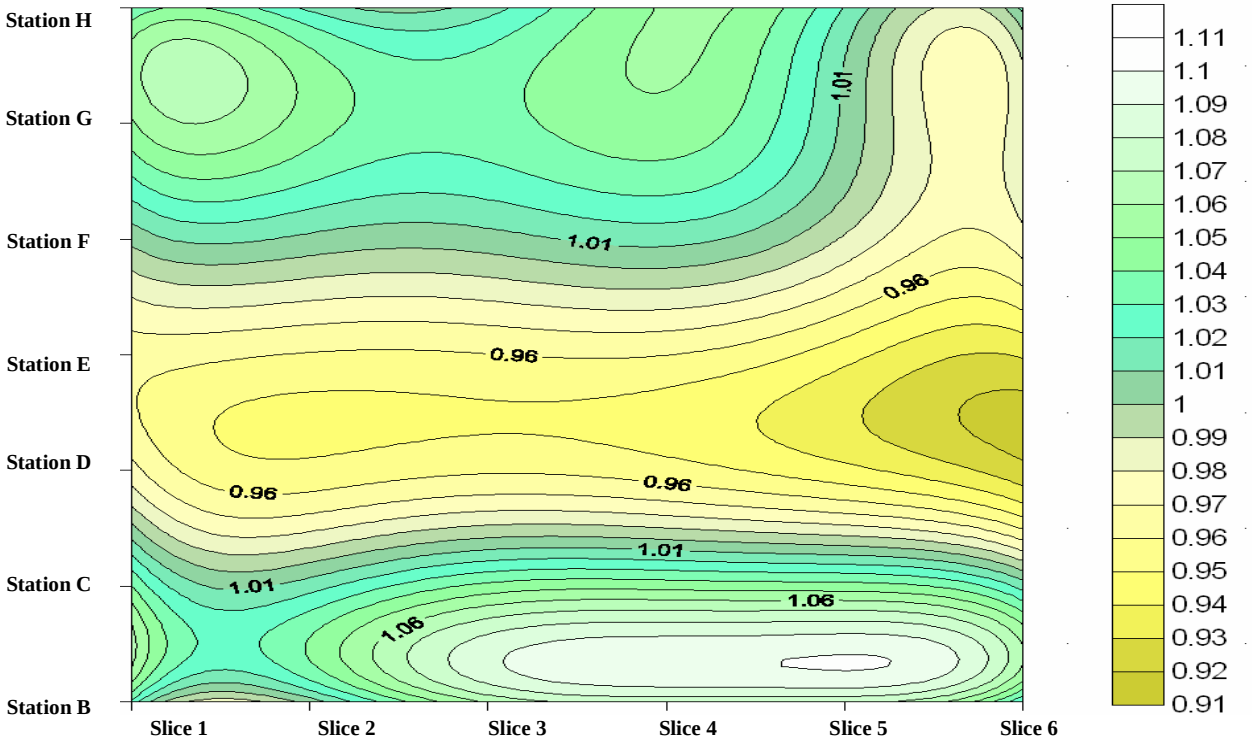

Figure 4-4(b): Spatio-temporal pattern of the hazard ratio for SV obtained from "pseudo" crash case dataset

4.2.6 Conclusions from the Simple Models

The matched case-control logistic regression was used as a simple analysis technique to detect traffic patterns that result in high potential of crashes on freeways. It was found that the coefficients of variation in speed measured at 5-minute intervals show slightly better association with crash occurrence than those measured at the 3-minute level. Also, combining observations from three lanes was concluded to be better than using only data from the lane where the crash occurred since it does not only captures across lane variation (or lack of it) in speed (or volume), but also allows us to use larger dataset for analysis. It also has an advantage in real-time application in case of a loop failure on a certain lane. The results show that even if the first time slice (0-5 minutes prior to a crash) is excluded due to practical considerations of the time required to act on the information and warn the drivers, 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. The hazard ratio values for these variables were significantly different from one around the crash location for three to four time slices (i.e., the precursors existed for about 15 minutes), that should provide enough time for prediction (and prevention) of crashes. Another significant feature of these findings is that they are based on accurately estimated time of the crash thereby evading the "cause and effect" fallacy. The results from the "Pseudo" matched case control dataset containing six non-crash cases with one of them randomly assigned as crash also prove the association between crash occurrence and the traffic variables identified here. Based on these findings we selected 5-minute combined lane dataset with only multi-vehicle crashes to develop our final model. The dataset had 1528 strata with each stratum consisting of one crash and five corresponding non-crash cases.

4.3 Matched Case-Control Logistic Regression: Multivariate Modeling

4.3.1 Methodology for Modeling and Classification

The purpose of the 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 study. In this section this extension of simple logistic regression to multivariate problem has been described 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 conditional likelihood for the j^{th} stratum is the probability of the observed data given the total number of observations and the number of crashes observed in the stratum. The probability of any observation in a stratum being a crash may be modeled using the following linear logistic regression model:

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

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

Note that the intercept term α_j summarizes the effect of control variables (used to form the strata) on the crash probability and would be different for different strata. In order to account for the stratification in the analysis, a conditional likelihood is constructed. The complex mathematical derivation of the relevant likelihood function is omitted here and the reader is referred to Collett

(1991) for more details. This conditional likelihood function is independent of the intercept terms $\alpha_1, \alpha_2, \ldots\ldots \alpha_N$ (Collett, 1991). So the effects of matching variables cannot be estimated and Equation 2 cannot be used to estimate crash probabilities. However, the values of the β parameters that maximize the conditional likelihood function would also be estimates of β coefficients in Equation 2. These estimates are log odds ratios and can be used to approximate the relative risk of a crash.

The log odds ratios can also be used for prediction purposes under this matched crash-non-crash analysis. Consider two observation vectors $x_{1j} = (x_{11j}, x_{21j}, \ldots, x_{k1j})$ and $x_{2j} = (x_{12j}, x_{22j}, \ldots, x_{k2j})$ from the j^{th} strata 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} may be derived from equation 2 and will have the following form

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

The right hand side of equation 3 depends only on β_j, therefore the estimate for log odds ratio may be obtained using estimated β coefficients. One may utilize the above relative log odds ratio for predicting crashes by replacing x_{2j} by 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 $\bar{x}_{2j} = (\bar{x}_{12j}, \bar{x}_{22j}, \bar{x}_{32j} \ldots, \bar{x}_{k2j})$ 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:

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

The above log odds ratio can then be used to predict crashes by establishing a threshold value that yields desirable classification accuracy.

4.3.2 Model Building: Data Analysis

The results from the analysis showed that three parameters, namely, *LogCVS, SV* and *AO* are most significantly associated with crash occurrence. These three parameters still correspond to 126 variables (measured from 7 stations during 6 time slices) as potential independent variables in the final model. Also, based on the results from the previous section we can discard *Station B, C* and *D*. Even though hazard ratio from these stations were significantly different from unity and also appeared to be different when compared between real and pseudo matched datasets, they are less significant than their counterparts belonging to *Station E, F,G* and *H.* This meant that any model comprising these factors together (From stations *B, C, D* as well as *E, F, G* and *H*) would invariably show the factors from way upstream stations as insignificant.

