Florida Department of Transportation
Research Project Work Order #4, Contract No. BD-548

Final Report

Relating Crash Occurrence to Freeway Loop Detectors Data, Weather Conditions and Geometric Factors

By

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

Anurag Pande
Ph.D. Candidate

Nizam Uddin, Ph.D.
Associate Professor

Jeremy Dilmore
Graduate Student

Rajashekar Pemmanaboina
Graduate Student

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

October 2005

September 20051. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle RELATING CRASH OCCURRENCE TO FREEWAY LOOP DETECTORS DATA, WEATHER CONDITIONS AND GEOMETRIC FACTORS		5. Report Date October 1, 2005 6. Performing Organization Code
7. Author(s) Abdel-Aty M., Pande A., Uddin N., Dilmore J., Pemmanaboina, R.		8. Performing Organization Report No.
9. Performing Organization Name and Address University of Central Florida Department of Civil & Environmental Engineering Orlando, FL 32816		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Agency Name and Address Florida Department of Transportation FDOT Research Center 605 Suwannee Street M.S. 30 Tallahassee, FL 32399-0450		13. Type of Report and Period Covered 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

In this research loop detector data patterns significantly associated with crash risk on the freeway are explored. Historical crash and loop detector data from the Interstate-4 (I-4) in Orlando have been used for this research. The possibility of using real-time weather information as part of the crash prediction system on I-4 is investigated. A “rain index” based on the archived rain data from certain locations around the freeway corridor is developed. It is then used, in addition to the loop data, in one of the existing crash prediction models. It was shown that including the “rain index” improves the classification performance of the model marginally. Seemingly Unrelated Negative Binomial (SUNB) crash frequency models were used to identify the geometric design elements of the freeway that could be used as inputs to the real-time crash prediction models. It was found that the presence of ramps is one of the most significant geometric design elements affecting crash risk on the freeway.
A section of the I-4 corridor was simulated using the microscopic traffic simulation model PARAMICS. Applying real-time crash prediction model(s) to the simulation, data a measure of crash risk was estimated. ITS strategies involving variable speed limits (VSL) were then employed in the simulation environment to examine their impact on the measure of crash risk. It was shown that at moderate-to-high speeds VSL may be used to improve the safety situation on the freeway in real-time.

17. Key Word	18. Distribution Statement
Real-time traffic, crash risk, simulation, PARAMICS, rainfall, weather, logistic regression, negative binomial models.	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
		109	

DISCLAIMER

"The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the State of Florida Department of Transportation."

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.

This project has been supplemental to the completed project BC-355 #8 and the ongoing project BD-550 #5. Therefore much of the work overlaps between this project and the other two. The main contributions of this project could be summarized in the following four areas:

- Investigating the factors from loop detector data that are significantly associated with crash risk
- Investigating the possibility of using real-time weather information as part of the real-time crash prediction system on I-4
- Identifying the geometric elements of the freeway that could be used in this system
- Exploring and initial analysis of the ITS strategies that could be used to improve the safety situation on I-4 in real-time, namely at this stage was variable speed limits (VSL)

The analysis showed that the coefficient of variation in speed, average occupancy and the standard deviation of volume in the 5 – 10 minutes before crash occurrence are the most significant variables that could lead to crashes on the freeway.

We have obtained detailed rain fall data from 5 weather stations in Central Florida and developed a rain index based on the archived rain data to investigate whether real-time rain data would be needed for implementation.

Investigating the geometric elements that are related to crash occurrence and could be used with association with real-time traffic conditions from loop detector data, indicated that the locations of the ramps are significant.

Finally, VSL were investigated using the micro simulation model (Paramics). This investigation showed that VSL can be used to reduce the crash risk in real-time. However, the analysis showed that this is most effective only in moderate-to-high-speed conditions. Also, we have noticed the possibility that the crash risk would relocate (migrate) to other locations other than the location that we intend to treat. The strategy to use VSL for real-time safety application is still in its infancy and would require more investigation.

TABLE OF CONTENTS

LIST OF ACRONYMS

AIC: Akaike Criterion

CVS: Coefficient of Variation in Speed

CVV: Coefficient of Variation in Volume

GIS: Geographical Information Systems

PCA: Principal Component Analysis

PDO: Property Damage Only

RCI: Roadway Characteristics Inventory

ROC: Receiver Operating Characteristics

SUNB: Seemingly Unrelated Negative Binomial

CHAPTER 1

STUDY AREA AND DATA PREPARATION

1.1 General

The final goal of this research is to develop a predictive system (excluding driver and vehicle characteristics) for crash occurrence on 36.25-mile 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.

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.

This section provides a brief overview of the data that has been collected as part of this project. For more details, the reader is referred to the final report of the previous project (Abdel-Aty et al., 2004).

1.2 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.

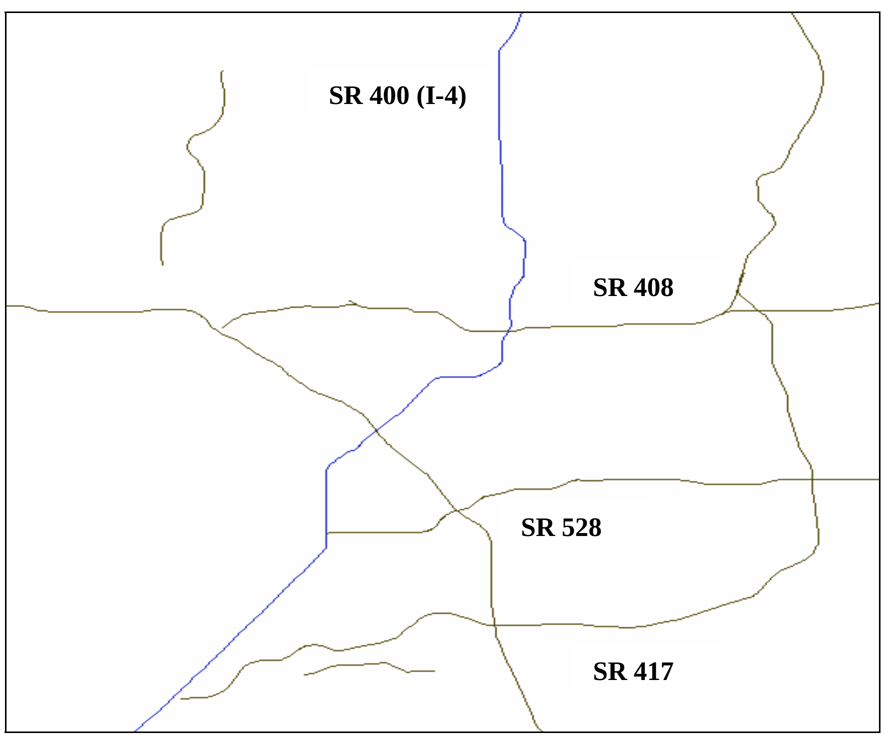


Figure 1-1: I-4 corridor under consideration along with other major roads

Figure 1-1 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 regional traffic management center (*RTMC*). The source of crash and geometric characteristics data for the freeway is *FDOT* (Florida Department of Transportation) intranet server.

1.3 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. The information extracted for each crash case to create a complete crash database for is shown in Table 1-1.

Table 1-1 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* (e.g., rear-end collision, sideswipe collision or vehicle hitting the guard rail). All other fields are self explanatory. The "milepost" field of the crash characteristics table (Table 1-2) was used to determine the loop detector station nearest to location of each crash and was referred to as the station of the crash. In this phase of the project not all available crash characteristics have been analyzed. None the less, they were made part of the database with future research in perspective.

1.4 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.

1.4.1 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 1-3.

Table 1-2 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	27	0	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	27	0	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	|	0	|	|	|	|	|	|	|	|	|	|
04/05/99	|	0	|	|	|	|	|	|	|	|	|	|
04/05/99	33	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/05/99	33	0	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	27	1	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	27	1	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	|	1	|	|	|	|	|	|	|	|	|	|
04/12/99	|	1	|	|	|	|	|	|	|	|	|	|
04/12/99	33	1	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/12/99	33	1	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	27	0	17:30:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	27	0	17:30:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	|	0	|	|	|	|	|	|	|	|	|	|
04/19/99	|	0	|	|	|	|	|	|	|	|	|	|
04/19/99	33	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
04/19/99	33	0	18:05:30	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
|	|	0	|	|	|	|	|	|	|	|	|	|
|	|	0	|	|	|	|	|	|	|	|	|	|
12/27/99	33	0	18:05:00	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx
12/27/99	33	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.

1.5 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 Inventory (*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 1-3. Geometric design of the freeway might differ from one direction to the other, hence the dataset has 138 (*69*2=138*) observations.

Table 1-3 Geometric design of the freeway at loop detector station locations

Loop	**Direction**	**Mile post**	**Radius (ft)**	**Number of Lanes**	**Median type and width (ft)**	**Distance to nearest upstream on ramp (miles)**	**Distance to nearest upstream off ramp (miles)**	**Distance to nearest down stream on ramp (miles)**	**Distance to nearest down stream off ramp (miles)**
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

1.6 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. We have developed a methodology 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.

CHAPTER 2

SIGNIFICANT TRAFFIC FACTORS FROM LOOP DETECTORS

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. For 5-minute level aggregation the arrangement of these time-slices and stations is shown in Figure 2.1. 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. 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.

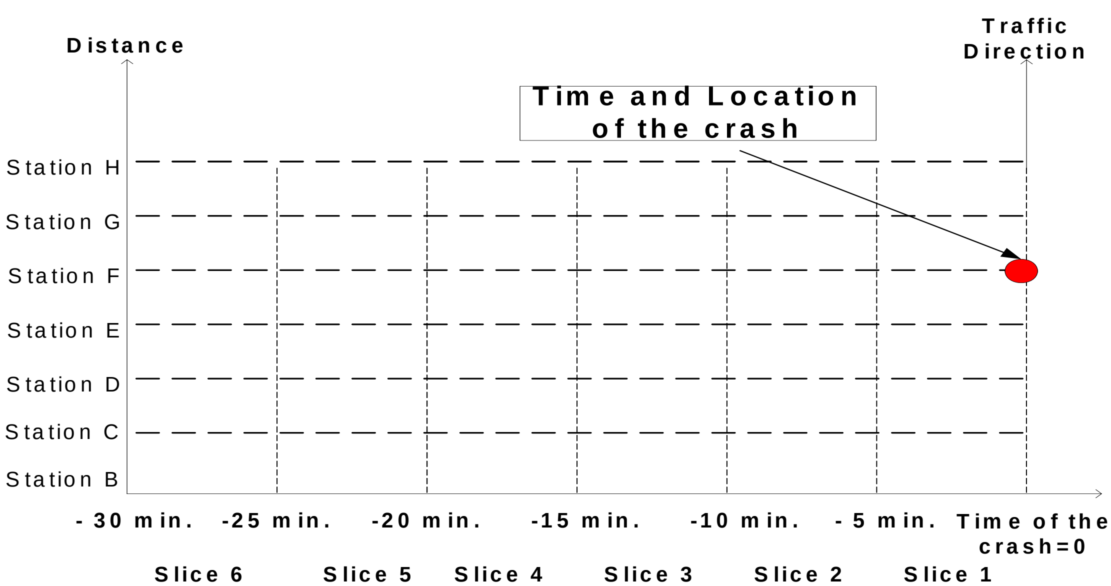


Figure 2-1: Time-space arrangement of all stations and time slices with respect to the crash site and the time of the crash

Using data only from 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 while the model is used for real-time prediction.

2.1 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 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 *3 parameters i.e., LogCVS, AO, SV)* single variable models; one model for each of the three variables *LogCVS, AO* and *SV* over every station *B-H* and the duration of time slice *1-6*. The outcome of these models was the hazard ratio value for these variables at various stations and time slices. 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). 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.

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.

2.2 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. The results show 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 very 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 combined lane data, the same trends in hazard ratio are essentially observed in a time-space framework, although we observed that the values part a little more from unity.

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).

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. 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.

In short, it can be suggested 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 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.

2.3 Spatio-temporal Variation of Crash Risk

As described 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 2.1(a), 2.2(a) and 2.3(a), respectively. Note that in Figure 2.1(a), and 2.2(a) the dark colored region represents high hazard ratios thereby identifying more risk while in Figure 2.3(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 2.1(b), 2.2(b) and 2.3(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 (2.1(a), 2.2(a) and 2.3(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 2.3(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*.

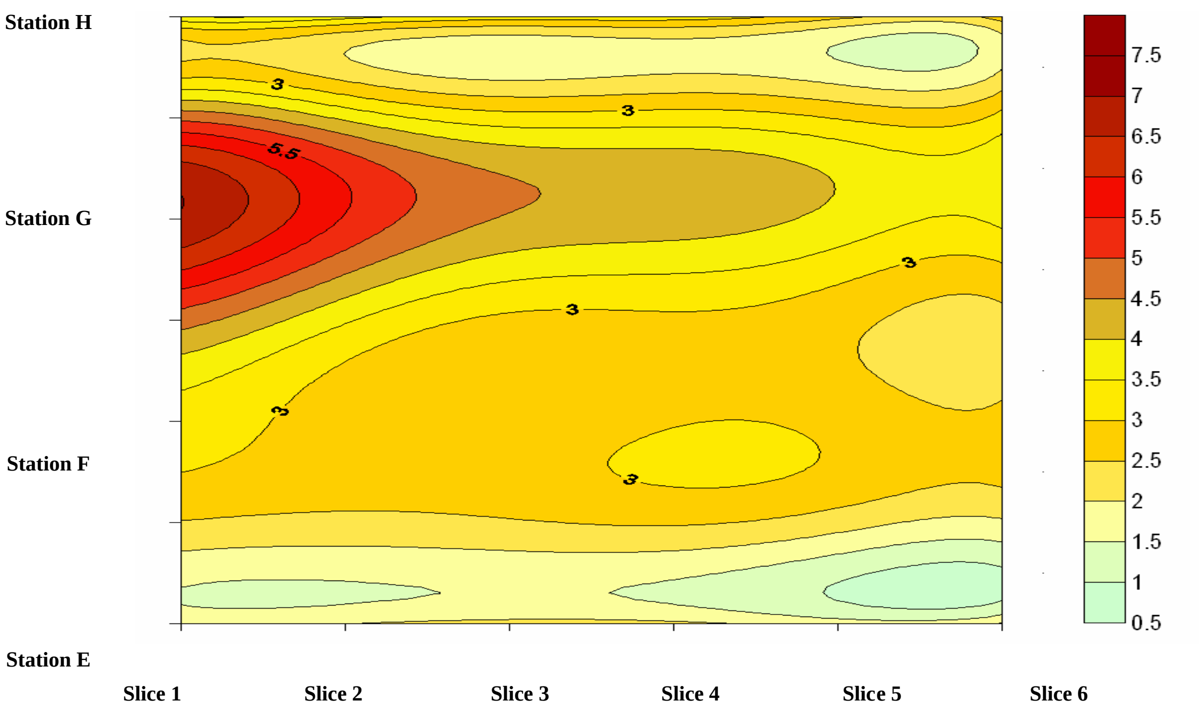
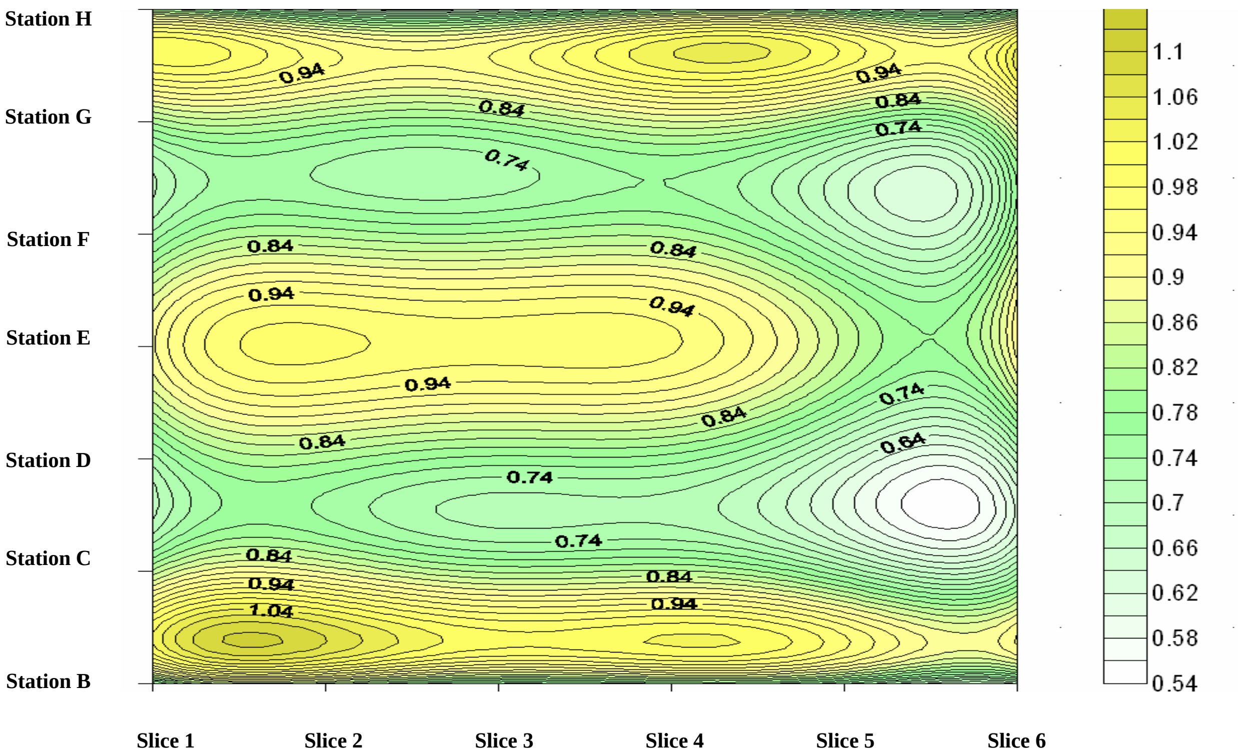