One might suggest that even if that's the case we should still examine both the full and the reduced models and make our decision based on the classification accuracy. This would not be a good idea since the modeling procedure requires all variables used in the model to be non-missing (i.e. complete case analysis) in order to use any observation from the dataset for model building. Now it should be understood that seven stations from which data are collected would not be simultaneously available at all the time (data limitation due to frequent loop failure). It means that some variables will be missing in certain observations. The number of observations used for model building can drastically be reduced if a lot of independent variables are used.

Also, even though time duration 1 (0-5 minutes) prior to crash exhibits significant hazard ratios (See Tables 1, 2 or 3), it is too close to the actual time of the crash and thus practically not useful for crash prediction models. This time duration is thus ignored from further considerations.

For each of the remaining time slices, we thus have $p = 12$ traffic flow variables; *LogCVS, SV,* and *AO* at each of the four loop detectors *E, F, G* and *H*. To identify all significant variables from these variables, the binary outcome variable y is now modeled using stratified conditional logistic regression method described above in the previous section. Note that this stratified analysis is widely known as matched case-control analysis. *SAS* Procedure *PHREG* is used with some modification of matched data to fit the proposed conditional logistic regression model. The procedure allows one to identify significant variables using standard automatic search techniques: *stepwise, forward* and *backward.* Full description of the three automatic search procedures can be found in Hosner and Lemeshow (1989). The beta coefficients and the hazard ratios are obtained for significant variables found under all three-search procedures.

These procedures 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. Now similar search procedures from subsequent time slices resulted in slightly different models involving variable measured during time slice 3, 4 and so on. The decision regarding selection of 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 4. First two variables have positive beta coefficients (and hazard ratio greater than 1), which mean that the odds of a crash increase as these variables increase while *SVG2* had negative beta coefficient implying increasing odds of crash as this parameter decreases. This indicates that high variation in speed at a freeway location coupled with high occupancy and low variation in volume downstream of this site increases the likelihood of having a crash at that location with in next 10 minutes. This structure of the final model also conforms to our findings reported earlier. This model is an improvement to the earlier model developed with only the 1999 data that was reported earlier to FDOT and published in Abdel-Aty et al. (2004).

Table 4-4: Final Model Description

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
LogCVSF2	1	1.21405	0.15548	60.9729	<.0001	3.367
AOG2	1	0.02466	0.00571	18.6747	<.0001	1.025
SVG2	1	-0.19124	0.04569	17.5216	<.0001	0.826

4.3.3 Classification Accuracy of the Model

As previously explained in the model building methodology, the odd ratio given by Equation 4 may be used to classify crash and non-crash cases. For this purpose, it was decided to evaluate the model first on the dataset used to develop this model. We first calculated the mean of the three variables *LogCVSF2, AOG2* and *SVG2* of all five non-crashes within each of the 1528 matched stratum of the dataset. For the j^{th}-matched set, the vector x_{2j} in Equation 4 may then be replaced by the vector of these non-crash means. The odds ratios for each observation in the data set are then calculated in Equation 4 utilizing the beta coefficients from Table 4-4 where the vector x_{1j} is the actual observation in the data set. An observation is classified as a crash if the corresponding odds ratio is greater than 1, and a non-crash if the odds ratio is less than or equal to 1. The classification table obtained this way for the 1:5 matched data set is shown in Table 4-5. It may be observed that more than 62.41% of crashes are identified using this threshold for the odd ratio. Note that this threshold (chosen to be equal to one here) may be varied in order to achieve desirable classification given the tradeoff between overall classification accuracy (crash and non-crash) and crash identification.

Table 4-5: Classification results from the dataset used to develop the model

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

4.4 Comparison of Classification Accuracy: Simple vs. Multivariate Models

While 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. Also, the classification accuracy for the models needs to be examined by evaluating the performance of the models at various threshold values for the odds ratio cut-off.

The methodology to classify patterns using the models is based on the ratio of odds of having a crash vs. no crash. It is used as the classification criterion and is obtained as per equation 4. The odds ratio was obtained for every observation of the dataset using coefficients from their respective models (i.e., single and multivariate) and then these observations may be classified using a threshold value of this odd ratio. The classification accuracy is sensitive to it and hence an arbitrary selection of the threshold is not preferable. Cumulative proportions of crashes above and non-crashes equal or below a range of these odd ratios were determined and are plotted against odd ratios in Figures 4-4 and 4-5 for a one-covariate model (with only *LogCVSF2* as input) and the multivariate model, respectively. For convenience, only odd ratio threshold less than or equal to 5 are shown on the horizontal axis. Also, of all simple models the classification plot is shown for only *LogCVSF2* which happens to be the single most significant variable.