Figure 2.1(a): Spatio-temporal pattern of the hazard ratio for LogCVS obtained from 5-minute combined lane dataset for multi-vehicle crashes



ure 2.1(b): Spatio-temporal pattern of the hazard ratio for LogCVS obtained from "pseudo" crash case dataset

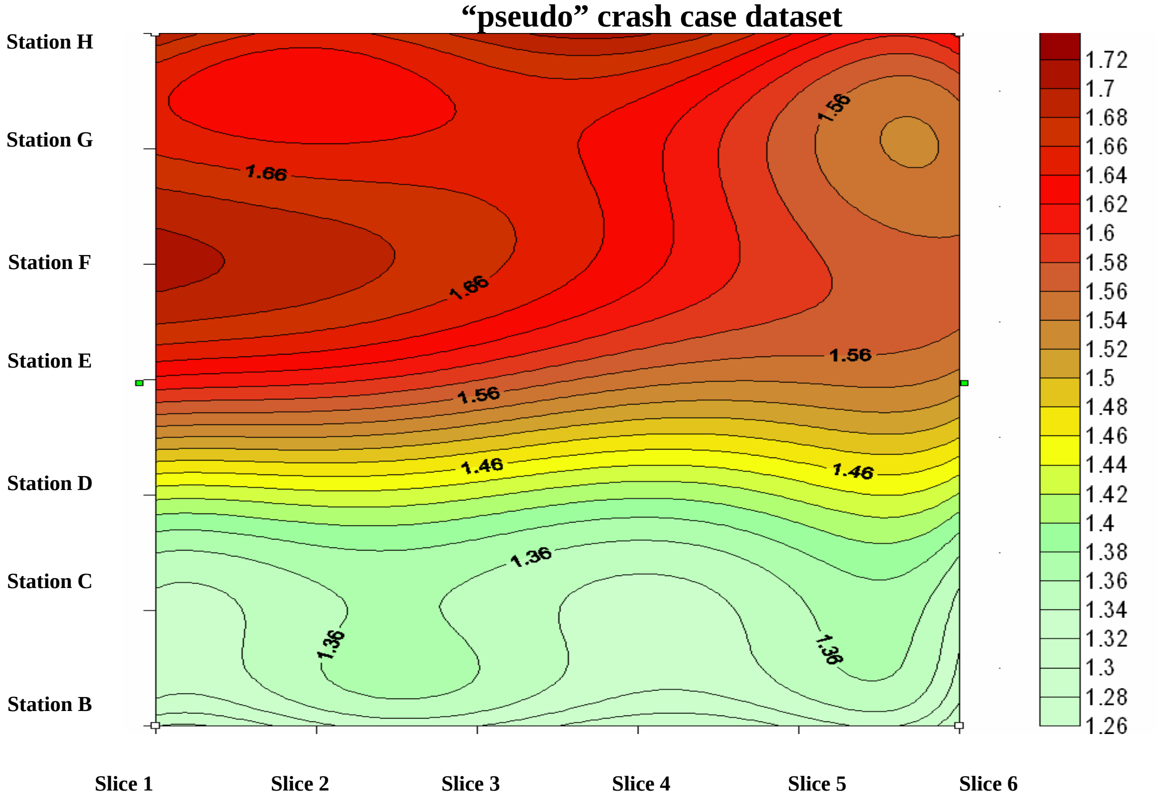


Figure 2.2(a): Spatio-temporal pattern of the modified hazard ratio (increase in crash risk there is ten unit increase in occupancy rather than one) for AO obtained from 5-minute combined lane dataset for multi-vehicle crashes

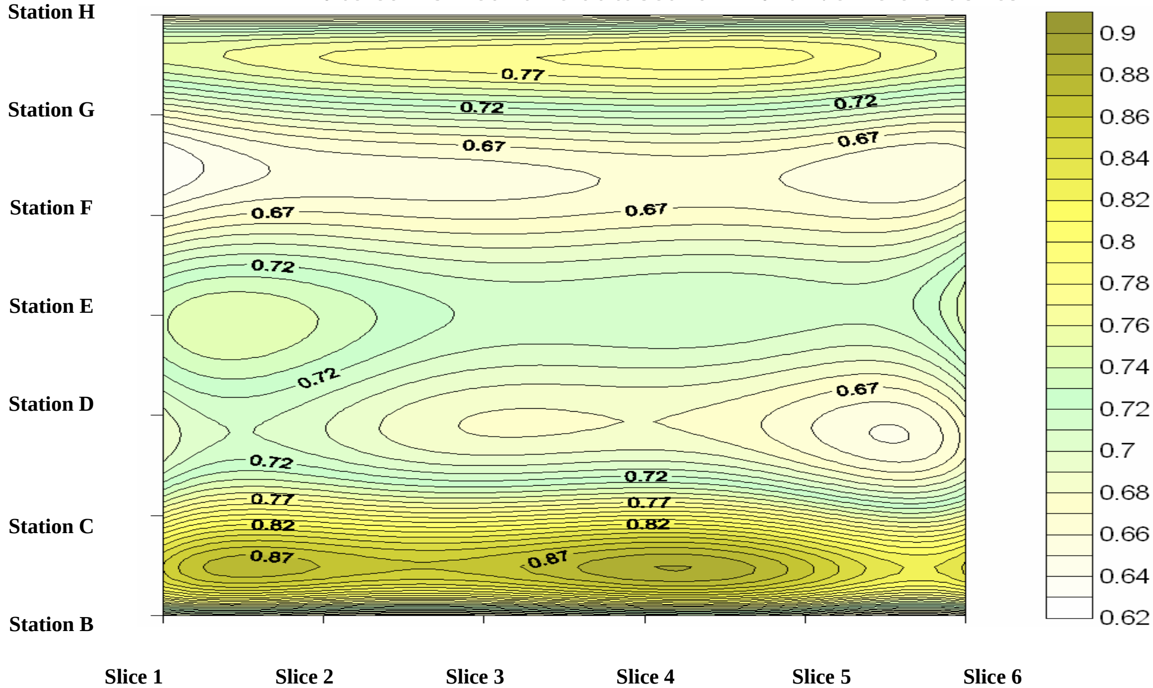


Figure 2(b): Spatio-temporal pattern of the modified hazard ratio (increase in crash hen there is ten unit increase in occupancy rather than one) for AO obtained from "pseudo" crash dataset

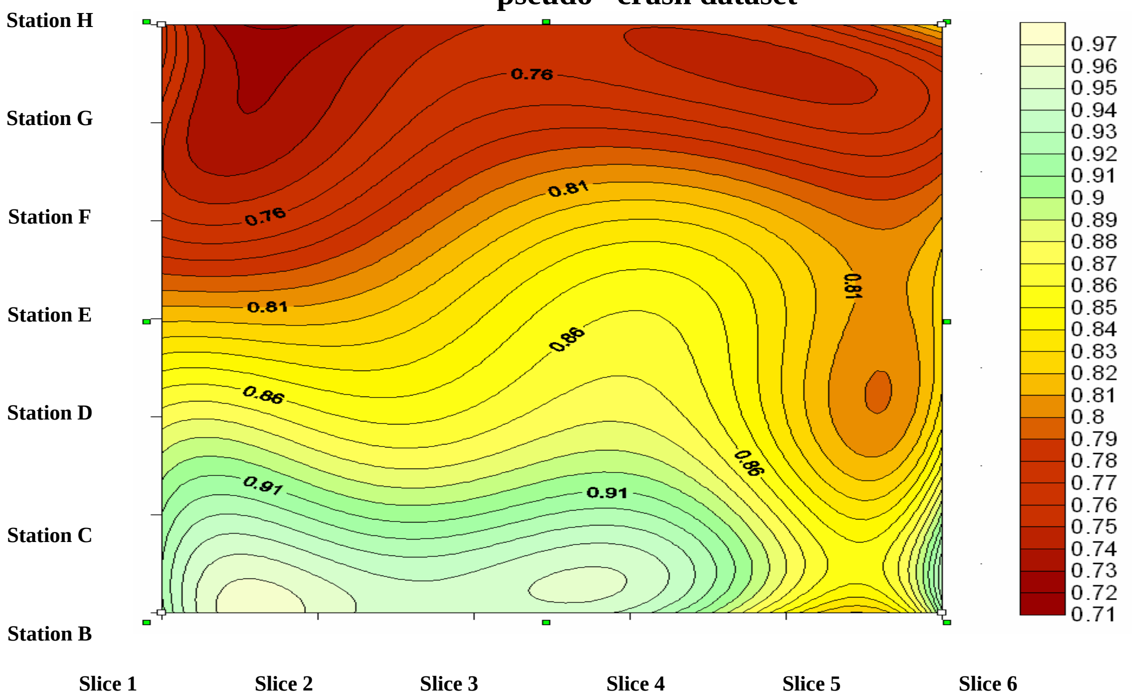


Figure 2.3(a) Spatio-temporal pattern of the hazard ratio for SV obtained from 5-minute combined lane dataset for multi-vehicle crashes

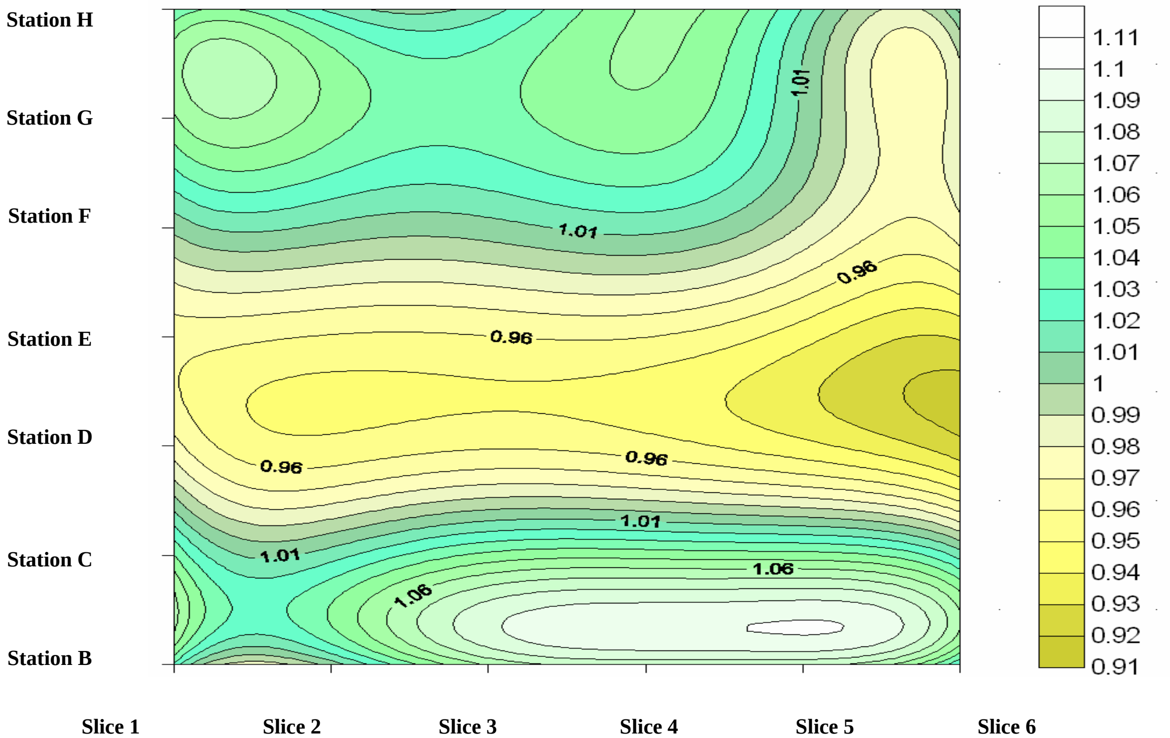


Figure 3(b): Spatio-temporal pattern of the hazard ratio for SV obtained from "pseudo" crash case dataset

2.4 Section Summary

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 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.

CHAPTER 3

REAL-TIME WEATHER INFORMATION

Since among the objective of this study is to identify the traffic and weather related factors (e.g., pavement conditions, visibility etc.) that affect the probability of crash occurrence, we were also interested in non crash situations (i.e., the traffic and rain conditions that do not lead to crashes). The crash police report identifies the weather condition when a crash occurs, however identifying the weather condition in the more than 47,000 non crash cases is also needed.

Along the I-4 corridor there are no weather monitoring stations, which can provide the exact rainfall information at a desired time and location. Alternatively, as mentioned above, the Florida crash database provides the exact weather condition at the time of only crashes on I-4. There is a need to identify rainfall information at a particular time and location on I-4 other than the time of crash occurrences.

The information on rainfall at the time of crash occurrence obtained from Florida crash database is provided in Table 3.1. Out of 1964 crash cases that happened during 1999 through 2001, 217 of them occurred during rain, which adds up to 11 percent of the total number of crashes. This is a significant percentage of crash occurrences during rain situation which explain the need to account for the rainfall condition for crash and non-crash cases which in turn helps to identify the effect of rainfall on crash occurrence.

Table 3.1: Number of crashes that occurred during rain in 1999 – 2001 on I-4

Rainfall occurrence during the crash cases			
Rain Situation	Frequency	Percent	Cumulative Frequency
No Rain	1747	88.95	1747
Rain	217	11.05	1964

Various agencies were contacted to obtain rainfall information. The main aim was to obtain rainfall information for I-4 at a desired time and location. Among the agencies contacted, *Florida Automated Weather Network's (FAWN)* and *National Oceanic and Atmospheric Administration (NOAA)* provided the rainfall data. *FAWN* website provided 15 minute data for two sites on the western side of Orlando. The sites are in Apopka and Avalon. *NOAA* provided access to their database that consisted of hourly rainfall totals. The hourly rainfall information for the weather stations located at Orlando International Airport, Executive Airport and Sanford Airport were obtained from NOAA.

In summary, rain data for five weather stations surrounding I-4 was successfully obtained. Two of them are located on the western side of I-4 and they provided 15 minute rainfall information from 1999 through 2002. The other three stations located on the Eastern side of I-4 provided hourly rainfall data from 1999 through 2002. A map showing the locations of the five weather stations surrounding I-4 is provided in Figure 3.1.

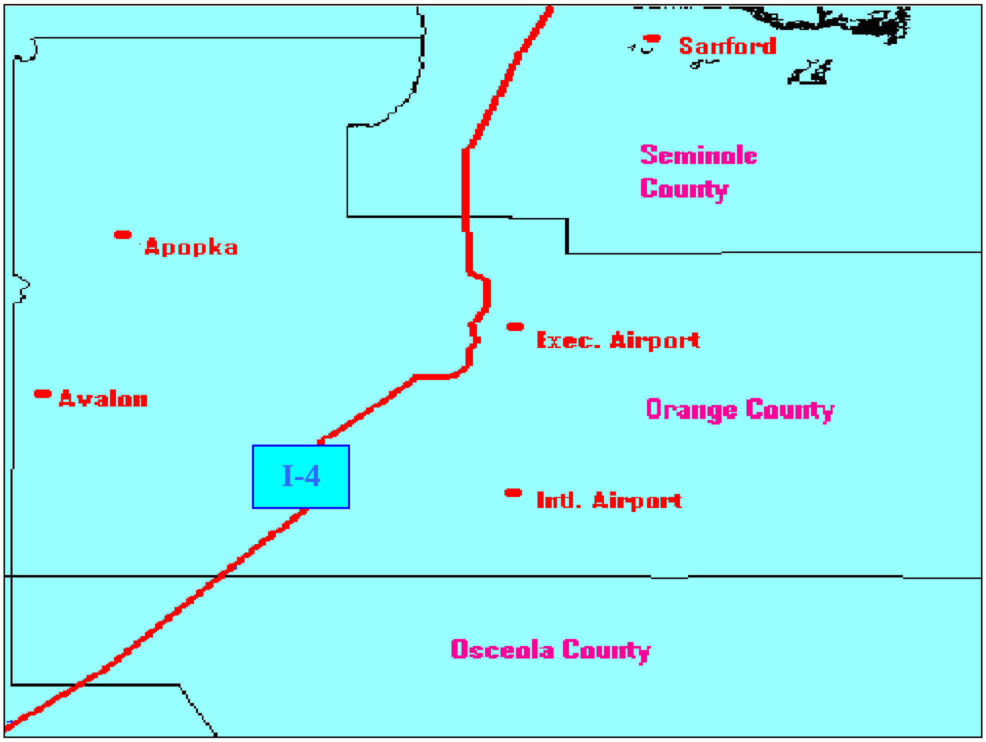


Figure 3.1: Map showing locations of the five weather stations surrounding Interstate 4 in Central Florida

3.1 Weather Model

3.1.1 Methodology and Data Preparation

As a result of not having rainfall information on I-4, logistic regression technique was used to fit a model to the data (crash cases) which uses the rainfall condition available for the crash cases as the response variable and the rainfall data at the same time of crash from the five weather stations situated on both sides of the I-4 corridor as the independent variables. The model developed with the crash cases, was then applied to a new data set (non-crash cases) to obtain the weather condition.

The goal of logistic regression is to identify the best fitting model that describes the relationship between a binary dependent variable (in general y=0 and y=1) and a set of independent variables. The dependent variable in the case of logistic regression is the probability (P) that the resulting outcome is equal to 1. So the model can be expressed as

$Y = \text{Logit}(P) = \text{Ln}\ \{P_i / 1 - P_i\} = \beta_0 + \beta_i X_i$, $i = 1,\ldots\ldots\ldots,n$ for a set of n independent variables.

So P_i can be written as

$P_i = \exp(\beta_0 + \beta_i X_i) / 1 + \exp(\beta_0 + \beta_i X_i)$

Where the logit is the log (to base e) of the odds that the dependent variable is 1, β_0 is the model constant and the β_i are the parameter estimates for the explanatory variables.

In this study the weather information provided by the Florida Crash Database is taken as the binary dependent variable and the rainfall information from the five weather stations surrounding I-4 are the independent or explanatory variables.

3.1.2 Dependent Variable

In the study area, a total of 1964 crashs were taken from the Crash Database for the years 1999 through 2001. Out of the three years, data from 1999 and 2000 (1296 crash cases) was used to calibrate the model and the year 2001 (668 crash cases) was used to evaluate the model. For each of the crash cases, the time, date and location of the crash and the weather condition are obtained. The study area has 69 dual loop detectors installed on a 36.25-mile stretch numbered from 2 to 71. For each crash case, the nearest loop station is identified as the crash location (station F). A sample of the information prepared as explained in above paragraph is provided in Table 3.2. The response variable with y = 1, when it rained and y = 0, otherwise. The time, date and location of the crashes are used in preparing the independent variables.

Table 3.2: Sample weather information extracted from the crash database

Serial No	Time of Crash	Loop Detector Station Nearest to the Crash Location (Station of the Crash)	Date of Crash	Weather condition
1	9:02:00	47	4/1/1999	CLEAR
2	8:50:00	49	4/1/1999	CLEAR
3	0:10:00	43	4/1/1999	CLEAR
4	16:45:00	42	4/1/1999	CLOUDY
5	14:45:00	34	4/1/1999	CLOUDY
6	17:15:00	59	4/2/1999	CLEAR
7	16:48:00	69	4/2/1999	CLEAR
8	15:30:00	11	4/6/1999	CLEAR
9	15:47:00	30	4/28/1999	RAIN
10	19:07:00	36	4/28/1999	RAIN

3.1.3 Independent Variables

For each crash case, rainfall information from each of the five weather stations is entered as the independent variables in the model at the same time as that of the crash occurrence. To relate the response variable with the independent variables in space also, an order for the independent variables is obtained based on the distance between a particular I-4 crash station and a weather station. Table 3.3 provides a sample of independent variables entered in the model.

Table 3.3: Sample information with dependent and independent variables used in the model

Time	station	Date	Weather	Rain_1*	Rain_2*	Rain_3*	Rain_4*	Rain_5*
6:15:00	43	10/15/2001	0	0	0	0	0	0
15:36:00	37	10/19/2001	1	0	0	0	0	0.01
12:39:00	49	10/19/2001	0	0.0001	0	0.0001	0.01	0
16:35:00	26	10/19/2001	0	0	0	0	0	0.02
14:10:00	60	10/19/2001	1	0	0.0001	0.17	0.0001	0
23:29:00	20	10/19/2001	0	0	0	0	0	0
17:15:00	9	10/19/2001	0	0	0	0	0	0
12:30:00	42	10/19/2001	0	0.0001	0.0001	0	0.01	0
16:03:00	26	10/19/2001	0	0	0	0	0	0.02
15:18:00	33	10/19/2001	0	0	0	0	0	0.01
18:20:00	4	10/21/2001	1	0.02	0.05	0.01	0	0.01
18:41:00	53	10/21/2001	1	0.01	0	0.01	0.05	0
3:58:00	10	10/22/2001	1	0	0.03	0	0	0
19:13:00	53	10/22/2001	0	0	0	0	0	0

* The units for rain_1 through rain_5 are inches/hour

- In Table 3.3, weather is the response variable with outcome of “1” when raining and “0” when not raining, and rain_1 – rain_5 are the independent variables with hourly rainfall information. Rain_1 contains the rain information at the nearest weather station from the corresponding crash station at the time and date of the crash. Rain_2 contains rain information at the second nearest weather station and so on. For example, the first independent variable for the crash that happened on 10/21/2001 at time 18:20:00 and at station 4 has 0.02 inches of rainfall and is the nearest weather station from crash station 4. Therefore, rain_1 to rain_5 are dynamic factors and change from one station to another on I-4 depending on its proximity to the weather stations. Therefore, the geographical co-ordinates for all the 69 I-4 stations and the five weather stations were obtained. Based on these co-ordinates, the distance between any I-4 station and each of the weather stations is obtained. A table is prepared which provides information on the order in which the weather stations are situated from each crash station based on distance. The nearest weather station is placed first, the second nearest second and so on. Tables were prepared for each of the five weather stations separately for each year (1999 – 2001), consisting of rainfall information. A sample of rain values at the Avalon station for the year 1999 is provided in Table 4.

Table 3.4: Sample Rainfall Information at the Avalon Weather Station

Time	Rainfall (inches/hour)
1/1/1999 10:00	0
1/1/1999 10:15	0.1
1/1/1999 10:30	0
1/1/1999 10:45	0.02
1/1/1999 11:00	0
1/1/1999 11:15	0
1/1/1999 11:30	0
1/1/1999 11:45	0
1/1/1999 12:00	0
1/1/1999 12:15	1.05
1/1/1999 12:30	0.08
1/1/1999 12:45	0.02
1/1/1999 13:00	0
1/1/1999 13:15	0

Using the information in Tables 3.2 and 3.4, the rain values are entered in Table 3.3 using a program developed in Visual Basic. For example, let us take a crash that happened on 10/21/2001 at time 18:20:00 and at station 4. We first retrieve the order of the weather stations located from the crash station 4. So the nearest weather station from station 4 is Avalon. Then we go the Avalon weather station table with rainfall values for 2001 and take the rain value for 20th October at time 18:20:00 and put this value in Table 3.3 for the rain_1.

3.1.4 Model Development

Once the response and independent variables are obtained, the next step would be to apply the logistic regression model. As stated earlier, the data from 1999 and 2000 which had 1296 crash cases was used to build the model. But it is probable that if it rains in one of the weather stations, it might also rain in the other stations, thereby making variables rain_1 - rain_5 correlated and violating the assumption of independence, which in turn reduces the efficiency of the model with erroneous parameter estimates. A chi-square test was conducted to check the independence of these variables. The results of this test shows that at a 95% confidence level, the test statistic and the p-value are provided 84.326 and 0.000, respectively. Since, the null hypothesis was rejected, the variables cannot be considered as independent.

To deal with the issue of non-independence, i.e., an approach to remove the redundancy in these variables, "principal component analysis" technique was applied to the variables before the regression analysis. Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The mathematical technique used in PCA is called Eigen analysis: we solve for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products, which in general called as the covariance matrix. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the covariance matrix and the total information provided by the original variables can be expressed as this trace.

So essentially by looking at each of eigen values, the percentage information provided by each of the principal components can be obtained.

In order to decide upon the number of principal components that are to be used as input (independent variables) to the logistic regression model, three rules are applied: 1) 80% rule: The minimum number of principal components to be used in the model has to retain at least 80% of the total information, 2) Average Eigen Value rule: All those principal components whose Eigen values are lesser than the average are to be excluded, and 3) Scree plot: It is the plot of Eigen values Vs the number the Eigen values. Exclude those principal components on the flat part of the curve, i.e., scree plot and retain those on the steep part.

The results of the PCA procedure are provided in Table 3.5 and 3.6, and Figure 3.3. Table 3.5 presents the covariance matrix of the independent variables from which the eigen values and eigen vectors are calculated. Table 3.6 provides the eigen values of the covariance matrix. Using these results, the number of principal components to be retained is determined. For rule 1, in Table 3.6, the shaded part under "cumulative" is around 90%. So 4 principal components are able retain at least 80% of the information. For rule 2, in Table 3.6, the average of Eigen values is 0.00830027 and only 2 Eigen values exceed this value. Thus two principal components have to be retained. For rule 3, looking at Figure 3.2 and retaining the eigen values on the steep part of the curve, four principal components can be retained. Since, two out of three rules say that four principal components can be retained, therefore, the first four principal components are used as the independent variables in the logistic regression model.

Table 3.5: Principal Component Analysis

Covariance Matrix						
		rain_1	**rain_2**	**rain_3**	**Rain_4**	**rain_5**
rain_1	**Rain_1**	0.00494	0.0011	0.00075	0.00065	0.00076
rain_2	**Rain_2**	0.0011	0.01091	0.00062	0.0005	0.001
rain_3	**Rain_3**	0.00075	0.00062	0.0141	0.00125	0.00158
rain_4	**Rain_4**	0.00065	0.0005	0.00125	0.00715	0.0005
rain_5	**Rain_5**	0.00076	0.001	0.00158	0.0005	0.00441

Total Variance	0.0415

Table 3.6: Eigenvalues from Principal Component Analysis

Eigenvalues of the Covariance Matrix Total = 0.04150134 Average = 0.00830027				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	0.01492	0.00389	0.3595	0.3595
2	0.01102	0.00405	0.2656	0.6251
3	0.00698	0.00217	0.1681	0.7932
4	0.00481	0.00103	0.1158	0.909
5	0.00378		0.091	1

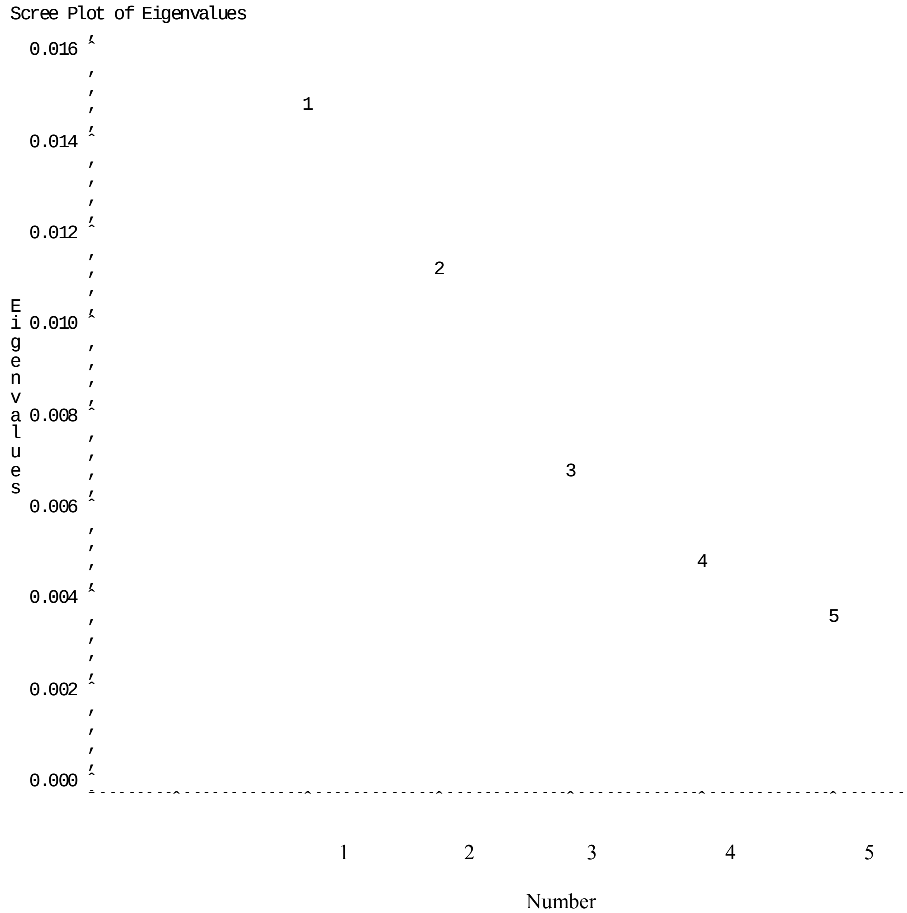


Figure 3.2: Scree Plot from Principal Component Analysis

With the four retained principal components of the variables rain_1 through rain_5, a simple logistic regression model was estimated. The parameter estimates of the four principal components used in the model are provided in Table 3.7. The model fit statistics of the logistic regression model as the Akaike Criterion value (AIC: the lower the better) and the log-likelihood value. The AIC value can be used to see if the regression technique chosen, works for the variables used. The low AIC value under the “intercept and covariates” heading, when compared with value under “covariates only” heading, proves the fact that logistic regression model is indeed a good fit for the variables. The same conclusion can be drawn from the log-likelihood values with a log-likelihood ratio test. The four principal components are significant at 95% confidence level. Also it can be seen that the first principal component is highly significant which confirms the fact that it contains a large portion of the total information. The model can be expressed as:

Probability that the outcome =1, i.e., it rained or "rain index"= $e^{(-2.1444 + 3.3260*\text{Principal component 1} + 1.2645*\text{ Principal component 2} + 1.5221*\text{ Principal component 3} + 2.3441*\text{ Principal component 4})} / 1 + e^{(-2.1444 + 3.3260*\text{Principal component 1} + 1.2645*\text{ Principal component 2} + 1.5221*\text{ Principal component 3} + 2.3441*\text{ Principal component 4})}$

This model provides the probability of rainfall at a given time and location on I-4.

Table 3.7: Logistic Regression Model Results

Parameter	Degree of Freedom	Estimate	Standard Error	P-value for the Test of Significance of the Parameter Estimate
Intercept	1	-2.1444	0.0925	<.0001
Principal component 1	1	3.3260	0.5910	<.0001
Principal component 2	1	1.2645	0.5834	0.0302
Principal component 3	1	1.5221	0.7354	0.0385
Principal component 4	1	2.3441	1.0342	0.0234

Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	910.324	864.188
-2 Log Likelihood	908.324	854.188

3.1.5 Rain Model Evaluation

As noted before, the year 2001 data was used to evaluate the model. The SAS "score" procedure was used for the purpose. This data set has 668 crash cases and is referred to as "validation data set". In model evaluation; the estimates from the model built with the data from 1999 & 2000 are applied to the validation data set to get the probability of rainfall. This probability is referred to here as "rain index" value. To know the prediction accuracy of the model which is applied to the validation data set, a cut-off was set above the 75th percentile (0.0985602) of the rain index values. The crash cases which have rain index values greater than 0.0985602 are assumed to have occurred during rain, i.e., predicted outcome. The Quantiles for the rain index values is provided in Table 3.8. The overall prediction accuracy for the model is 88.02%. The prediction accuracy for the cases with "rain" is 77.78% and the prediction accuracy for the cases with "no rain" is 89.26%. Since, the overall prediction accuracy of the model is high, therefore the model can be used for obtaining the rain index values for a desired time, date and station on the 36.25-mile stretch of Interstate 4 in central Florida for the non-crash cases. The logistic regression model was applied to the database to obtain the rain index values for 2035 crash (where loop detector data were also available) and 47612 random non-crash cases.

Table 3.8: Quantiles for the rain index values

Quantiles	
Quantile	Estimate
100% Max	0.9849514
99%	0.5901525
95%	0.1470778
90%	0.1032220
75% Q3	0.0985602
50% Median	0.0985602
25% Q1	0.0985602
10%	0.0985602
5%	0.0985446
1%	0.0910038
0% Min	0.0187801

3.2 Crash Prediction Model

Initial analysis showed that three parameters, namely, *LogCVS* (standard deviation/mean speed), *SV* (standard deviation of volume) and *AO* (average occupancy) 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 preliminary results we can discard *Station B, C* and *D* (since they are less significant than traffic measurements from *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.

Also, even though time duration 1 (0-5 minutes) prior to crash exhibits significant hazard ratios, 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 *12* traffic flow variables; *LogCVS, SV,* and *AO* at each of the four loop detectors *E, F, G* and *H*. Using the logistic regression technique, 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), were retained in the model. All other variables are found to be statistically insignificant. Thus the final model includes variables *LogCVSF2* and *AOG2* and *SVG2*. The details of the final predictive model are provided in Table 3.9. First two variables have positive beta coefficients, which mean that the likelihood of a crash increase as these variables increase while *SVG2* had negative beta coefficient implying decreasing likelihood 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.

To simplify the process, the rain index values were directly used in the safety model, instead of setting a cut-off value and then determining how many cases occurred during rain. The 'rain index', a measure for the probability of having rain, was defined on Page 38. This can be justified because the rain index values are continuous and indicate the probability of rainfall at a particular location. Also setting a cut-off value may force some cases to have a "rain" situation when it is actually a "no-rain" case and vice-versa. This might undermine the actual effect of rainfall in the safety analysis. The rain index does also indicate a measure of intensity of rain which might show a visibility problem in addition to slippery pavement situation. Table 3.10 depicts the coefficient estimates when the Rain Index is entered in the model.