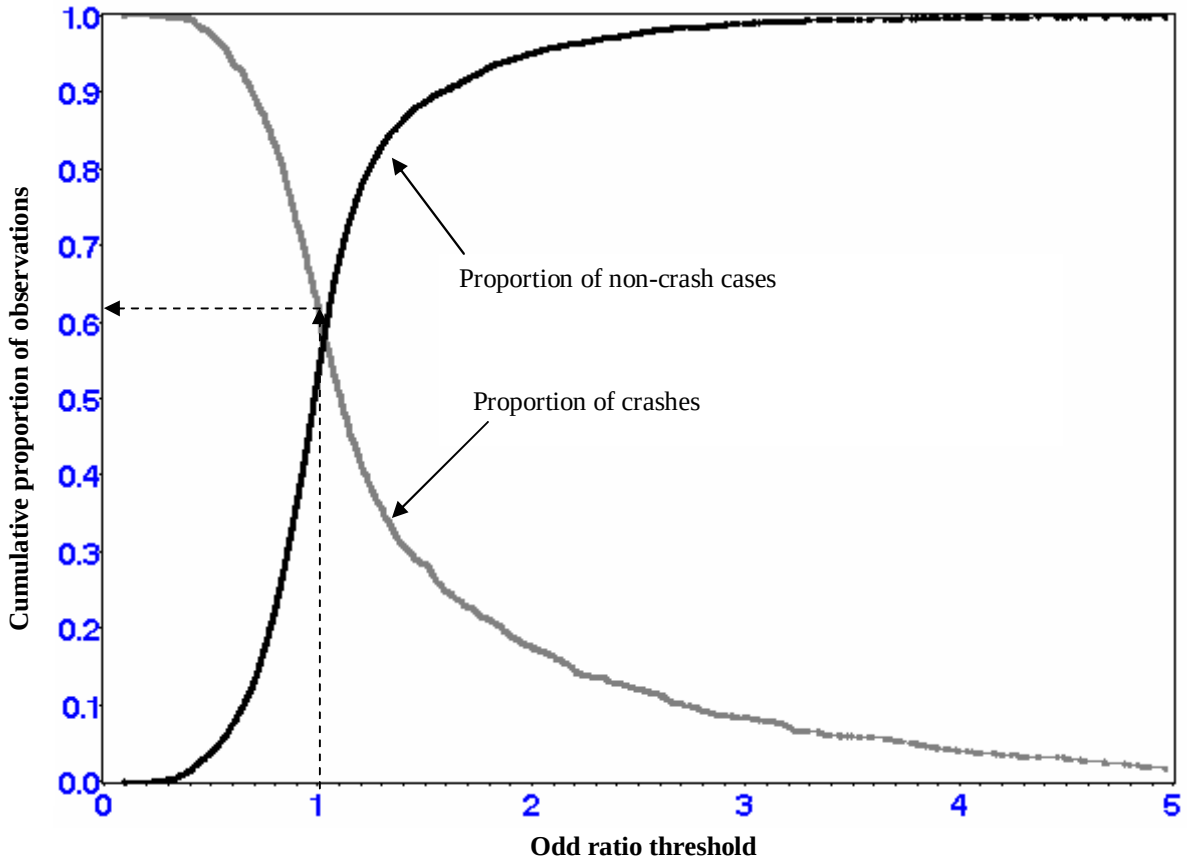

Figure 4-4 Classification performance of the multivariate model: Cumulative proportion of crashes above and non-crash cases below a range of odd-ratio threshold values (Grey Curve: Proportion of crashes and Black Curve: Proportion of non-crash cases)

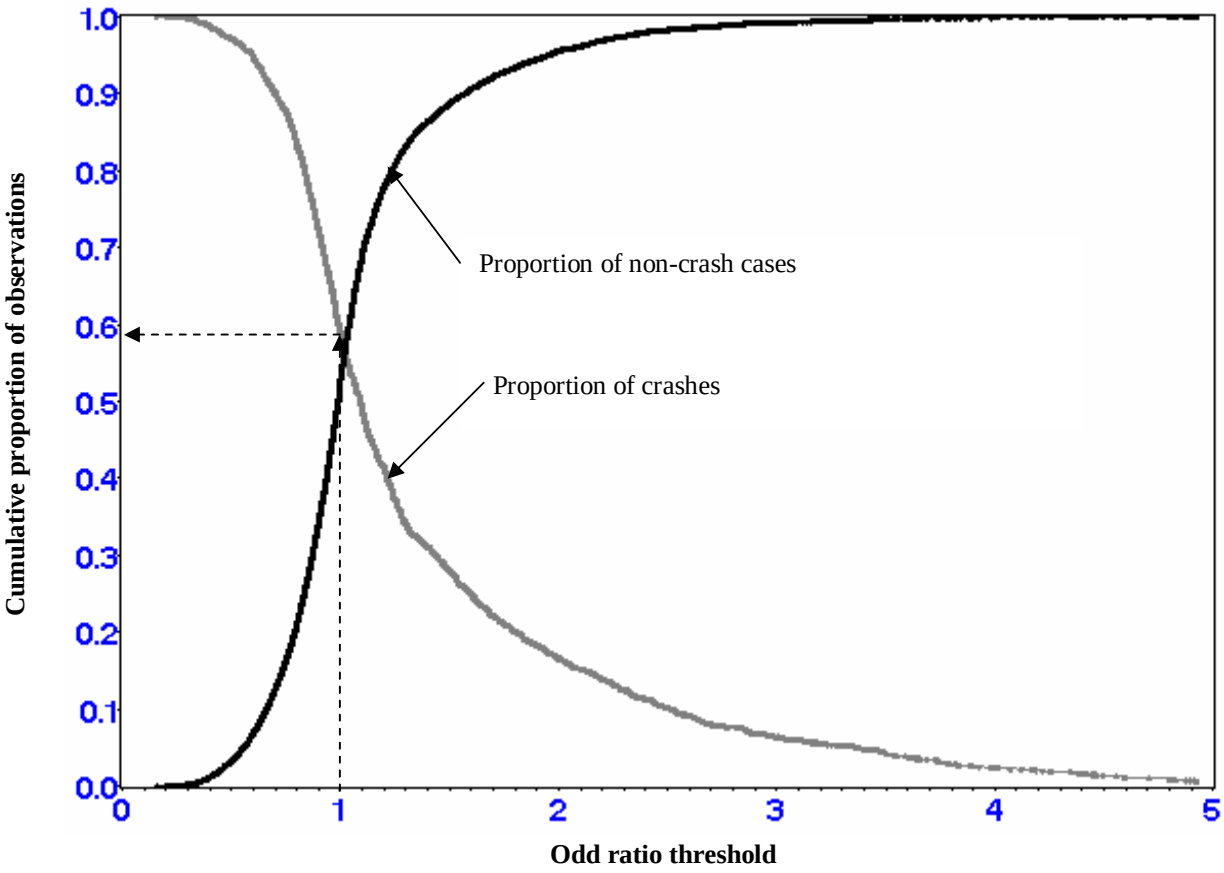

Figure 4-5 Classification performance of the simple model with *LogCVSF2* as covariate: Cumulative proportion of crashes above and non-crash cases below a range of odd-ratio threshold values (Grey Curve: Proportion of crashes and Black Curve: Proportion of non-crash cases)

As expected, the cumulative proportion of crash cases decreases as odd ratio increases and the cumulative proportion non-crash cases increases as odd ratio increases. The grey (lighter) curve indicates the cumulative proportion of crash cases that are greater than the corresponding odd ratio on the horizontal axis and the black (darker) curve indicates the cumulative proportion of non-crash cases less than or equal to the corresponding odd ratio. One may choose a threshold value of odd ratio along the horizontal axis and determine the proportions of crashes and non-crash cases that would be correctly classified by the corresponding model. For example, if odd ratio of one is chosen as the cut off point, then about *62%* of crashes and *53%* of the non-crash cases in the dataset would be correctly classified by the multivariate model presented above. The crash identification is only *59%* when the single covariate model with *LogCVSF2* is used for classification (with odd ratio cutoff set at 1.0). A multivariate model, therefore, is recommended for reliable classification of the patterns.

These graphs are useful in selecting an odd ratio value that would satisfy the requirement of a desired accuracy. Note, however, that both (crash and non-crash) 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 of odd ratio may be set as threshold so that most of the crashes are identified even if that increases the number of "false-alarms" because speed is known to be positively associated with the severity of crashes.

Once an appropriate cut-off for odd ratio is determined, the models may be applied to identify crash prone locations on the freeway in real-time using the methodology discussed in the next

chapter. However, as we shall see later, the online application proposed at this stage would mostly be suitable for traffic management authorities and therefore the decision of threshold doesn't seem as critical. The threshold of unity provides reasonable balance between the two conflicting attributes (i.e., overall classification and crash identification) and hence is recommended as the cut-off value.