Table 3.9: Parameter estimates of the crash prediction model without the "Rain Index"

Parameter	Estimate	Standard Error	Pr > ChiSq
Intercept	-3.4388	0.1308	<.0001
AOG2	0.00964	0.00307	0.0017
SVG2	-0.1299	0.025	<.0001
LOGCVSF2	0.5366	0.0979	<.0001
AIC	11994.87		
-2LogL	11986.87		
Pearson Chi-square	35906.0802		

Table 3.10: Parameter estimates of the crash prediction model with the "Rain Index"

Parameter	Estimate	Standard Error	Pr > ChiSq
Intercept	-3.5853	0.1344	<.0001
AOG2	0.00891	0.00308	0.0038
SVG2	-0.1284	0.025	<.0001
LOGCVSF2	0.5333	0.0979	<.0001
Rain Index	1.3924	0.267	<.0001
AIC	11951.37		
-2Log Likelihood	11941.37		
Pearson Chi-square	35679.7239		

When comparing the AIC values of two models, the model with the lower AIC value is chosen over the other model. In this case the model with the "Rain Index" variable has an AIC of 11951.37 which is lower than the AIC value of 11994.87 for the model without it. Also, the Log-likelihood ratio test indicates that the model that includes the Rain Index is better (G = -11941.37 – (-11986.87) = 45.5). The lower Pearson Chi-square of the model with the rain index also indicates a better statistical fit.

Receiver operating characteristic curve which is widely known as the ROC curve, originated from signal detection theory, shows how the receiver operates the existence of signal in the presence of noise. It plots the probability of detecting true signal and the false signal for an entire range of possible cut-off points. The area under the ROC curve, ranging from zero to one, provides a measure of the model's ability to differentiate between those cases which are crashes versus those which are not. The higher the value for the area under the ROC curve, better the prediction accuracy. SAS reports four measures of association between the predicted probabilities and observed responses. The measures lie between 0 and 1, with large values suggesting a strong association. These associations are provided in Table 3.11 for the model without and with "Rain Index".

Although the associations are not of interest in the analysis, the measure of association "C" is actually the area under ROC curve. Looking at Tables 3.11, the higher "C" value for the model with "Rain Index" indicates better prediction accuracy for this model. It may be observed, however, that the difference in the "C" values for the two models is only 0.009. Marginal difference in performance measures for the two models indicates that even the best efforts to account for rain fall information do not significantly improve the performance of the crash 'prediction' model. Possible reason for that could be that the parameters (speed, volume and occupancy) measured at loop detectors are able to capture the impact of the wet weather on the traffic characteristics and as it pertains to crash occurrence.

Table 3.11: Measures of association between the predicted probabilities and observed responses for the models

	Without Rain Index	With Rain Index
Somers' D	0.163	0.178
Gamma	0.173	0.189
Tau-a	0.013	0.014
C	0.581	0.589

3.3 Section Summary

The developed rainfall index has a positive impact on the probability of crashes. The final model developed could be used to calculate the probability of observing a crash vs. not. A threshold value for this probability may then be set in order to determine whether the location has to be flagged as a potential “crash location”. On-line traffic and rainfall data could be used for real-time crash prediction. However, our contacts with researchers at the University of North Florida who are conducting a research project sponsored by the I-Florida program at district 5 to use the real-time GIS analysis to study the impact of rainfall rates on highways in north east Florida, revealed that the last rain station on I-4 would be around the St Johns’ river bridge which is outside our study area in Central Florida. This might make the use of real-time weather information in a real-time crash prediction framework unrealistic, at least in the current scenario. Moreover, it was noticed that the difference in the classification performance of the models with and without the rain index (the measure developed here to represent the weather conditions on the freeway) was marginal. Therefore, while model(s) that include the rain index, derived in this study, could be used once weather stations are available to record (and report) the real-time rainfall information for the study corridor, we do not expect significant improvement in classification performance of the models.

CHAPTER 4

STUDYING THE GEOMETRIC ELEMENTS

Geometric and roadway factors considered for all the 138 loop detector stations in both directions include; radius of the freeway section, median type, median width, pavement condition, pavement surface type, pavement roughness index, and the presence of off or on-ramps within the influence area of each crash station. The influence area of a crash station or loop station was taken as the sum of half the distances between that loop and the loops on each side. Graphical description of the influence segment for a loop detector station (for instance, station 6) is provided in Figure 4.1.

Figure 4.1 also provides a visual representation of the instrumented Interstate 4 with installed dual loop detectors in the East and West directions. Other factors such as the shoulder width, shoulder type, etc was not considered as there was no variability in these factors along the study section of I-4.

As indicated previously, loop detector data were available for the years 1999 through 2002. It would be practically infeasible to take speeds and volumes for the entire year to represent the traffic characteristics of a crash station. So a typical month in the year 2002 was chosen for that purpose, the latest among the years during which the data was collected. In this year, all Tuesdays, Wednesdays and Thursdays in the month of February were chosen to get the desired traffic factors. To remove the abnormal traffic pattern caused by the crashes that have occurred during these days, loop detector data one hour before and one hour after the crash occurrence was discarded. Now loop data for the remaining time during all these days was combined at every crash station. The raw data obtained from the loop detector stations was for 30 second intervals. Since, 30 second data is a short interval data and due to the possibility that no visible traffic pattern can be captured during this interval, loop data was aggregated to 5, 10, and 15 minute intervals. Data was aggregated to a maximum of 15 minute interval to keep the aggregation level as microscopic as possible.

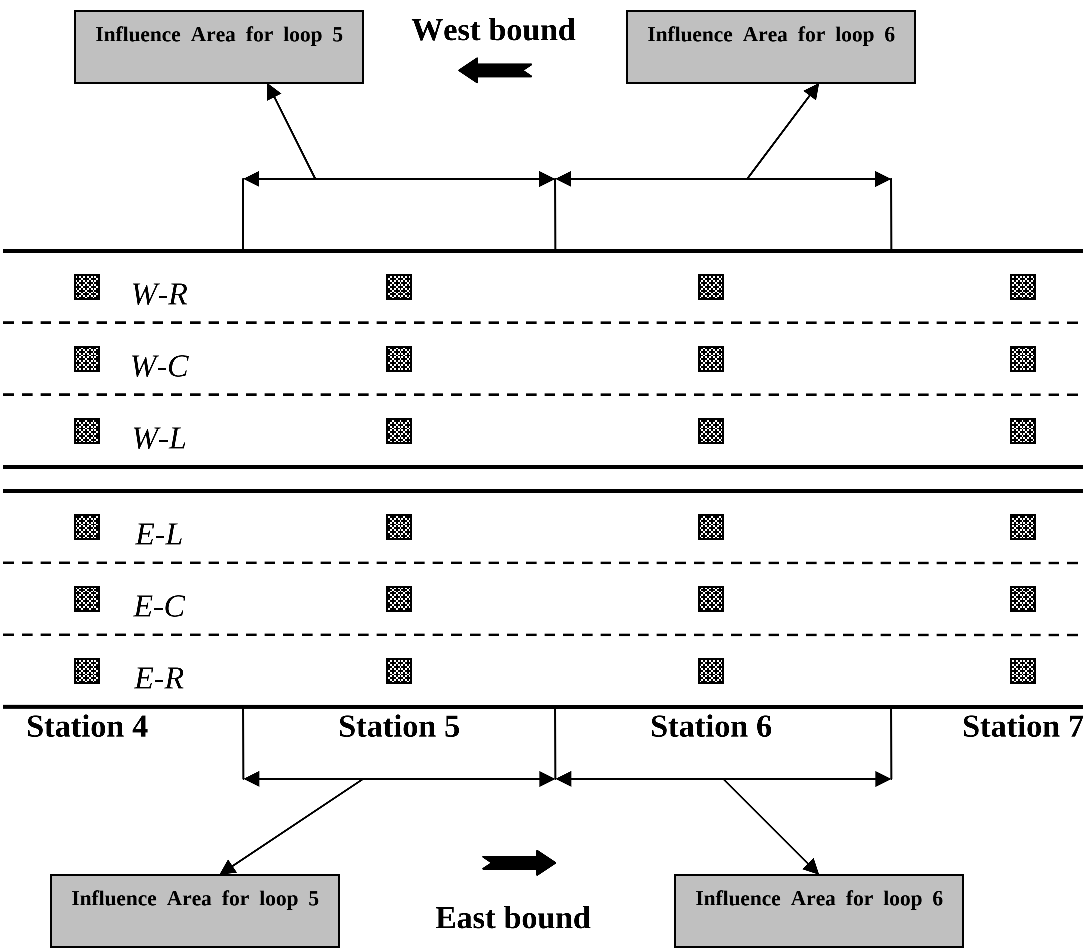


Figure 4.1: Influence segment for each loop detector station

Volumes at 30 second level for each crash station were taken for the period specified in the aforementioned paragraph, and aggregate fifteen minute volumes were calculated. The maximum of these aggregate fifteen minute volumes at each crash station was taken as the peak fifteen minute volume for that station. This was considered as the peak volume at each crash station. Fifteen minute peak volumes were considered to capture the effect of actual peaking condition on crash occurrence. It is important to note that the peak fifteen minute volumes were not taken specifically for the morning or evening peak periods, and the whole day was considered for both directions.

The following factors were considered for the study, which used the loop data as explained in the above paragraphs.

4.1.1.1.1 Average speed

The raw data from the loop detectors is obtained for an interval of 30 seconds. This data was aggregated to 5, 10, and 15 minute intervals, and average speed across all the lanes were taken to represent a particular crash station. Since traffic factors (speed and volume factors) in one lane are correlated with the factors in the other lanes, the average across all the lanes was calculated. The 75^{th} percentile of average speed values at every station is taken as the variable for consideration in the models. There can always be a question on how to decide what percentile would actually represent a particular traffic factor at a crash station. The most logical explanation could be as follows: If we take the 50^{th} percentile, we might actually under represent the traffic at the station, since there is 50^{th} percentile vehicle population (not always true) exceeding the value considered. If we take 90^{th} percentile, we might over represent the traffic, since the vehicles do not travel at such high speeds always. To statistically prove this fact, all the three percentiles were tried and it was found that there was no significant difference among the three percentiles. Hence, based on the above discussion it was decided to use the 75th percentile in the analysis.

4.1.1.1.2 Standard deviation of speed and volume

Raw data was aggregated for 5, 10, and 15 minute intervals and standard deviation of speeds and volumes was taken. As in the case of average speed, standard deviation was also taken across all lanes in one direction. The 75th percentile of standard deviation values at every station is taken as the variable for consideration in the model.

4.1.1.1.3 Coefficient of Variation of speed and volume

Coefficient of variation can be seen as a measure of deviation of the selected variable from its mean. It is defined as:

Coefficient of variation = Standard deviation/ expected mean

Standard deviation and average speed and volume for 5, 10 and 15 minute intervals were used to obtain the coefficient of variation for these factors. Again 75th percentile of these values was used in the models (the 75th percentile value taken here is after calculating the coefficient of variation first, and then obtaining its 75th percentile).

4.2 Seemingly Unrelated (SUR) Negative Binomial Modeling Approach

The crash frequency models in the study were developed using negative binomial regression, due to the fact that Poisson regression cannot account for the over-dispersion in the data. Then seemingly unrelated negative binomial models were developed for different crash categories using traffic, roadway and geometric characteristics. The Negative Binomial model has the following form:

$$\lambda_i = EXP\ (\beta X_i + \varepsilon_i)$$

Where λ_i is the expected number of crashes per period at location *i*, X_i is the vector of explanatory variables, β is the vector of estimable parameters, and $EXP\ (\varepsilon_i)$ is a gamma distributed error term with mean 1 and variance α^2. The addition of this error term allows the variance to differ from the mean in the following way:

$$VAR[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha \{E[y_i]\}^2$$

Where $VAR[y_i]$ is the variance and $E[y_i]$ is the mean of the model distribution

Every model is associated with an error term which can be related to many factors. In case of models developed in road safety field, two types of error terms are correlated: omitted variables and measurement errors. Omitted variables may be unintentionally or intentionally excluded mainly due to data unavailability. Also it is impractical to assume that each and every variable affecting crashes to be included in crash models. Measurement errors are the most common components of error terms since there always exists unreliability in the measurement of variables. For instance, inaccurate computation of AADT or any other traffic variable is a measurement error. SUR models come into the picture when we deal with a system of equations where error terms are correlated across the equations. The effects of omitted variables are carried to the error terms of each model. When estimating various crash types (for example multiple or single vehicle crashes), it is likely that error terms (mostly the omitted variables) across these two models will be correlated. Unlike simultaneous models, seemingly unrelated regression deals with a set of equations not because they interact, but because the error terms are related. Let us assume that the effect of omitted variables is represented by the term ϕ, and is consigned to the new combined error term χ, as shown in the following equations.

$$\lambda_i = Exp(\beta_i + \varepsilon_i + \phi_i)$$

$$\lambda_i = Exp(\beta_i + \chi_i)$$

It was assumed that the original error term ε is not related to the existing variables and includes general random error terms like measurement errors.

Two decisive factors were used to keep different variables in the models: 1) A p-value less than 0.1 for the coefficient of estimated variable corresponding to 90% confidence level, and 2) the magnitude and sign of the coefficient of estimated variable is in agreement with the expected or theoretical sign for these factors. For the best model selection between two models, Akaike Information Criteria (AIC) was applied. The best model is decided by the lowest value of AIC. AIC for a model is defined as:

$AIC = -2\ Log\ (L) + 2K$

Where

$Log\ (L)$ is the log likelihood of the estimated model and

K is the number of estimated parameters.

4.2.1 Development of SUR Models using aML Software

aML Software was used which had the capability to solve the seemingly unrelated negative binomial models. aML uses the iterative process, Gauss-Newton likelihood maximization algorithm to obtain the model convergence. More about this algorithm can be found in the user's guide and manual of aML software (Lillard and Panis, 2003). The approach used to solve the SUR models in aML needs some explanation (Lillard and Panis, 2003). Negative binomial models, in plain form, do not feature an explicit residual. That doesn't mean that there is no stochasticity; the model is parameterized as a probability statement, and the residual is implicit in deviations from the predicted probabilities. To capture the correlation of disturbance terms across sets of equations in SUR modeling, an explicit residual term has to be added in individual models. Thus, there is both an implicit and an explicit residual in the individual negative binomial models. Precise identification of both these residuals can be facilitated by making available two or more outcomes per observation. Essentially multiple outcomes contain information about the extent to which a particular observation is different from other observations, so that the explicit residual is identified. So the crash data which was initially combined for four years at each crash station was divided based on year at each station. This would make observations for each of the four years with the same crash station highly correlated. But during modeling the records for a particular crash station for 1999, 2000, 2001, and 2002 are all part of the same group and given the same identification number. Unobserved factors could also be related to both dependent variables (e.g., peak and off peak or single and multiple crash frequency). In aML while modeling interrelated equations, the correlation will be strongly identified once you tie all records pertaining to a particular crash station together via a common identification number. In the process of seemingly unrelated negative binomial (SUNB) estimation, the aML software provides dispersion factors, standard deviation, and correlation coefficient for disturbance terms. The present analysis deals with SUNB estimation of two models at a time, and so the correlation matrix for the disturbance terms, has the following form:

Correlation matrix: $R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$

Where ρ is the correlation coefficient for the error terms, and defined as {[COVARIANCE (U1, U2)]/ [σ_1 σ_2]}. U1 is the error term representation for the 1st model and U2 for the 2nd model. σ_1 and σ_2 are the standard deviation values for the first and second model, respectively. The present study represents σ_1 and σ_2, as SIGMA_U1 and SIGMA_U2, ρ as RHO_U1U2, for the disturbances terms U1 and U2.

4.3 Model Estimation and Results

As mentioned in the previous section, there are five main crash categories, based on the type of crash, availability of daylight, severity of crash, peak condition, and pavement condition (dry or wet). Before proceeding with the estimation of SUNB models, models for each of the sub-categories in each main category are estimated. The estimation results of the individual models were used to obtain the starting values for SUNB models. The individual models are described first followed by the description of simultaneous models.

Table 4.1 provides an explanation of the various variables included in model development. In Table 4.1, AVGS, STDS, CVS, STDV, and CVV are traffic factors obtained from raw 30 second loop detector data. These variables were taken representing the whole day in case of multiple/single vehicle, dry/wet pavement, and injury/PDO crash models. In case of peak and off-peak period, daytime and dark hour crash models, all the microscopic traffic factors were taken separately for peak and off-peak periods, and day and dark hours, and not considered for the whole day. All these variables have been tried for 5, 10 and 15 minute aggregated intervals.

On the whole there are five different main categories. Each main category has two different sub categories. Thus, there are ten different individual models. Each of these ten models is presented in the following sub-sections. Before going into the specifics of individual models, details of the steps to arrive at existing models are to be discussed. For all the models, a comparison between AADT and VMT was made based on AIC values. In all the models, AADT was found to be significant. So AADT was included in the subsequent model estimation. As was the intent of the study, the use of microscopic or disaggregate traffic measures was evaluated. For this purpose, AADT and PEAKFIFT variables were compared keeping all other variables the same. Here AADT indicates a macroscopic variable while PEAKFIFT indicates a microscopic variable. In the present analysis, PEAKFIFT was not found to be significant, but AADT was found to be significant in most cases. Although PEAKFIFT was found not to significantly affect the crash occurrence, other microscopic traffic factors, e.g., average speed, standard deviation of speed/volume, and coefficient of variation of speed/volume, were found to notably influence the crash occurrence at the 90% confidence level. These statistical measures corresponding to a single factor, (for instance vehicle speed which was tried for 5, 10, and 15 minute aggregation levels) as expected will be highly correlated when used simultaneously in the model estimation. Hence, these factors were used separately in the models and the best among various models was selected based on the lowest AIC value.

Table 4.1: Code Sheet for all the variables used in the Model

Variable	Type	Code	Explanation of variables
Frequency of crashes at each loop detector station for different crash categories	Response	FREQ	
Radius Category	Qualitative	RADCAT	> 3000 ft – 0 <=3000 ft – 1
Number of lanes	Quantitative	LANES	
Median Type	Qualitative	MTYPCAT	Without barrier – 0 With barrier – 1
Median Width	Quantitative	MEDWID	
Pavement Condition	Quantitative	PAVCOND	3 – 5 scale. With 5 being very good and 3 being fair.
Pavement Surface Type Category	Qualitative	PSURCAT	0 – Asphalt 1 – Concrete
Pavement Roughness Index	Quantitative	PRI	40 – 78. It is the calibrated roughness measurement to the nearest inch per mile.
Off-ramp(s) presence within the influence area of the loop detector	Qualitative	OFFRCAT	0 – absent 1 – present
On-ramp(s) presence within the influence area of the loop detector	Qualitative	ONRCAT	0 – absent 1 – present
Annual Daily Traffic Volumes	Quantitative	AADT	
Peak Fifteen minute volumes	Quantitative	PEAKFIFT	
75% percentile of Average Speed	Quantitative	AVGS	
75% percentile of Standard Deviation of Speed	Quantitative	STDS	
75% percentile of Coefficient of Variation of Speed	Quantitative	CVS	
75% percentile of Standard Deviation of Volume	Quantitative	STDV	
75% percentile of Coefficient of Variation of Volume	Quantitative	CVV	

4.3.1 Category 1

Based on the type of crash, there are two sub-categories, multiple and single vehicle crashes. A SUNB model was estimated for this category. Two left-hand side (dependent) variables were considered: frequency of multiple and single vehicle crashes. The right-hand side (independent) variables consisted of traffic, roadway and geometric factors.

4.3.1.1.1 Individual Multiple Vehicle Crash Model

Before arriving at the final model, two models were attempted, one with AADT and another with PEAKFIFT keeping all other variables same. PEAKFIFT was not found to be significant in the model. Hence the model with AADT was chosen as the final individual multiple vehicle crash model. As for other traffic variables extracted from loop detector data, 5, 10, and 15 minute aggregations were tried. And for each aggregation, standard deviation of volume and speed, or coefficient of variation of volume and speed were used separately to avoid the correlation among these measures. The multiple vehicle crash model was selected based on the criteria illustrated in modeling approach section, where different decisive factors were explained. It has a log-likelihood value of -1252.372. The estimation results are provided in Table 4.2.

4.3.1.1.2 Individual Single Vehicle Crash Model

As explained in multiple crash model estimation, before arriving at the final model, two models were tried, one with AADT and another with PEAKFIFT keeping all other variables same. Neither PEAKFIFT nor AADT was found to be significant in the model. Also there were no significant microscopic traffic variables. It has a log-likelihood value of -708.321. The estimation results are provided in Table 4.3.

Table4.2: Estimation results for individual multiple vehicle crash model

Parameter	Estimate	Standard Error
CONSTANT	-0.21278	1.02616
RADCAT	0.342407	0.194975
MTYCAT	-0.43596	0.172878
PSURCAT	0.747703	0.227456
OFFRCAT	0.424278	0.122057
ONRCAT	0.447878	0.125682
AADT	0.265633	0.086238
ALPHA 1	0.157812	0.030254

Log Likelihood: -1252.372 No. of observations: 552

Table 4.3: Estimation results for individual single vehicle crash model

Parameter	Estimate	Standard Error
CONSTANT	0.496707	1.059095
RADCAT	0.293304	0.201705
MTPYCAT	-0.30851	0.1753
OFFRCAT	0.494572	0.131351
ONRCAT	0.224393	0.126161
ALPHA 2	0.154817	0.086474

Log Likelihood: -708.321 No. of observations: 552

4.3.1.1.3 Seemingly Unrelated Negative Binomial Model for Multiple and Single Vehicle Crashes

As mentioned previously, SUNB estimation was performed for multiple and single vehicle models. Dispersion parameters (ALPHA 1 and ALPHA 2), standard deviation for disturbance terms (SIGMA_U1 and SIGMA_U2), and correlation coefficient (RHO_U1U2) were evaluated. The estimation results are provided in Table 4.4.

As shown in Table 4.4-c, the correlation between the disturbance terms is relatively high with a value of 0.75. This implies that the omitted variables are allocated across disturbances for multiple and single vehicle crashes. Therefore, the use of the SUNB model estimation is justified and facilitated in efficient parameter estimates. SIGMA_U1, SIGMA_U2 and RHO_U1U2 are part of the correlation matrix estimated for the SUNB model.

Table 4.4-a: SUNB model estimation results for multiple vehicle crash model

Parameter	**Estimate**	**Standard Error**
CONSTANT	-0.23624	1.036856
RADCAT	0.331074	0.192185
MTYPCAT	-0.41066	0.170384
PSURCAT	0.782405	0.247369
OFFRCAT	0.404423	0.128966
ONRCAT	0.431569	0.122117
AADT	0.264062	0.093646
ALPHA 1	0.158552	0.029867

Log Likelihood: -1951.216 No. of observations: 552

Table 4.4-b: SUNB model estimation results for single vehicle crash model

Parameter	Estimate	Standard Error
CONSTANT	0.526154	1.064154
RADCAT	0.298596	0.199444
MTYPCAT	-0.32581	0.176071
OFFRCAT	0.477845	0.132039
ONRCAT	0.228205	0.125955
ALPHA 2	0.156048	0.085439

Log Likelihood: -1951.216 No. of observations: 552

Table 4.4-c: Model estimation results contd.

Parameter	Estimate	Standard Error
SIGMA_U1	0.571609	0.052894
SIGMA_U2	0.402274	0.077594
RHO-U1U2	0.748625	0.13375

4.3.2 Category 2

This category comprises two sub categories, of which one has peak period crashes and the other has off-peak period crashes. A SUNB model was estimated for this category. Two left-hand side (dependent) variables were considered: frequency of peak and off-peak period crashes. The right-hand side (independent) variables consisted of traffic, roadway and geometric factors. The correlation between the error terms for the SUNB model was very high (very close to 1), and caused difficulty in estimating peak and off-peak period crashes simultaneously. An attempt was made to estimate the simultaneous equations by setting the correlation to unity produced slightly improved standard errors. The separate negative binomial models showed the significance of roadway curvature, pavement surface type, off and on ramps, AADT and the 15 minutes coefficient of variation of speed, on the frequency of peak period crashes. Roadway curvature, median type, pavement surface type, off and on ramps and AADT were significant in the off-peak model. While AADT is significant in both models, a microscopic variable representing the variation in speed at the 15 minutes level was also significant in the peak period crash model.

4.3.3 Category 3

This category also comprises two sub categories, dry pavement and wet pavement crashes. A SUNB model was estimated for this category. Two left-hand side (dependent) variables were considered: frequency of dry and wet pavement crashes. The right-hand side (independent) variables consisted of traffic, roadway and geometric factors. The correlation between the error terms for these two models was very high (very close to 1), and therefore caused difficulty in estimating dry and wet pavement crashes simultaneously. The separate models showed the significance of roadway curvature, median type, pavement surface type, off and on ramps and AADT in the frequency of dry crashes. Only curvature, on and off ramps and AADT were found significant in the wet pavement crash model.

4.3.4 Category 4

Based on availability of daylight, there are two sub-categories, daylight and dark hour crashes. A SUNB model was estimated for this category. Two left-hand side (dependent) variables were considered: the frequency of daylight and dark hour crashes. The microscopic traffic factors included in these models were obtained separately for daylight and dark hours. For instance CVS for day hour crash model was taken only during the day time with sun light availability. For the purpose of obtaining these microscopic traffic parameters, daylight was considered from 5:30 A.M. to 7:00 P.M. during summer and 6:30 A.M. to 5:30 P.M. during winter.

The individual daytime crash model had a log-likelihood value of -1211.34. The estimation results are provided in Table 4.5. The individual dark hour crash model had a log-likelihood value of -859.649. The estimation results are provided in Table 4.6.

Table 4.5: Estimation results for individual day time crash model

Parameter	Estimate	Standard Error
CONSTANT	0.223723	1.434843
RADCAT	0.3781	0.230821
MTYPCAT	-0.33195	0.193563
PSURCAT	0.881797	0.340125
OFFRCAT	0.390121	0.164871
ONRCAT	0.432593	0.15517
AADT	0.125771	0.153611
CVS_15	0.414549	0.259527
ALPHA 1	0.15796	0.03657

Log Likelihood: -1211.34 No. of observations: 552

Table 4.6: Estimation results for individual dark hour crash model

Parameter	Estimate	Standard Error
CONSTANT	0.199007	0.904107
RADCAT	0.376247	0.168859
MTYPCAT	-0.44595	0.158141
PSURCAT	0.205027	0.215089
OFFRCAT	0.485975	0.114783
ONRCAT	0.294935	0.109333
AADT	0.207726	0.07536
ALPHA 2	0.098216	0.053079

Log Likelihood: -859.649 No. of observations: 552

4.3.4.1.1 Seemingly Unrelated Negative Binomial Model for Daylight and Dark Hour Crashes

As mentioned previously, SUNB estimation was performed for daylight and dark hour crash models. Dispersion parameters (ALPHA 1 and ALPHA 2), standard deviation for disturbance terms (SIGMA_U1 and SIGMA_U2), and correlation coefficient (RHO_U1U2) were evaluated. The estimation results are provided in Table 4.7. As shown in Table 4.7-c, the correlation between the disturbance terms is substantially high with a value of 0.95. This entails that the omitted variables are shared across the model disturbances for day and dark hour crashes. Therefore the use of SUNB estimation is warranted and assisted in efficient parameter estimates. SIGMA_U1, SIGMA_U2 and RHO_U1U2 are part of the correlation matrix estimated for the SUNB model. Through the estimation of SUNB models, the errors were minimized and the reliability of the models was increased which is shown by smaller standard errors. A comparison of standard errors between individual models and SUNB models would help in understanding the efficiency gained. This comparison table is provided in Table 4.8. As observed from Tables 4.8 a and b, most of the parameter coefficients in the SUNB models have smaller standard errors.

4.3.5 Category 5

This category includes two sub categories, PDO and injury crashes. A SUNB model was estimated for this category. Two left-hand side (dependent) variables were considered: PDO and injury crashes. The correlation between the error terms for these two models was very high (very close to 1), and therefore caused difficulty in estimating PDO and injury crashes simultaneously. The separate negative binomial models depicted that roadway curvature, pavement type, AADT and on and off ramps significantly influence the frequency of both PDO and injury crashes. In addition, the median type affected only the frequency of PDO crashes.

Table 4.7-a: SUNB model estimation results for day time crash model

Parameter	Estimate	Standard Error
CONSTANT	0.445938	1.040764
RADCAT	0.274204	0.176907
MTYPCAT	-0.37914	0.160429
PSURCAT	1.065064	0.229792
OFFRCAT	0.379462	0.115442
ONRCAT	0.540367	0.132584
AADT	0.126856	0.073355
CVS_15	0.438534	0.251081
ALPHA 1	0.15	0.033303

Log Likelihood: -2113.48 No. of observations: 552

Table 4.7-b: SUNB model estimation results for dark hour crash model

Parameter	Estimate	Standard Error
CONSTANT	0.498033	0.884796
RADCAT	0.352927	0.155771
MTYPCAT	-0.45057	0.148744
PSURCAT	0.354589	0.205383
OFFRCAT	0.467791	0.108309
ONRCAT	0.406194	0.112068
AADT	0.177486	0.068655
ALPHA 2	0.09618	0.052127

Log Likelihood: -2113.48 No. of observations: 552

Table 4.7-c: SUNB model estimation results contd.

Parameter	Estimate	Standard Error
SIGMA_U1	0.597484	0.053107
SIGMA_U2	0.395009	0.059129
RHO_U1U2	0.950000	0.105333

Table 4.8-a: Comparison of standard errors for day time crash model

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.230821	0.176907
MTYPCAT	0.193563	0.160429
PSURCAT	0.340125	0.229792
OFFRCAT	0.164871	0.115442
ONRCAT	0.15517	0.132584
AADT	0.153611	0.073355
CVS_15	0.259527	0.251081

Table 4.8-b: Comparison of errors for dark hour crash model

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.168859	0.155771
MTYPCAT	0.158141	0.148744
PSURCAT	0.215089	0.205383
OFFRCAT	0.114783	0.108309
ONRCAT	0.109333	0.112068
AADT	0.07536	0.068655

4.3.5.1.1 Measurement of Goodness-of-fit

There seems to be no universally accepted goodness of fit for seemingly unrelated negative binomial models. There are two alternative methods, for estimating the goodness-of-fit of SUNB models. There are 1) R_p^2 statistic, and 2) G^2 statistic.

R_p^2 is given as:

$$R_p^2 = \frac{\sum_{i=1}^{n}\left[\frac{y_i - \lambda_i}{\sqrt{\lambda_i}}\right]}{\sum_{i=1}^{n}\left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}}\right]}$$

G^2 is given as:

$$G^2 = \sum_{i=1}^{n} 2\{y_i \ln(y_i / \lambda_i) - (y_i - \lambda_i)\}$$

In the above equations λ_i is the expected number of crashes for a particular observation y_i, as defined by the model. For instance, the expected number of crashes in multiple vehicle crash model can be shown as:

$$\lambda_i = EXP(-0.21278 + 0.342407 * RADCAT - 0.43596 * MTYPCAT + 0.747703 * PSURCAT + 0.424278 * OFFRCAT + 0.447878 * ONRCAT + 0.265633 * AADT)$$

G^2 and R_p^2 are calculated separately for the individual models first, and then for the SUNB models. R_p^2 statistic was computed for all the individual and SUNB models. The values were very close and thus making it difficult to differentiate between individual and SUNB models. So the other statistic G^2 was used to identify the model with the better goodness-of-fit. Table 4.9 provides the details of the G^2 statistic for various individual and SUNB models for the five categories.

Based on the smallest values of G^2, the following conclusions can be drawn:

- Both multiple and single vehicle crash models were improved by SUNB estimation
- Both peak and off-peak period crash models were improved by SUNB estimation
- Peak period crash model improved substantially, while there was little improvement in the off-peak period crash model
- There was no improvement in both daytime and dark hour crash models with SUNB estimation

Even though goodness-of-fit statistics do not show improvement in all models with SUNB estimation, a good explanation behind estimation of SUNB models arrives from the significant correlation coefficient between the error terms arising from the omitted variables. For instance, in category 4, both the daytime and dark hour crash models did not improve upon SUNB estimation. Nevertheless these models have small standard errors and the correlation coefficient was substantially high.

Table 4.9: Goodness-of-fit statistics for different crash categories

GOODNESS-OF-FIT TABLE				
	Individual Model	**G-square Statistic**	**SUNB Model**	**G-square Statistic**
Category 1	Multiple Vehicle	**3264.45**	Multiple Vehicle	**3123.83**
	Single Vehicle	**3168.36**	Single Vehicle	**3143.04**
Category 2	Peak Period	**7359.85**	Peak Period	**3175.8**
	Off-peak Period	**1134.36**	Off-peak Period	**1128.95**
Category 4	Daytime	**3257.90**	Daytime	**5826.67**
	Dark Hour	**4217.21**	Dark Hour	**5388.35**

4.4 Section Summary

This research investigated a technique to address the problem of correlation between the error terms, when the crashes are divided into different logical categories (e.g., single and multiple vehicle crashes), while modeling crash frequencies. The results showed that accounting for the correlation factor between error terms is imperative while modeling crash frequencies for different crash categories. This resulted in better models in terms of improved parameter estimates and better goodness-of-fit of the models, while allowing for more accurate identification of factors related to the different crash categories.

The first models, included multiple and single vehicle crashes, had a significant correlation coefficient which lead to the main justification of estimating SUNB models for this category. Also the goodness-of-fit of both multiple and single vehicle crash models was improved. The significant factors in the multiple vehicle crash model were roadway curvature, median type, pavement surface type and presence of on-ramps/off-ramps and AADT. In the case of the single vehicle crash model, the significant factors were roadway curvature, median type, and presence of on-ramps/off-ramps. Therefore, the common factors influencing both multiple and single vehicle crashes were road curvature, median type, and presence of on-ramps/off-ramps. However, the effect of off-ramps was more profound compared to the on-ramps in the single vehicle model, as could be observed by the value of parameter coefficient. In the multiple vehicle model both were comparable. The results indicated that an increase in AADT caused more multiple vehicle crashes, while AADT had no effect on single vehicle crashes. This can be justified because the increase in volume increases the probability of interaction among vehicles, which is generally related to more multiple vehicle crashes. Single vehicle crashes on the other hand are believed to occur because of speeding, which is more of a driver related factor.