4.5 Concluding Remarks

Before implementation of the models (simple or multivariate) to provide real-time information to the drivers many issues will need to be investigated including the means of notifying the drivers of the potential of a crash, and the expected reaction of drivers to such warnings. The authors believe that the problem of intervening with measures such as variable speed limits or warning signs, and thereby reducing the freeway crash potential is a non-trivial one and demands further investigation (will be addressed in phase 2). However, the models developed here may still be used to classify the patterns in the loop detector data so that the traffic management authorities can anticipate impending hazard and can prepare accordingly.

In this chapter the final model was developed based on our findings from Abdel-Aty et al., 2004 presented earlier at the 83rd annual meeting of the Transportation Research Board (analysis with crash data only from year 1999) and the analysis presented in this report. In the next chapter, methodology for real-time application of the final model in association with simple models is proposed.

CHAPTER 5

IMPLEMENTATION PLAN

The models described in the previous chapter include the simple (one covariate) models as well as the final multivariate model. The classification accuracy of these models was also discussed. The implementation strategy for these models is proposed here. The models would help the traffic management authorities to identify which locations on the freeway are currently crash prone. This information may then be used by the authorities to have the crash mitigation set up ready around these freeway locations so that if (and when) a crash does occur, it can be immediately responded to. This way its impact on freeway operation can be minimized. This would be the first step toward the proactive traffic management system that this research aims to develop. More aggressive intervention strategies such as variable speed limits, flashing warning on the VMS (Variable Message Sign) etc. require more focused research efforts and are currently being evaluated as part of the next phase of this project.

The two set of models (simple and multivariate, respectively) will be utilized in the two phased implementation plan proposed here. The first phase constitutes a preliminary risk assessment using the simple models. When some locations appear to have a high likelihood of crash occurrence within the next 10-15 (slice 3) minutes based on the results from the simple models; data from corresponding stations is subjected to the multivariate model to obtain the final crash prediction for these freeway segments.

5.1 Simple Models Implementation

5.1.1 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 the previous chapter was the hazard ratio for 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 covariate were zero.

For a real-time application, the instrumented freeway corridor can be divided in to 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 report. 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* (2 upstream stations, station closest to the crash and 2 stations downstream, respectively). Note that based on the analysis, 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 station *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 values of hazard ratio corresponding to *LogCVS* measured at these five stations (*D-H*) at all six time slices are shown in Table 5-1. Among all possible parameters *LogCVS* was chosen because the plot depicting spatio-

temporal variation of crash risk (Figure 4-2) showed stark contrast between station of the crash and other locations.

Table 5-1: Hazard ratios from single covariate models consisting of *LogCVS* from five stations and six time slices

Hazard ratio corresponding to station	Hazard ratio to asses the crash risk 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	3.132	2.430	3.074	2.735	2.499
E	4.436	3.335	3.025	3.257	2.664	2.426
F	7.237	5.580	4.485	3.801	3.654	3.809
G	4.705	3.899	3.037	3.519	3.209	2.964
H	3.976	3.635	3.476	3.139	2.623	2.871

With the hazard ratio for *LogCVS* from station *D to* station *H* (shown in Table 5-1) 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 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 multiplied by the corresponding hazard ratio to obtain the measure of crash risk for a period up to 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, i.e. 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 Chapter 4. Based on the changing patterns depicted by the continuously updated plots, freeway locations with high crash risk may be identified in real-time.

5.1.2 Simple Models: Illustration

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 non-crash case. Table 5-2 shows a sample of *LogCVS* calculated as a moving average from real-life historical traffic speed data from a set of five detectors, 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 (Station of the crash was 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 section 4.2.2. The detailed snapshot of the data used to get this sample is shown in the Appendix.

Table 5-2: 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	LogCVS
4/6/99 4:19:30 PM	32 (D)	34	1.42
4/6/99 4:20:00 PM	32 (D)	34	1.42
4/6/99 4:20:30 PM	32 (D)	34	1.45
4/6/99 4:19:30 PM	33 (E)	34	1.60
4/6/99 4:20:00 PM	33 (E)	34	1.65
4/6/99 4:20:30 PM	33 (E)	34	1.67
4/6/99 4:19:30 PM	34 (F)	34	1.42
4/6/99 4:20:00 PM	34 (F)	34	1.43
4/6/99 4:20:30 PM	34 (F)	34	1.52
4/6/99 4:19:30 PM	35 (G)	34	1.56
4/6/99 4:20:00 PM	35 (G)	34	1.57
4/6/99 4:20:30 PM	35 (G)	34	1.59
4/6/99 4:19:30 PM	36 (H)	34	1.71
4/6/99 4:20:00 PM	36 (H)	34	1.69
4/6/99 4:20:30 PM	36 (H)	34	1.74

Table 5-1 depicted the hazard ratios corresponding to station *D-H* at all six time slices. In Table 5-3 (a)-5-3 (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 5-3 (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, forth 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 asses the risk for various future time periods the same value of *LogCVS* is used, however the value of hazard ratio is as per the time slice, e.g., for next 10 minutes hazard ratio corresponding to slice 2, for next 15 minutes hazard ratio corresponding to slice 3 and so on.

Table 5-3 (b) is generated through a similar procedure, the only difference being that the values for *LogCVS* are now updated as per the most recent speed observations. In table 5-3(c) the value of the independent covariate (*LogCVS*) are further updated with the most recent speed observations. It may be noted that in Table 5-3(a) the values of *LogCVS* are highlighted in yellow to associate them with the observations in Table 5-2 from the same point in time (4:19:30 PM). Similarly, in Table 5-3(b) and Table 5-3(c), the updated values for *LogCVS* are highlighted red and green to associate them with their respective times of observation (4:20:00 and 4:20:30 PM respectively) in Table 5-2.

Table 5-3 (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 *LogCVS* 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 5-3 (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 *LogCVS* 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

Table 5-3 (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 *LogCVS* 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

Three contour plots depicting the variation in crash risk generated from this data are shown in Figure 5-1 (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 5-1 (a) comes from the corresponding cells of Table 5-3 (a) and the plot is updated into Figure 5-1 (b) as soon as the new set of readings are recorded after 30 seconds. The values for contour variable in the updated plot, Figure 5-1 (b), are given by Table 5-3 (b) which eventually turns into Figure 5-3 (c) after 30-seconds taking input from Table 5-3 (c). The updated patterns do not differ a lot from their predecessor since most of the observations contributing to 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 corresponding matched non-crash case (On April 27, 1999 from the same set of stations) shown in figure 5-2 (a) – (c). The detailed snapshot of data used to generate these plots may also be found in the Appendix.