Modeling simultaneously the frequency of peak and off-peak period crashes improved the goodness-of-fit for the SUNB peak period crash model when compared to the individual model. The goodness-of-fit for the off-peak period crash model also increased. The significant factors in peak period crash model were road curvature, pavement surface type, presence of on-ramps/off-ramps, AADT, and coefficient of variation in speed during peak period aggregated for 15 minute interval. In the case of off-peak period crash model, the significant factors were road curvature, median type, pavement surface type, presence of on-ramps/off-ramps and AADT. Therefore, the common factors influencing both these crashes were road curvature, pavement surface type, presence of on-ramps/off-ramps, and AADT. Median type was found to affect only off-peak period crashes, while the coefficient of variation in speed is found to affect only peak period crashes. We observe higher coefficient of "variation in speeds" during "peak periods" where vehicles travel at low speeds, which is the cause of crash occurrence.

SUNB was used to model the daytime and dark hour crash frequencies. The significant factors in the day time crash model were road curvature, median type, pavement surface type and presence of on-ramps/off-ramps, AADT, and coefficient of variation in speed aggregated for 15 minute interval. In the case of dark hour crash model, the significant factors were road curvature, median type, pavement surface type, presence of on-ramps/off-ramps and AADT. Thus, the common factors influencing both these crashes were road curvature, median type, pavement surface type, presence of on-ramps/off-ramps, and AADT. Coefficient of "variation in speed" was found to affect only "daytime crashes", which is reasonable. During daytime peak traffic conditions occur, causing higher coefficient of variation in speed, which in turn causes crashes.

To summarize, radius (i.e., horizontal curvature of the roadway) category, presence of on-ramps, and presence of off-ramps appeared in all the models. AADT was also found to influence all the crash categories except for single vehicle crashes. This could indicate that AADT is still an acceptable measure of traffic volume. Median type appeared in all models except for wet pavement and injury crash models. A reasonable explanation can be put forth as follows: medians without barrier as explained in Souleyrette et al., resulted in more crashes, injury crashes might be strongly associated with other factors so that median type is not significant in such crashes. Pavement surface type was found significant in all models except for single and wet pavement crash models. Coefficient of variation in speed was found to influence only peak and daytime crash models. These conditions, i.e. peak and daytime traffic conditions, cause higher coefficient of variation in speeds which result in more crash occurrences. CVS was the only microscopic factor that was significant in these models although other microscopic factors were significant in the preliminary analysis.

Using the crash frequency models developed in this work, and using specific traffic volume values from archived loop detectors, the risk at each section of the freeway could be evaluated. Different scenarios could be adopted based on typical traffic volume counts by time of day, day of week, season, etc. Higher risk locations on the freeway might change by time and day based on the specific traffic volume. This could help traffic management centers draw a detailed picture of the risk on the freeway, and therefore allocate the response and resources. A possible extension to this work is the possibility that similar models could be implemented real-time to indicate an increase in the risk level at different locations of urban freeways as a function of changing traffic volumes given the roadway characteristics of each location. Future work could attempt to add more independent variables in the models to avoid the difficulties in estimating SUNB models with high correlation between the error terms.

CHAPTER 5

USING VARIABLE SPEED LIMITS FOR REAL-TIME SAFETY IMPROVEMENT

This research evaluates Intelligent Transportation System (ITS) implementation through Variable Speed Limits (VSL) strategies to improve the safety of a freeway once a relatively high potential of a crash is detected. VSL are ITS devices, commonly used to calm traffic in an attempt to relieve congestion and enhance throughput. A different aspect of VSL can be realized in the improvement of traffic safety. Through the use of multiple microscopic traffic simulations, best practices can be determined, and a final recommendation of VSL strategies with a safety perspective can be made.

5.1 Application of VSL

Variable speed limits are used to increase average headways and reduce variances in speed (Borrough, 1997, Ha et al., 2003, Pilli-Sivola, 2004). However, less variability of speed leads to fewer short headways, and lower mean speeds according to Ha et al. (2003). This translates into fewer crashes (Smulder 1990). A study in Finland by Rämä (1999) shows that VSL lead to lower speeds and less variability. Borrough (1997) found that the use of VSL and strong enforcement (video cameras) greatly reduced the number of crashes (28% over 18 months). The effect was attributed to not only a smoothing of traffic conditions through longer following distances, but also through reducing the number of lane changes during congestion (Borrough, 1997).

Lee et al. (2004) used VSL to try and reduce crash potentials. Lee et al. (2004) only simulated a one mile long stretch that included just one ramp, and placed Variable Message Sign (VMS) just upstream of a particular location of interest. They attempted few scenarios and did not address all the factors that are related to VSL application. In this study, a 20 mile stretch is simulated. The larger network allows more flexibility in the implementation of VSL. Rather than having just two locations that effect traffic speeds, up to 12 miles are investigated via 24 VSL test cases with both upstream and downstream introduction of VSL. Also speed limits are decreased, increased, or simultaneously decreased and increased (up and down stream, respectively) to investigate all possible cases. This introduces a more dynamic approach for implementation taking into consideration all the possible scenarios extending significantly over Lee et al. (2004) where they investigated either increasing or decreasing speeds in order to improve safety.

5.2 Micro Simulation

When studying VSL, microsimulation appears to be an ideally suited tool to evaluate ITS technologies especially like VSL (Lee et al., 2004). Generally, data is collected using loops (Placer, 2001; Sisiopiku, 2001; Senn, 2004; Borrough, 1997; Portaankorva, 2002). or speed radars (Sisiopiku, 2001, Pilli-Sivola, 2000, Peltola, 1997) and field studies are undertaken to test strategies. In the transportation simulation field there is a general agreement that microscopic simulation, i.e. a computational resolution down to the level of individual travelers, is now a viable alternative and may be the only answer to a wide variety of problems.

Simulation has some desirable qualities that make it useful. First, it is cheaper than field-testing in most cases. The upfront cost of software and hardware does not compare to the cost of outfitting a road network, or the loss in confidence of the driving public on an always-changing driving environment. Second, impossible scenarios maybe carried out. Third, time can be sped up to yield future results now. The effect of a change many months away can be determined in few hours. Fourth, safety issues can be safely tested without the potential of harming any drivers.

For Variable Message Signs which are closely related to VSL, the simulated technology in this instance, only two software packages are compatible. The AIMSUN 2 software has VMS capability as does PARAMICS (Boxill and Yu, 2000). While either would look to be a perfect choice PARAMICS was chosen due to its scalability and proven background on freeways and urban roads. Furthermore, PARAMICS was preferred over CORSIM since it has advantages in implementing ITS applications (e.g., variable speed limits and ramp metering) through its Application Programming Interface (API).

5.3 Safety Measure

Abdel-Aty et al., (2005), Lee et al., (2003) and Pande et al., (2005) developed statistical models to get a measure of real-time crash potential. Lee et al. (2004) applied the model from one of their previous studies (Lee et al., 2003) to a small simulated freeway network (1 mile) to measure the crash risk before and after the application of certain VSL strategies

The model developed by Abdel-Aty et al. (2005) is used to assess crash likelihood for the simulated network used in this study. The modes were developed for the same segment of Interstate-4 being simulated here, making them the most appropriate choice. Also, real-time crash "prediction" models were separately developed for a moderate-to-high-speed and low-speed traffic speed regime and the threshold for separating the two regimes was set at 37.5 mph based on visual examination of traffic speed distributions. Above this speed, a moderate-to-high-speed model, which takes average occupancy and flow as input, is used. Below this speed, a low-speed model, involving average volume, occupancy, and coefficient of variation in speed variation as inputs, is used. These models may be used to assess the crash potential at any given location in real-time using loop detector data. Since the input parameters to these models were measured 5-15 minute before the crash; there would be time to introduce strategies at locations experiencing crash prone conditions before they culminate into a crash. The moderate-to-high and low-speed models are shown in Equations 5.1 and 5.2, respectively (Abdel-Aty et al., 2005).

$$\begin{aligned} Crash\_Likelihood = & -0.93423 \text{LogAOF2} + 1.14584 \text{LogAOH3} - 0.22878 \text{SVH2} - \\ & 0.10055 \text{AVG2} + 05932 \text{AVE3} \end{aligned} \tag{5.1}$$

LogAOF2: Log of average occupancy at the station of interest 5 to 10 minutes before the time of interest
LogAOH3: Log of average occupancy one mile downstream of the station of interest 10 to 15 minutes before the time of interest
SVH2: The standard deviation of volume one mile downstream of the station of interest 5 to 10 minutes before the time of interest
AVG2: The average volume half mile downstream of the station of interest 5 to 10 minutes before the time of interest
AVE3: The average volume half mile upstream of the station of interest 10 to 15 minutes before the time of interest

$$\begin{aligned} Crash\_Likelihood = & \ 2.64827 \text{LogCVSF2} + 0.88842 \text{LogCVSF3} + 1.33966 \text{LogAOE2} + \\ & 0.97766 \text{LogAOH3} - 0.43603 \text{SVH2} \end{aligned} \tag{5.2}$$

LOGCVSF2: The log of the standard deviation of speed divided by the average speed at the station of interest 5 to 10 minutes before the time of interest
LOGCVSF3: The log of the standard deviation of speed divided by the average speed at the station of interest 10 to 15 minutes before the time of interest
LogAOE2: Log of average occupancy half a mile upstream of the station of interest 5 to 10 minutes before the time of interest
LogAOH3: Log of average occupancy one mile downstream of the station of interest 10 to 15 minutes before the time of interest
SVH2: The standard deviation of volume one mile downstream of the station of interest 5 to 10 minutes before the time of interest

The models shown above would provide a measure that may be used to evaluate the impact of our application experiments on the safety situation of the freeway. This measure is specific for every location and a decrease in this measure signifies a decrease in the risk of crash and vice-versa. It should be noted, however that it can only be used to compare crash risk at the same station before and after implementing certain strategies. In other words, based on this measure we can not infer for example, whether station 33 is at higher risk than station 34 or 37; though we can find out, for example, if the crash risk at station 33 has been reduced after application of a particular VSL strategy.

5.4 Study Corridor

Interstate 4 (I-4) is the main arterial for the Orlando metropolitan area. It has spurred development along its path to both the north and south of downtown Orlando. Data from dual loop detectors of I-4 are simulated in this study by replicating the loading conditions, geometric features, and loop locations. Loops collect data on average vehicle counts, average speed, and lane detector occupancy, across three lanes on I-4 in each direction every 30 seconds at 69 stations spaced every half a mile for about a 36-mile stretch. Over the next several years, many Florida Department of Transportation projects plan to expand the number of lanes in an effort to boost capacity of the roadway. In addition, Variable Speed Limits are scheduled to be added to the roadway to promote better management of the stretch.

The section of I-4 under study runs from just south of downtown Orlando north into the surrounding suburban cities. It spans Loop Detector Stations 1 to 69 from South to North. The section has 17 interchanges with 59 ramps, curves with radii varying from 1910 to 85,944 ft, and speed limits from 50 mph to 65 mph. The roadway maintains a consistent elevation throughout its length.

Of the twenty mile section three different locations: station 33, station 47, and station 61 are investigated for VSL implementation. The three locations represent different geometric features and traffic conditions.

- Station 33 is investigated in the east bound direction in the high volume test case (i.e., low speed or congested situation). It has a high number of crashes (around 90 crashes over 4 years) and is located in the downtown area.
- Station 61 is investigated in the westbound direction in the low loading test case (i.e., high speed or non congested situation). It is located well outside of downtown on the north side of Orlando and has a relatively high number of crashes for areas outside of downtown, but compared to the downtown area the number of crashes is low.
- Station 47 in the westbound direction is used to investigate the low loading case (also high speed). It is just north of downtown and has a high number of crashes (55 crashes over 4 years) amongst stations in westbound direction.

All three stations are located near or on curves and near ramps.

5.5 Calibration and Verification

The literature shows many techniques and values that have been used to calibrate PARAMICS. In studies by Bertini et al. (2002), Abdulhai et al. (2002), Trapp (2002), and Stewart (2001), the multiple factors used to calibrate their networks are laid out. Each mentions the use of flow and travel time, which across literature appear to be the factors of choice. However, only Bertini et al. (2002) showed the calibration values used in simulating their network, and mentioned adjusting two parameters, neither of which directly affects queuing behavior. Abdulhai et al. (2002) mentioned that they calibrated for vehicle bunching, which based on their description is excessive queuing, by decreasing the memory speed, furthering the idea that queuing behavior is an important factor to consider.

Cheu et al. (2002) and Bertini et al. (2002) noted that in PARAMICS the mean target headway and mean driver's reaction time need to be calibrated based on the area's drivers for accurate flow characteristics, but did not reveal the values they used. Gardes et al. (2002), Abdulhai (2002), and Lee et al. (2001) stated their values for mean headway and driver reaction as 1 and 0.6 second, 1.65 and 0.42 second, and 0.615 and 0.415 second, respectively. Other factors involved in calibrating PARAMICS are the time step, aggressiveness, and minimum gap values (Abdulhai et al., 2002, Cheu et al., 2002).

Initially, the mean target headway and mean driver's reaction time are changed to values found in the literature. The values along with the corresponding errors are shown in Table 1. The errors are determined by comparing mainline counts on the simulated network to mainline counts on the roadway. It is important to stress that these counts are used as a starting point for verification. The error values are not of a concern since they are based on Average Daily Traffic, which is an aggregate measure of traffic flow that is used with the K-factor to obtain a traffic volume value at the peak period.

Table 5.1: Core variable specification and their resulting error percentages

Run Number	Headway	Reaction Time	Queue distance	Queue speed	Error Rate
1	1.00 sec	0.60 sec	35.41 ft	4.470 mph	17.14
2	1.65 sec	0.42 sec	35.41 ft	4.470 mph	20.5
3	0.61 sec	0.42 sec	35.41 ft	4.470 mph	14.54
4	0.50 sec	0.50 sec	35.41 ft	4.470 mph	14.36

After calibration showed that headway and reaction time of 0.5 seconds and 0.5 seconds have the lowest error rate, speeds were inspected, revealing a reoccurring backward shockwave at one on-ramp. A strong backward shockwave at the head of the queue is expected when a crash is cleared or some other type of bottleneck is removed from a system; but when the bottleneck does not move and persists over the entire simulation, this type of shockwave is not expected. To correct the simulation, the driver behavior characteristics in the simulation environment were changed. These changes included varying the Mean Target Headway that the drivers maintain and Mean Reaction Time of the drivers to determine their effect on the development of shock-waves.

PARAMICS runs were inspected with different values of core parameters: target headway, mean reaction time, queuing speed, and queuing distance. Behavior characteristics, aggressiveness and awareness, were also tested, but no effect was seen from changing their values. Each runs' effect on vehicle behavior was qualitatively compared by the modeler to determine an approximate value for each parameter.

By inspecting the resulting speeds when the core numbers are close to the approximate values, it is possible to determine if the detector is experiencing no congestion, mild congestion, or heavy congestion. Comparing the results to where congestion is expected, the best core numbers were determined. The congestion location standard is determined by field. Note that speed variation and queue formation are the most important factors in the equations used to capture the crash potential (Abdel-Aty's et al., 2005). Therefore, it is important to achieve realistic speeds and queuing behavior, which in turn explains why speed profiles and the nature of the queues were seen as crucial to the objectives of this simulation study.

The values that produced the best result were mean headway of 1.0 second, a mean reaction time of 0.42 seconds, a queuing speed of 8 mph and a queuing distance of 9 ft. The values for mean reaction time and mean headway most closely resemble Lee et al's (2001) values of 0.42 and 0.62 second, respectively. The mean headway is shorter for Lee et al. (2001) because the study used the software's default queue distance and queue speed. Through the multiple rounds of verification, the queue distance and the mean headway appeared to interact when the headway was small, so a different headway from that of Lee et al.(2001) could be justified due to the different queue distance. Headways of 1.0 second are the default offered by PARAMICS and is supported by Gardes et al. (2002).

5.6 Test Cases

Five minutes moving averages of speed at each station were determined based on two loading conditions: a peak loading and an off-peak loading. The vast majority of detectors during the peak loading were found to be in a low-speed condition, while the vast majority of detectors during the off-peak loading were found to be in a moderate-to-high-speed condition, as the condition determined according to those defined by Abdel-Aty et al. (2005). As a result a detector from the peak loading was chosen to be used for low-speed testing and a detector from the off-peak loading was chosen to be used for high-speed testing.

Extensive testing with both speed regimes was attempted. It was concluded that VSL does not affect the crash likelihood at low speed regimes. This is intuitive, since the freeway is congested, and therefore the speed is already low, and increasing the speed limit would have no effect. An example comparing the crash likelihood, calculated by Equation 2, for station 33 East-bound is shown in Figure 5.1

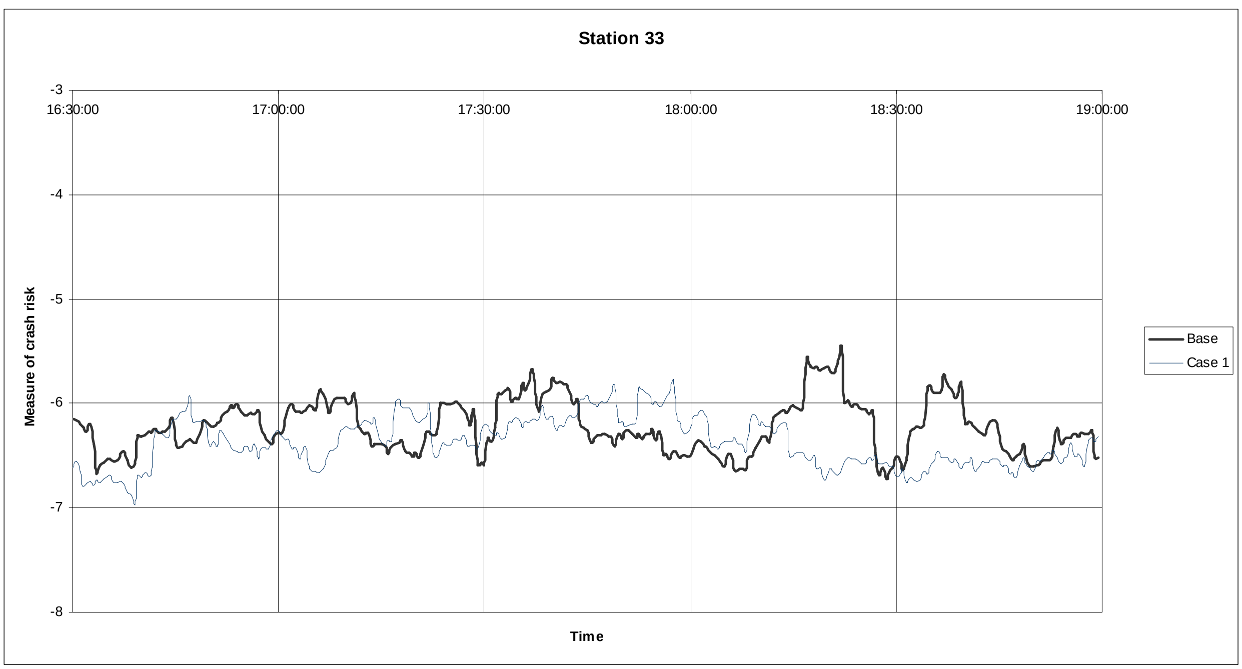


Figure 5.1: Station 33 crash potential for the base case and VSL test case under the low-speed traffic regime with matched seeds

.

Figure 5.1 shows that there is almost no difference between the base case (no VSL) and the test case (introducing VSL). Multiple scenarios over 20 different but matched seed values were introduced and no effect was noticed. Therefore moderate-to-high speed regime (i.e., average speed above 37.5 mph) was focused upon thereon.

5.7 High Speed Test Case

For the moderate-to-high-speed test case a complex scenario that involves lowering and raising the speed limit was used. Figure 5.2 depicts the terms that were used in the moderate-to-high-speed case. Notice that distances can be listed as either number of miles or number of detectors plus the direction. For example, 1 mile upstream of detector 47 or 2 detectors upstream of detector 47 would be specifying the same location.

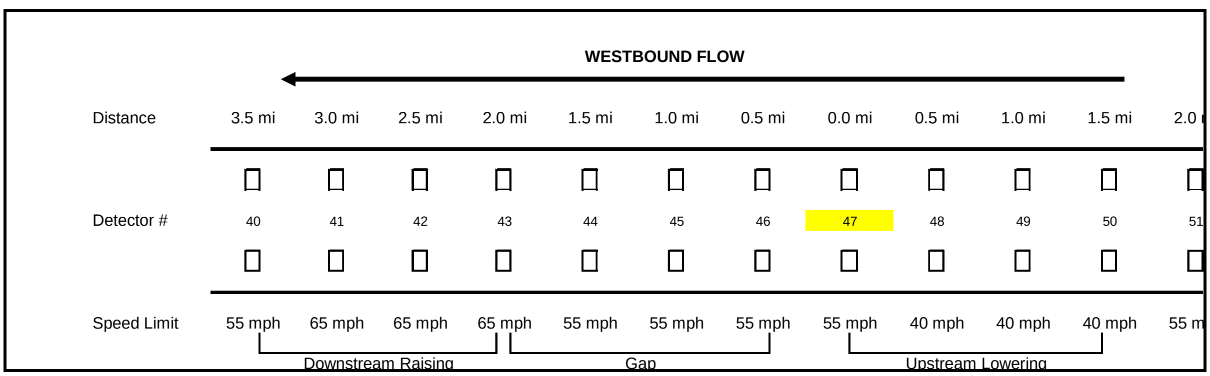


Figure 5.2: Sample scenario for a westbound roadway with detector 47 as the detector of interest.

The moderate-to-high-speed implementation of VSL is unique in that it applies both upstream and downstream changes, rather than only upstream or only downstream (which were also attempted). Along with this, the implementation strategy involved raising or lowering speed limits, instead of only lowering speed limits. Careful Crash Risk analysis (Abdel-Aty et al., 2004; 2005) showed an increase in the potential of crashes on freeways when queues start forming. Queuing causes backward shockwaves while speeds were still high upstream. Therefore, the idea behind downstream increase of speed limits comes from the paradigm that there is a group of cars approaching an existing queue. With this in mind, it is logical to attempt to slow cars upstream to keep them from hitting the queue and raise the speed limit downstream to help the queue dissipation. A typical speed profile of this approach is depicted in Figure 5.3.

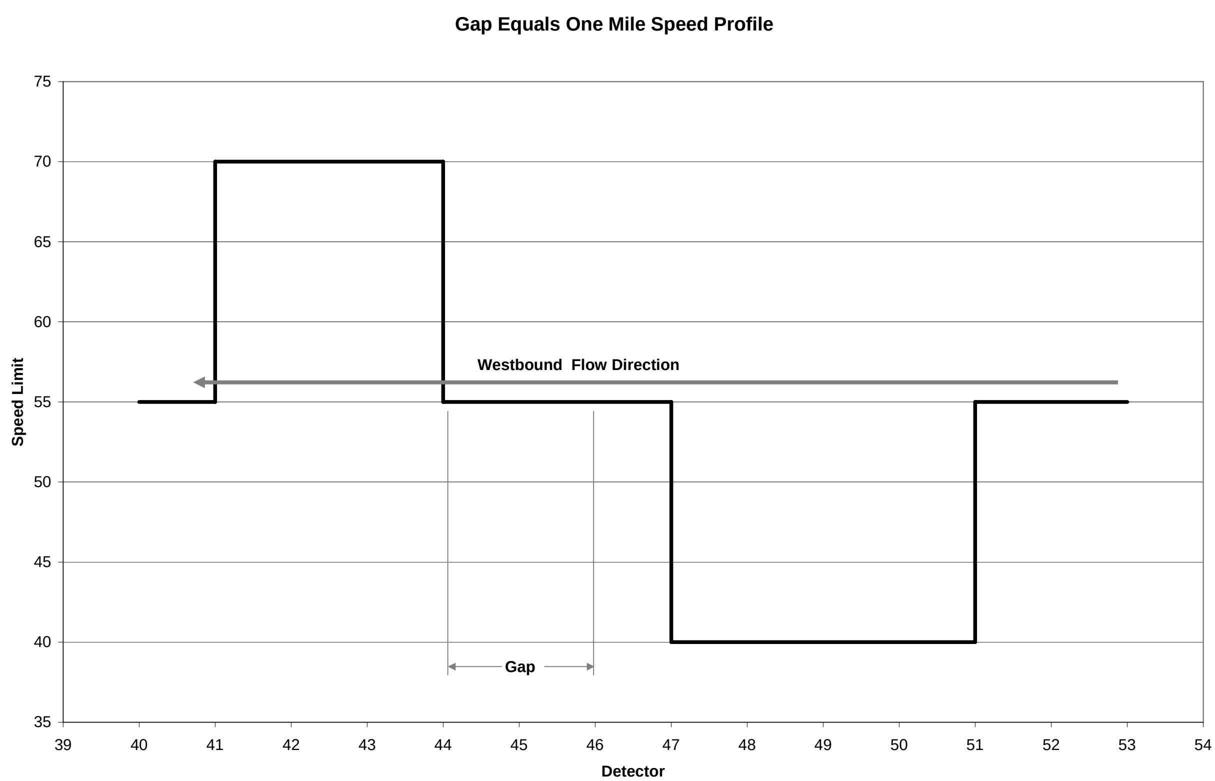


Figure 5.3: A sample speed limit profile across space.

5.8 Design of Experiment

A three-hour simulation, with a 15-minute initialization period, was used to evaluate VSL effectiveness around detector 47 in the westbound direction. In order to see the effect of crash potential changing over time, speed limits were changed 30 minutes into the simulation and were maintained for only 30 minutes. After that time the speed limits were changed back to their base value.

Speed limits were changed using multiple links files that were accessed at different predetermined times. The speed control setting, used via programmer function used by Lee et al. (2004), was compared to the multiple links files method and similar results were seen, validating the use of the method.

A full factorial experiment was attempted, but because a suitable measure of effectiveness could not be determined, an alternative stepwise approach was used akin to Forward Sequential Selection.

The High Speed Test case served as the starting point of the experiment. Four separate steps were conducted with the best case scenario from the previous set of scenarios being used as the input for the next test. Figure 5.4 gives the layout of the tests and which are referenced in Tables 5.2, 5.3, 5.4, and 5.5, respectively.

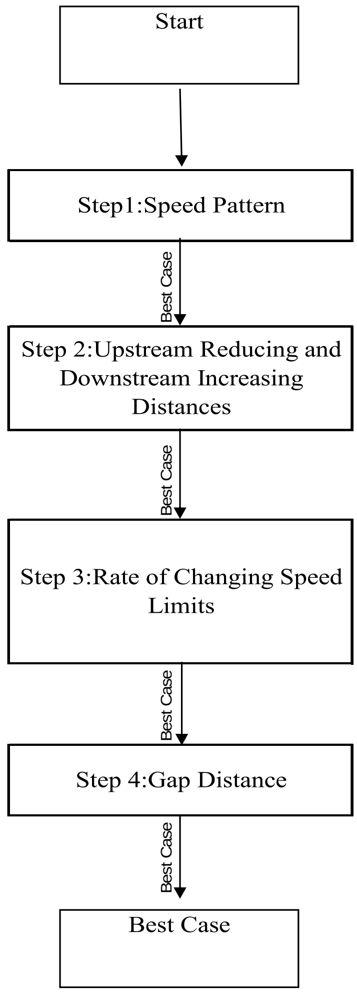


Figure 5.4: Steps followed to evaluate the effect of using VSL in the high-speed case.

In Step 1, the objective was to find the Best strategy of Speed Limit Patterns. The purpose is three fold.

- To find out if simultaneously lowering speeds upstream of station of study and raising them downstream of station of study is better than just lowering speeds upstream of station of study or just raising them downstream of station of study. The location of the change refers to whether the change is going to take place in the 5.5 miles downstream of the detector, or for the same stretch downstream plus an additional 4 miles upstream

- To find if gradually changing speed limits in space is better than changing them abruptly. Gradual implementation involves making the first half a mile either a 5 mph change or a ten mph change based on whether the total change is 10 mph or 15 mph. In other words for a 15 mph change, speed limit would be lowered by 15 mph from a location 3.5 miles upstream, but from 4.0 miles to 3.5 miles the speed limit would be lowered 10 mph only. In the abrupt case all 4 miles would be lowered 15 mph. The amount of the change is limited to a 10 mph or 15 mph change. In either case both upstream and downstream changes would be made in the same amount.
- To find if a 10mph change is better than a 15mph change is speed limits.

The combination of these factors gives rise to 3 x 2 x 2= 12 cases which are summarized in Table 5.2 (the variables in the different scenarios are spatial implementation rate, change amount, and location of the change).

Table 5.2: Different scenarios based on changing speed limit patterns(Step 1)

	Pattern of Change	Amount of Speed Change	Location of Change
Case 1	Abrupt	10 mph	Up and Downstream
Case 2	Abrupt	15 mph	Up and Downstream
Case 3	Gradual	10 mph	Up and Downstream
Case 4	Gradual	15 mph	Up and Downstream
Case 5	Gradual	10 mph	Downstream
Case 6	Gradual	15 mph	Downstream
Case 7	Abrupt	10mph	Downstream
Case 8	Abrupt	15mph	Downstream
Case 9	Gradual	10mph	Upstream
Case 10	Gradual	15mph	Upstream
Case 11	Abrupt	10mph	Upstream
Case 12	Abrupt	15mph	Upstream

The last 6 cases were not showing any significant effect and are hence not depicted in the analysis process. All the cases in Step 1 were implemented on the test case and the Best Case Scenario was identified. Figure 5.5 shows the Crash risks for the first six cases in Table 5.2 and the Base case scenario.

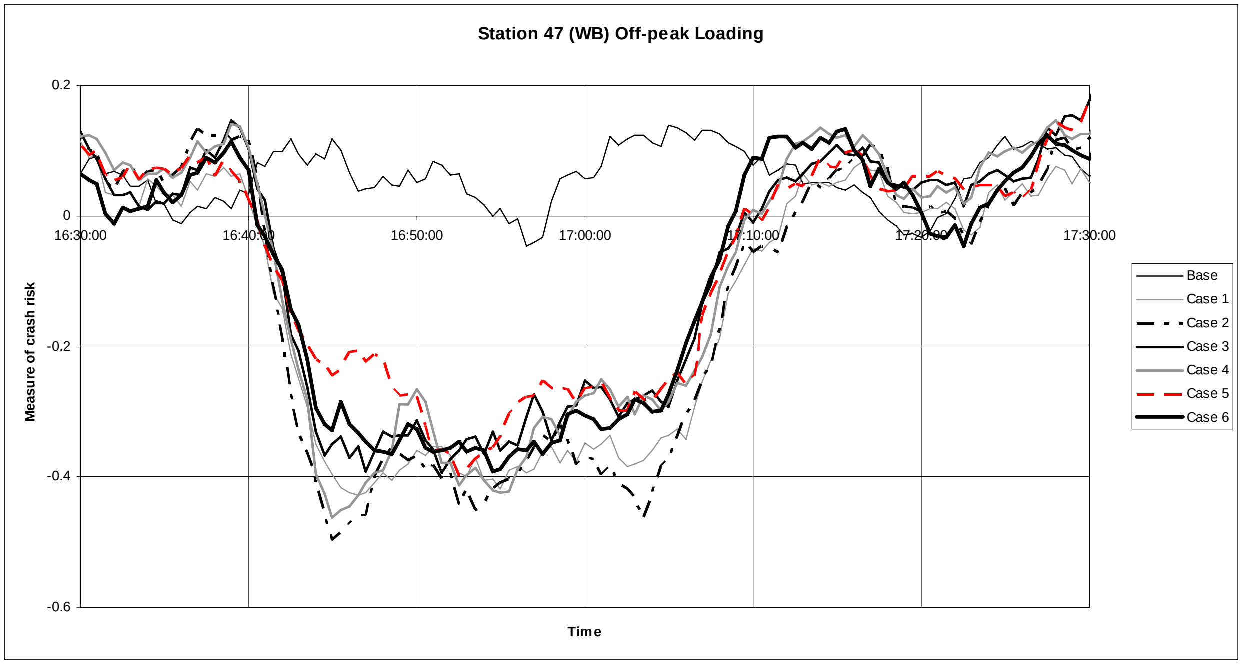


Figure 5.5: Crash potentials at Station 47 for the Pattern Scenarios.

The first set of scenarios tested compared the effect of changing speed limits by different values across one time period, as shown in Table 5.2. Figure 5.5 shows that the abrupt cases outperform the gradual cases and that 15 mph differences outperform 10 mph differences. Also, upstream decreases in speed increase the effectiveness of the VSL, especially at the beginning and ending periods of the change, making Case 2 the best case.

Step 2 is used to identify the optimal spatial location of changing speed limit signs. In this test, the best case from Step 1, viz. Case 2 is used and the location of the change in speed limit is altered to evaluate its effect. Table 5.3 shows the 3 x 3 = 9 possible cases where we can see Case 1 is the same as Case 2 of Step 1.

Table 5.3: Different scenarios based on changing distance for speed limit change (Step 2)

	Upstream Reducing Distance*	Downstream Increasing Distance*
Case 1	4 mi	5.5 mi
Case 2	6 mi	5.5 mi
Case 3	2 mi	5.5 mi
Case 4	4 mi	4 mi
Case 5	6 mi	4 mi
Case 6	2 mi	4 mi
Case 7	4 mi	2 mi
Case 8	6 mi	2 mi
Case 9	2 mi	2 mi

*-From the station under consideration

The results showed that all of the distance cases are approximately equally effective in reducing the crash potential at Station 47. Case 9, which has the shortest downstream and upstream lengths, 2 miles each, is therefore chosen as it affects the minimum length of the freeway.