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. It should be noted that the difference between crash and non-crash 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 to consider it as a warning.

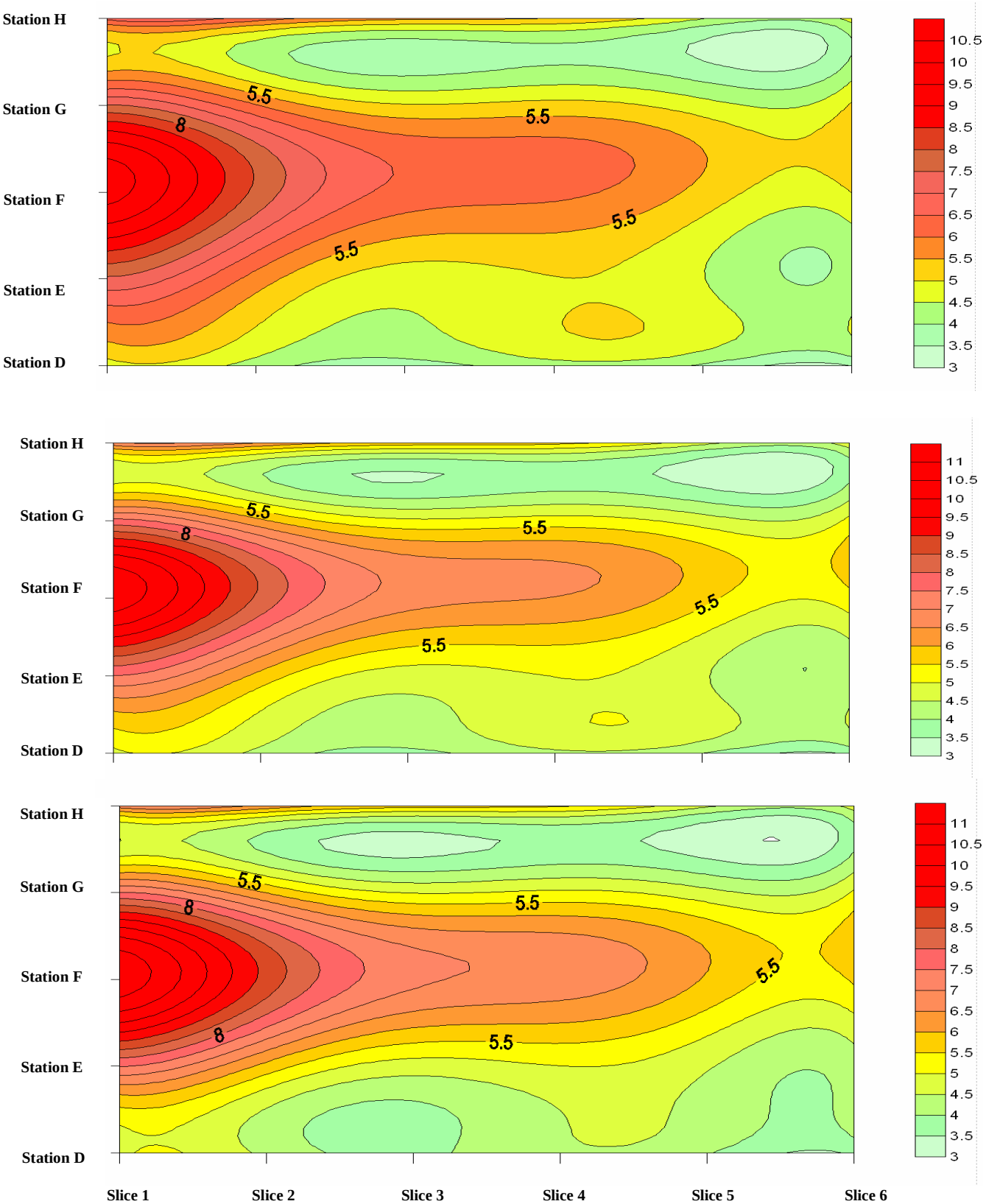

Figure 5-1 (a-c): illustrative pattern of variation in measure for risk of observing a crash in the segment belonging to station F updated every 30-second 15 minutes prior to a crash

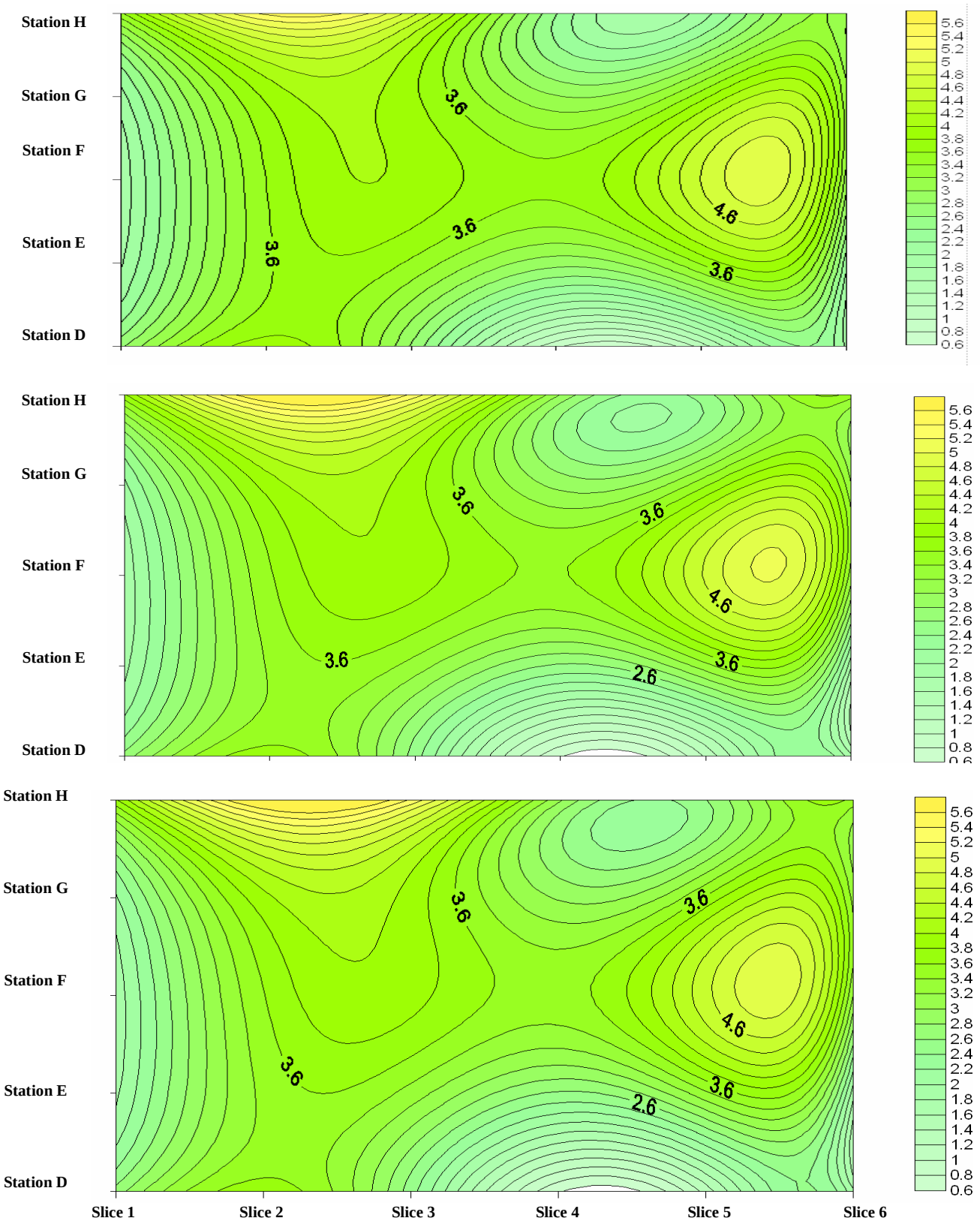

Figure 5-2 (a-c): Illustrative pattern of variation in measure for risk of observing a crash in the segment belonging to station F updated every 30-second for a non-crash scenario

The simple models are proposed to be applied before the multivariate model which employs data from three stations 5-10 minutes (slice2) prior to crash occurrence. Keeping this in perspective an effective online application strategy would be to examine closely the plots in the region where the abscissa encompasses slice 3, since the hazard ratio corresponding to slice 3 would give an indication for crash occurrence with in next 10-15 minutes. If the sequential patterns appear hazardous as is the case with those depicted in Figure 5- 1 (a - c) then the multivariate model can be employed to classify the patterns to assess the crash occurrence with in next 5-10 minutes (note that the factors appearing in the final model use data from slice 2). The application for the multivariate model is described in the following section.

5.2 Application of multivariate models

5.2.1 Procedure and Data Requirement

Following the detection of hazardous patterns through the contour plots the multivariate model may be applied for classification of patterns into leading or not leading to a crash. As explained in the previous chapter, the log odds calculated from the following equation may be used to classify the patterns into crash and non-crash cases.

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

For this purpose, we first calculate the mean for the three covariates included in the final model *LogCVSF2* and *AOG2* and *SVG2* on all five non-crashes within each matched set of the *1:5* matched dataset. For j^{th}-matched set, the vector $\bar{x}_{k2j}$ in Equation 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 odds ratio for the purpose of identifying a crash. Equation 4 with estimated values of the parameters can be rewritten as:

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

The *RHS* of above equation is the odds ratio and it may be noted that β_p (model coefficients) in equation 4 have been replaced with the parameter estimate for coefficients of *LogCVSF2, AOG2, and SVG2* from Table 4-4. The averages ($\bar{x}_{k2j}$) have been replaced with the respective means of these covariates over non-crash cases in the matched dataset. The values for the three parameters obtained from the loop detectors in real-time would be used as independent variables in this equation to obtain the odds ratio of having a crash vs. not having a crash. As explained in the previous chapter if the resultant odd ratio exceeds unity then the patterns would be classified as crash.

These odds ratio may also be updated in a way similar to the contour plots. To update the odds ratios after every 30-second period the last set of observations in the 5-minute period may be replaced by the data most recently recorded.

5.2.2 Multivariate Model: Illustration

Table 5-4 shows the historical values for the three covariates included in the final model starting ten minutes prior to the same crash which was used to illustrate the application of the simple models. These values are calculated on a continuous basis i.e., the averages and standard deviations are calculated as a moving average. The procedure to obtain the input parameters through the moving average is same as described in the implementation plan for the simple

models. In the first row the three input parameters (5-minute average occupancy, 5-minute standard deviation of volume and 5-minute coefficient of variation in speed) are obtained using the last 10 observations (5 minutes*2 observations every minute) from corresponding stations, in the subsequent rows the parameters are updated using the two most recent observations while discarding the first of the set of ten observations. The odds ratios are then calculated as per equation 5. The odds ratios of having a crash vs. not having a crash near *station F* for the observed values of independent variables are also shown in the table.

It may be seen that at three instances the odds ratio is greater than unity and hence the model classifies the data patterns as "crash". It is expected since it is already known that a crash did occur following these data patterns.

Since the final model included the parameters from time slice 2, the odds of crash occurrence within next 10 minutes are assessed. Data from *station F* and *G* (station of the crash and the one immediately following it in the downstream direction) may be collected and updated continuously every 30-second as shown in Table 5-4 to obtain the risk for crash occurrence within next ten minutes.

Table 5-4: The output from the final multivariate model when applied on the historical loop data continuously updating every 30-second starting ten minutes prior to the crash

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

5.3 Concluding Remarks

This chapter presents a simple approach to implement a two-step methodology to apply the models developed in this study. It is shown with the help of illustrative examples that the loop data patterns emanating in real-time may be used to assess whether or not a crash is likely to occur on certain freeway segments. However, there are some issues which can not be resolved at this stage; one of them is the determination of the threshold cut-off value separating crash and non-crash cases. The cut-off value will really depend upon the desirable proportions of conflicting attributes namely, crash identification and false alarms. We have proposed a universal cut-off of unity for odds ratio to separate crashes from non-crashes since these models at this stage are proposed to be used only by the traffic management authorities. The application would be much more critical in the future while we examine more aggressive intervention strategies such as warning the drivers through variable message signs etc., to reduce the potential of crash occurrence. Then the possibility of having separate cutoffs under different operation regimes (e.g. congested or free flow) or at different locations will also need to be explored. The other unresolved issues involve the question of how the model output may be utilized. These issues will be part of the next phase and are discussed in the next chapter.

CHAPTER 6

CONCLUSIONS AND FUTURE SCOPE

6.1 Summary and Conclusions

The objective of this ambitious and innovative research endeavor was to develop a methodology to link ITS-archived data with historical crashes on instrumented I-4 corridor and it has been achieved with fair amount of success. The research group has assembled a detailed crash database for all crashes that occurred on I-4 in the years 1999-2002 (with a plan for extending it to 2003 for the next phase of the project), linking them to the archived loop detector data, and also to the geometric characteristics at the crash locations. It was proved statistically that turbulence in traffic conditions before a crash (both time and space) is associated with crash occurrence. This means that we can predict crashes if this turbulence is observed in the future.

The methodology for analysis was case-control logistic regression with a matched study design. The matched design of the study accounts for the external factors such as the freeway geometry, time of the day and day of the week. A series of crash prediction models were estimated based on the statistical link between crash occurrence and the turbulence in the traffic flow observed through the loop detectors. The simple models (involving one covariate) were estimated first, following the exploratory analysis. Also a multivariate logistic regression model was estimated following a step-wise procedure. For the final model, 5-minute average occupancy and 5-minute standard deviation of volume observed at the downstream station, during the slice of 5-10 minutes prior to the crash along with the 5-minute coefficient of variation in speed at the station closest to the location of the crash during the same time slice were found to affect the crash occurrence most significantly. The final model developed was used to calculate the log-odds

ratio of observing a crash vs. not observing a crash. A threshold value for this ratio may then be set in order to determine whether the location has to be flagged as a potential "crash location". 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.

It should be noted that even though the simple models did not achieve classification accuracy as well as the final model, the advantage of using those models is that they have very tolerant data requirements. Besides, it was shown that the results from these models could be used to obtain a spatio-temporal variation of the crash risk. A real-time application plan for these models was demonstrated in the report. Essentially the plan proposed here states that a preliminary assessment of the freeway conditions may be made using the plots generated by simple models and then if the conditions appear to be hazardous the data may be 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 their crash mitigation squad on alert so that the impacts of crash occurrence may be minimized. Also, if there are some freeway segments where the models trigger the warning more often than the other locations, these segments may be closely watched through the freeway cameras. This will help recognize any problems associated with these locations such as weaving sections, ramps, etc.

6.2 Future Scope

The first phase of the research summarized in this report points toward much more work that needs to be done beyond this phase to incorporate other data elements that are related to traffic safety.

The existing models crash prediction level is around 62%. This is acceptable since we are only capturing the traffic and geometric factors in this current effort. It is very well known from the traffic safety literature that the drivers' characteristics and environmental factors are very important to crash occurrence phenomena. Therefore, we don't envision better prediction percentages without accounting for these elements. A temporal system of models on I-4 to relate crashes to ITS-archived data has been proposed as an extension to this study. We envision different models for different times of day and possibly day of week. A clear idea about the driver population by time of the day and day of the week and freeway location would enhance the prediction level. Different models applied in different situations might be needed to operate in a multi model system to achieve the best prediction level.

Our experience from this phase has shown us also that it is not sufficient to know the environmental conditions, such as rain, in crash situations (which is the case now), but we need also to know the environmental conditions in non-crash situations. Therefore, we have proposed in the extension to the current project to obtain detailed rainfall data from the National Weather Services (NWS) as well as other sources for the whole study period. Not only rain data will be needed but rain intensity data might be needed. Light rain might not be correlated with crash occurrence, but heavy rain could likely be.

The research group also believes that the overall system of crash prediction will not be as simple as initially envisioned, but multiple modeling techniques will have to be used and implemented simultaneously in real-time to achieve better prediction levels. The modeling methodology would encompass statistical as well as Neural Networks in a hybrid manner to achieve the best

prediction level. We expect also that this expanded effort and data will allow us not only to identify crashes but also the crash type (harmful event). It is likely that the traffic conditions that affect for example a rear-end collision would be different than those affecting a side swipe or single vehicle collision and the kind of intervention needed to calm these conditions would be different. Another important issue that has to be investigated is whether and to what extent if at all false crash alarms will exist.

Finally, the implementation plan proposed in this report adds to the incident detection capabilities of traffic management authorities. At the most it can be effective in identifying the locations that experience the crash prone conditions more often than the others. However, one must understand that more proactive intervention plan such as the variable speed limits, flashing warning on the Variable Message Sign etc. demand a more careful analysis which is beyond the scope of this phase. In other words, once the system identifying the potential for a crash is developed, there would be a plethora of questions that need to be answered regarding how to reduce the impending hazard to avoid the crash. Is it by posting messages on the electronic VMS signs or by using the Variable Speed Limit approach? What is the message that should be displayed, or what is the new speed limit that would be displayed? How would drivers react? And how would there reaction or new speed limits affect the potential of the crash. These are critical issues that will be investigated to guarantee the success of the on-line crash prediction system. Even if there is a perfect system, we cannot assume that it would have the desired effect without fully understanding how to convey the warning and how this warning would affect the crash potential.

The aforementioned issues are discussed in detail in our proposal to the Florida Department of Transportation for the extension of this project (phase 2). These issues logically follow the findings presented in this report. It is evident from the fact the proposal for the second phase of this project has been accepted by the *DOT*.

The research effort documented in this report has put *UCF* and the state of Florida *DOT* in the forefront of developing real-time proactive crash prediction models. Our extensive review of the literature shows that very limited work is starting in this area of research, and mostly theoretical without solid ideas to implement it. The research group is of the opinion that while the current phase was critical for a clearer understanding of this innovative research problem it will be the next phase in which we deliver the final product to achieve the goals of this endeavor.

APPENDIX

Table A-1: A detailed snapshot of the raw 30-second data 15 minutes prior to a crash from a set of five stations

		Station of		Speed			Volume			Occupancy		
Date-Time	Station	Crash	Direction	Left Lane	Center Lane	Right Lane	Left Lane	Center Lane	Right Lane	Left Lane	Center Lane	Right Lane
4/6/99 4:15 PM	32 (D)	34	E	Missing	31	39	Missing	14	9	Missing	22	9
4/6/99 4:15 PM	32 (D)	34	E	Missing	35	38	Missing	12	12	Missing	21	13
4/6/99 4:16 PM	32 (D)	34	E	Missing	37	41	Missing	14	14	Missing	19	16
4/6/99 4:16 PM	32 (D)	34	E	Missing	34	45	Missing	14	11	Missing	19	9
4/6/99 4:17 PM	32 (D)	34	E	Missing	12	32	Missing	8	15	Missing	42	22
4/6/99 4:17 PM	32 (D)	34	E	Missing	25	37	Missing	16	13	Missing	32	16
4/6/99 4:18 PM	32 (D)	34	E	Missing	17	35	Missing	13	14	Missing	33	14
4/6/99 4:18 PM	32 (D)	34	E	Missing	26	40	Missing	17	12	Missing	30	11
4/6/99 4:19 PM	32 (D)	34	E	Missing	26	42	Missing	12	12	Missing	22	12
4/6/99 4:19 PM	32 (D)	34	E	Missing	24	36	Missing	11	10	Missing	24	11
4/6/99 4:20 PM	32 (D)	34	E	Missing	34	43	Missing	16	6	Missing	19	5
4/6/99 4:20 PM	32 (D)	34	E	Missing	25	46	Missing	9	8	Missing	16	7
4/6/99 4:19 PM	33 (E)	34	E	3	12	27	3	10	14	29	24	16
4/6/99 4:20 PM	33 (E)	34	E	4	11	15	7	10	11	48	24	23
4/6/99 4:20 PM	33 (E)	34	E	5	13	14	7	10	13	44	23	30
4/6/99 4:19 PM	34 (F)	34	E	33	32	37	18	15	16	21	18	6
4/6/99 4:20 PM	34 (F)	34	E	23	23	33	14	17	14	25	29	5
4/6/99 4:20 PM	34 (F)	34	E	10	14	12	10	8	9	42	28	3
4/6/99 4:19 PM	35 (G)	34	E	7	20	11	10	13	9	41	23	31
4/6/99 4:20 PM	35 (G)	34	E	19	32	33	13	15	16	26	16	17
4/6/99 4:20 PM	35 (G)	34	E	13	22	19	9	14	11	25	24	20
4/6/99 4:19 PM	36 (H)	34	E	15	20	15	13	14	15	37	30	36
4/6/99 4:20 PM	36 (H)	34	E	28	22	25	14	12	17	15	25	22
4/6/99 4:20 PM	36 (H)	34	E	33	29	32	15	17	11	14	26	11

Table A-2: A detailed snapshot of the raw 30-second data under normal conditions on freeway from a set of five stations

Date-Time	Station	Station of Crash	Direction	Speed			Volume			Occupancy		
				Left Lane	Center Lane	Right Lane	Left Lane	Center Lane	Right Lane	Left Lane	Center Lane	Right Lane
4/27/99 4:15 PM	32 (D)	34	E	Missing	41	47	Missing	9	8	Missing	15	6
4/27/99 4:15 PM	32 (D)	34	E	Missing	46	44	Missing	11	7	Missing	13	7
4/27/99 4:16 PM	32 (D)	34	E	Missing	47	48	Missing	11	11	Missing	10	10
4/27/99 4:16 PM	32 (D)	34	E	Missing	46	45	Missing	15	7	Missing	14	9
4/27/99 4:17 PM	32 (D)	34	E	Missing	45	45	Missing	14	7	Missing	13	8
4/27/99 4:17 PM	32 (D)	34	E	Missing	42	43	Missing	13	3	Missing	18	3
4/27/99 4:18 PM	32 (D)	34	E	Missing	46	48	Missing	14	12	Missing	14	9
4/27/99 4:18 PM	32 (D)	34	E	Missing	48	46	Missing	12	6	Missing	10	5
4/27/99 4:19 PM	32 (D)	34	E	Missing	45	48	Missing	14	6	Missing	14	5
4/27/99 4:19 PM	32 (D)	34	E	Missing	47	49	Missing	12	9	Missing	11	7
4/27/99 4:20 PM	32 (D)	34	E	Missing	47	49	Missing	11	9	Missing	10	8
4/27/99 4:20 PM	32 (D)	34	E	Missing	44	45	Missing	15	8	Missing	13	8
4/27/99 4:19 PM	33 (E)	34	E	62	57	64	8	12	11	4	6	5
4/27/99 4:20 PM	33 (E)	34	E	60	62	64	7	8	10	3	4	5
4/27/99 4:20 PM	33 (E)	34	E	52	55	63	11	13	18	8	7	8
4/27/99 4:19 PM	34 (F)	34	E	60	55	59	17	18	16	11	13	6
4/27/99 4:20 PM	34 (F)	34	E	59	58	60	9	9	14	6	6	5
4/27/99 4:20 PM	34 (F)	34	E	60	55	62	14	13	19	9	11	7
4/27/99 4:19 PM	35 (G)	34	E	51	57	59	14	15	22	9	10	12
4/27/99 4:20 PM	35 (G)	34	E	48	56	57	15	15	13	10	9	7
4/27/99 4:20 PM	35 (G)	34	E	45	50	57	17	19	14	11	14	7
4/27/99 4:19 PM	36 (H)	34	E	55	53	50	17	14	9	9	14	5
4/27/99 4:20 PM	36 (H)	34	E	40	32	34	13	12	11	10	18	11
4/27/99 4:20 PM	36 (H)	34	E	43	46	41	13	14	9	9	15	7

REFERENCES

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

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

3. 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. Presented at the *83rd Annual Meeting of the Transportation Research Board* (TRB), Washington D.C., 2004.

4. Adeli, H., and Karim, A., Fuzzy-wavelet RBFNN model for freeway incident detection. *Journal of Transportation Engineering*, Vol. 126, No. 6, 2000, pp. 464-471.

5. Agresti, A., Categorical data analysis, 2nd Ed. *John Wiley and Sons, Inc.*, 2002.

6. Collett, D., Modelling binary data. *Chapman and Hall*, 1991.

7. Garber, N., and Ehrhart, A., The effect of speed, flow, and geometric characteristics on crash frequency for two-lane highways. *Transportation Research Record, No. 1717, Transportation Research Board, National Research Council*, Washington, D.C., 2000, pp. 76-83.

8. Golob, T. F., and Recker, W. W., A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A*, 2003, In press.

9. Golob, T. F., and Recker, W. W., Relationships among urban freeway accidents, traffic flow, weather and lighting Conditions. *California PATH Working Paper UCB-ITS-PWP-2001-19, Institute of Transportation Studies.* University of California, Berkeley, 2001.

10. Golob, T. F., Recker, W. W., and, Alvarez, V.M., A tool to evaluate the safety effects of changes in freeway traffic flow. Presented at *the 82nd annual meeting of Transportation Research Board,* Washington, D.C., 2003.

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

12. 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.

13. Kockelman, K. M., and Ma, J., Freeway speeds and speed variations preceding crashes, within and across lanes. Presented at *the 83rd annual meeting of Transportation Research Board,* Washington, D.C., 2004.

14. Lee, C., Saccomanno, F., and Hellinga, B., Analysis of crash precursors on instrumented freeways. *Transportation Research Record, No. 1784, Transportation Research Board, National Research Council,* Washington, D.C., 2002, pp. 1-8.

15. Lee, C., Saccomanno, F., and Hellinga, B., Real-time crash prediction model for the application to crash prevention in freeway traffic. Presented at *the 82nd annual meeting of Transportation Research Board,* Washington, D.C., 2003.

16. Lee, C., Hellinga, B., and Saccomanno, F., Assessing benefits of variable speed limits. Presented at *the 83rd annual meeting of Transportation Research Board,* Washington, D.C., 2004.

17. Madanat, S., and Liu, P., A prototype system for real-time incident likelihood prediction. *IDEA Project Final Report (ITS-2), Transportation Research Board, National Research Council,* Washington, D.C., 1995.

18. 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.

19. Pande, A., Classification of real-time traffic speed patterns to predict crashes on freeways. MS Thesis, *University of Central Florida*, 2003.

20. Pande, A., and Abdel-Aty, M., Spatio-temporal variation of risk preceding crash occurrence on freeways. *Manuscript prepared for publication*, 2004.

21. Stamatiadis, N., and Deacon, J., A., Quasi-induced exposure: methodology and insight. *Accident Analysis and Prevention, Vol. 31,* 1999, pp. 705-718.

22. Xiao, J., Kulakowski, B., T., and El-Gindy, M., 2000. Prediction of wet-pavement accidents: fuzzy logic model. *Transportation Research Record, No. 1717, Transportation Research Board, National Research Council*, Washington, D.C., 2000, pp. 28-36.