Step 3 tests the temporal implementation strategy of the VSL on the best case scenario from Step 2 i.e. Case 9. The first case that is considered is the abrupt change from the standard speed limits to the new speed limits. The other cases involve changing the speed limit stepwise from the normal speed limit to the crash prevention speed limit. The step size varies by the time period between changes and the value of the speed. The speed limits can be stepped either every five minutes or every 10 minutes and the speed change be 5 mph or 10 mph. As with the Step 2, if the scenario from the earlier tests makes one or more of the cases irrelevant, then the case will be disregarded. See Table 5.4 for the scenarios that were possible and which had not been rejected in the previous steps.

Table 5.4: Different scenarios based on changing speed limits over time (Step 3)

	Time Step for Change	Speed Step for Change	Number of Time Steps
Case 1	Abrupt(Immediate)	15mph	0
Case 2	5 min	5mph	2
Case 3	10 min	5mph	2
Case 4	5 min	10 mph	1
Case 5	10 min	10 mph	1
Case 6	5 min	5 mph	1

As shown in Figure 5.6, all of the cases reduce crash potential quickly and efficiently at the detector of interest. Case 3, which has 10 minutes between changes and 5 mph steps, is selected as the best case as it shows the largest reduction in crash potential.

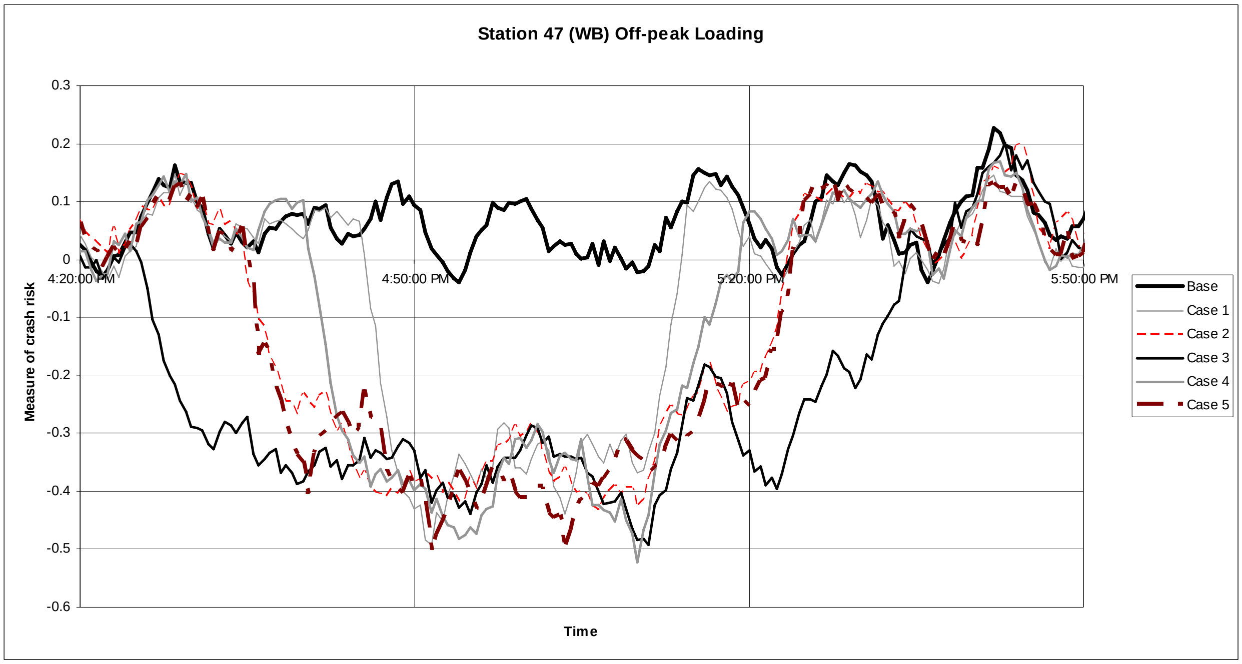


Figure 5.6: Crash potentials at Station 47 for the rate of changing speed limits scenarios.

Step 4 tests the best case from the previous step for the ideal gap distance that needs to be provided between the Station of Interest and the Downstream Location where the Speed limits can be raised again. In this Step it is assumed that in order to affect the high density traffic, the front of the high density needs to run faster than the back, thereby increasing the distance that the vehicles occupy, lowering their density. To test this, four values of the gap are tested. See Table 5.5 for the scenarios.

Table 5.5: Different scenarios based on changing the gap (Step 4)

	Gap Distance
Case 1	0 mi
Case 2	1 mi
Case 3	2 mi
Case 4	3 mi

Figure 5.7 shows the crash potential for these four cases and the Base Case Scenario. It clearly shows that Case 1, the zero gap case, is the best case. In fact, Case 1 is the only case where the VSL have an effect.

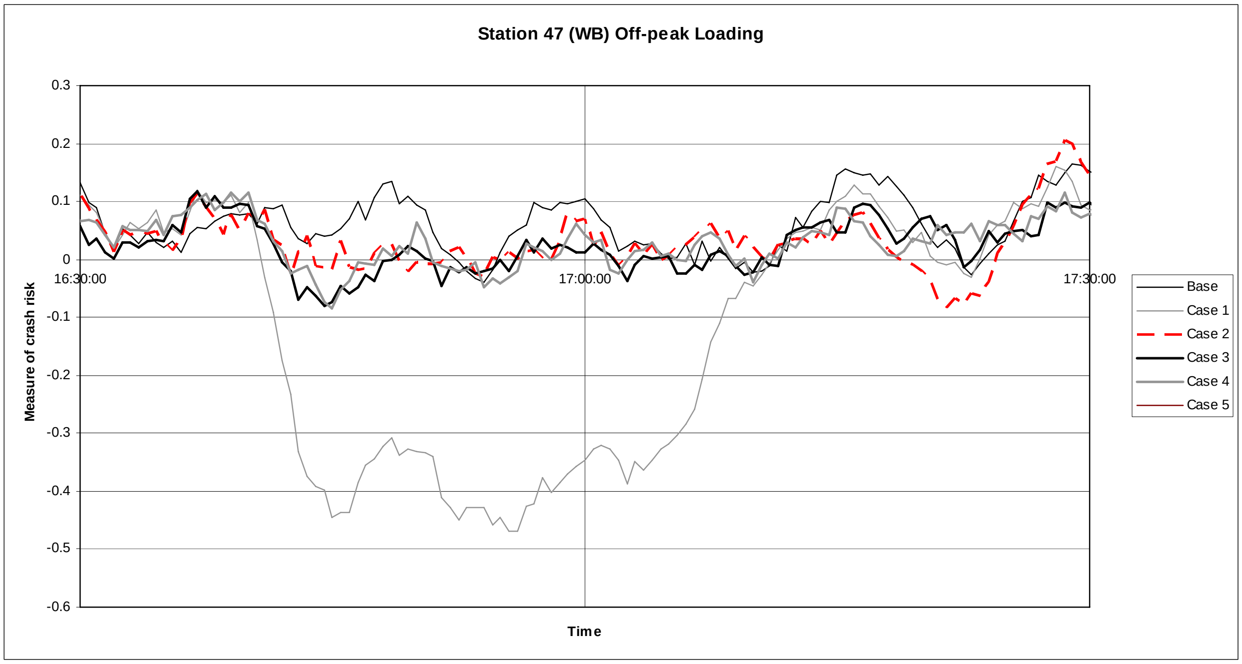


Figure 5.7: Crash potentials at the detector of interest, Station 47, for the gap scenarios.

5.9 Summary of the Safety effect of VSL

Various cases for the moderate-to-high-speed scenario have determined the best scenario for implementing VSL. In the first round of tests, the pattern of variable speed limit was tested. The rate of change in speed limit over space was either abrupt or gradual and the maximum speed change was either 15 mph or 10 mph. The results showed that an abrupt change of 15 mph produced the best result (although the 10 mph change was also effective).

In the second round of scenarios, the length upstream and downstream of the detector of interest was tested. The different lengths for both the upstream and downstream changes were 2 miles, 4 miles, and 5.5 miles. The results showed that a distance of 2 miles upstream and 2 miles downstream were adequate.

In the third round of scenarios, the rate of implementation of speed limits over time was tested. Changes in speed limit of 5 mph and 10 mph were tested along with time between changes of 5 minutes and 10 minutes. The results showed that a change of 5 mph every 10 minutes produced the best results. In the final test, the effect of gap distance on crash potential was tested. Gap distance of zero miles, 1 mile, 2 miles, and 3 miles were tested. The results showed that a gap of 0 miles produced the best results.

The best case is, therefore, an abrupt 15 mph change in speed limit both decreasing 2 miles upstream and increasing 2 miles downstream, implemented in 5 mph increments every 10 minutes. The change in crash potential can be seen in Figure 5.8.

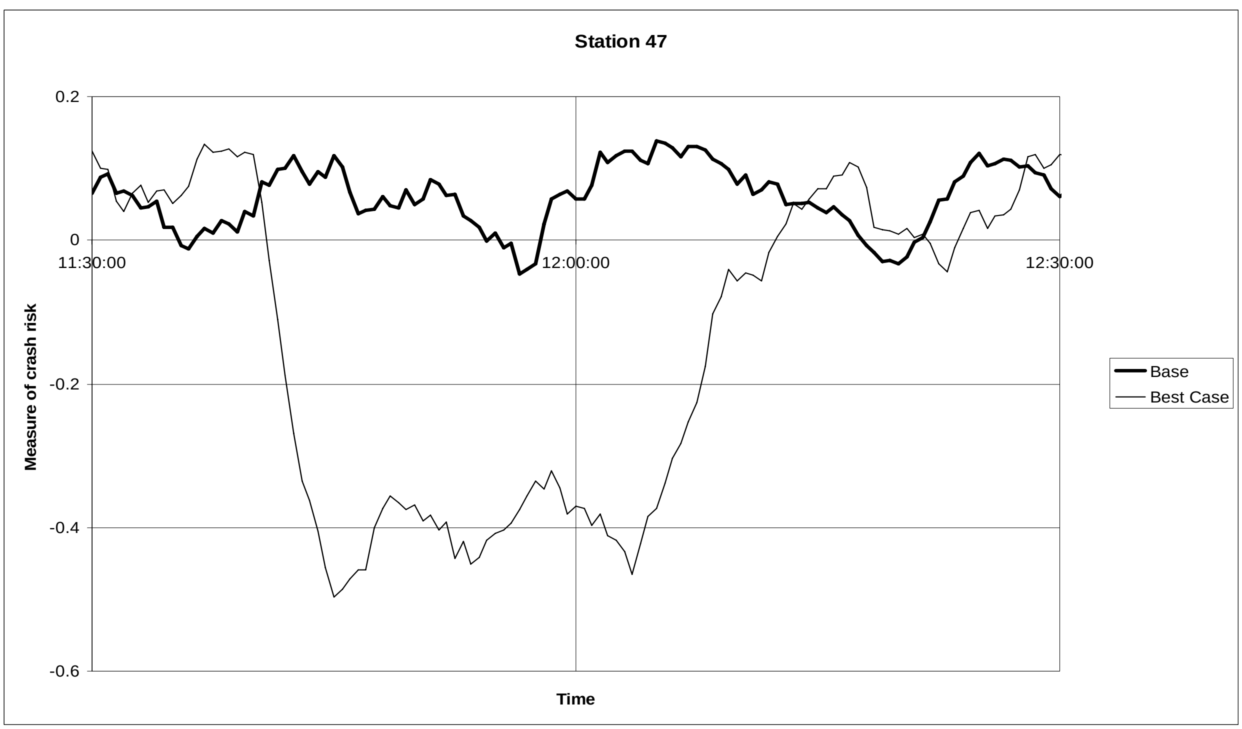


Figure 5.8: Crash potentials at the detector of interest for Best Case and Base Case

Since, the results from simulation need to be verified using a number of different random seeds, 19 such matched cases were run for the base case and the best case scenario.

After running nineteen matched cases of a base scenario and a best-case scenario the results are analyzed using a t-test. The half width is determined by multiplying the t-value by the standard deviation divided by the square root of the number of observations (19 in this case). For the differences that are significant the percent change of each are calculated. The percent change is the difference between the best case and the base case (less their half width) normalized by the base case value. The significant result can be realized for station 47 which shows a decrease in Crash likelihood of 122%.

5.10 Travel Time Analysis

Initially, 20 replications of the base case and the VSL implemented cases were run. The travel time for the entire network, from 16:30 to 17:30, was then calculated for each of the 40 runs. Eventually 140 runs (70 each) were conducted for the base and best implementation scenario, and the 95% confidence interval of the 70 differences is calculated. The effect is concluded by determining if the confidence intervals overlap. A paired t-test was performed for the 70 observations.

Because the scenarios have a matched seed number, a paired t-test can be used to build a 95% confidence interval for the different runs. Table 5.6 shows the travel time for the corridor in 70 simulation runs for the base case (no VSL) and the best scenario (using VSL). Table 6 shows that there is a significant consistent reduction in travel time, because the t-test reveals a significant negative value of travel time in the system, due to the implementation of VSL. This indicates that by changing the speed limits in the manner outlined above we did not only achieve a reduction in the crash potential but also achieved an improvement in the efficiency of the freeway. The t-test proved a significant reduction in the total travel time of the best case compared to that of the base case (t-stat = 14.81, df = 69, $p < 0.0001$). The mean difference was found to be 189.89 min.

Table 5.6: Travel time saving (minutes)

Base Scenario	Best Scenario	Difference (Best Case-BaseCase)	Base Scenario	Best Scenario	Difference (Best Case-BaseCase)
36776.3	36655	-121.3	36603.5	36456.5	-147
36868.9	36754.5	-114.4	36712.5	36563	-149.5
36510.6	36526.3	15.7	36759.9	36577.1	-182.8
36632.9	36529.3	-103.6	36696.4	36616.8	-79.6
36775.7	36763.6	-12.1	36712.6	36510.2	-202.4
36718.4	36563	-155.4	36814	36703.7	-110.3
36823.4	36644.9	-178.5	36892.2	36756.5	-135.7
36683.6	36456.4	-227.2	36856.1	36637.5	-218.6
36789.9	36612.5	-177.4	36959.4	36711.4	-248
36838.6	36628.1	-210.5	36701.5	36513.8	-187.7
36954.3	36618.8	-335.5	36757.6	36583.1	-174.5
36943	36487.8	-455.2	36834.7	36615.3	-219.4
36733.1	36598.8	-134.3	36837.3	36687.3	-150
36844.5	36645.4	-199.1	36749.2	36652.4	-96.8
36810.1	36743	-67.1	36854.9	36510.9	-344
36799.5	36720	-79.5	36814	36703.7	-110.3
36626.4	36444.7	-181.7	36892.2	36756.5	-135.7
36818.2	36795.3	-22.9	36856.1	36637.5	-218.6
36733.7	36551.4	-182.3	36959.4	36711.4	-248
36886.8	36419.8	-467	36701.5	36513.8	-187.7
36590.2	36406.5	-183.7	36757.6	36583.1	-174.5
36763.5	36534.7	-228.8	36834.7	36615.3	-219.4
36951.2	36689.4	-261.8	36837.3	36687.3	-150
36719.8	36642.7	-77.1	36749.2	36652.4	-96.8
36715.3	36444.4	-270.9	36854.9	36510.9	-344
36807.6	36529.6	-278	36714.7	36707.3	-7.4
36682.2	36526.9	-155.3	36760.5	36573.1	-187.4
36748.3	36526.8	-221.5	36884.7	36699.5	-185.2
36635.9	36468.8	-167.1	36799.5	36783.1	-16.4
36776.8	36612.7	-164.1	36678.8	36533.7	-145.1
36835.6	36673.2	-162.4	36785	36530.8	-254.2
36685.1	36591.5	-93.6	36695.5	36457.7	-237.8
36845.1	36649.3	-195.8	36819.6	36425.5	-394.1
36660.6	36503	-157.6	36799.7	36441.8	-357.9
36884.8	36475.7	-409.1	37099.7	36561.9	-537.8

5.11 Section Summary

The objective of this research was to use Micro simulation to explore an ITS strategy for improving safety on Interstate 4 (I-4), specifically by using variable speed limits (VSL). Through the implementation of variable speed limits on I-4, via simulation, a best case for improving safety was determined. Multiple scenarios with characteristic trends in them were used to generate general "rules," or best practices, that result in an optimal safe condition. To measure the crash risk separate models were used for low-speed and moderate-to-high-speed traffic conditions. To achieve these traffic regimes in the simulation environment different loading conditions may be used. These loading conditions are associated with the number of vehicles released into the network by the simulation program. Low loading (to simulate moderate-to-high-speed conditions) and high loading (to simulate low-speed conditions) were considered in this study.

The objective of reducing the risk is successfully achieved for the moderate-to-high-speed case. Through four rounds of testing a final best model was obtained. As a result, this best practice determination for existing VSL can be thought of in a new light. Instead of just using VSL before or during periods of high congestion, VSL can be thought of as applicable during off-peak periods as well. Instead of aiming to reduce congestion only, the VSL can be used to effectively reduce the hazard at certain locations. This study also concluded that by using VSL in off-peak conditions, travel time is positively affected. When implementing VSL, the recommendations of this study are as follows:

- Gradually introduce speed changes in time (5 mph every 10 minutes)
- Abruptly introduce speed changes in space (No gap distance)

- Use upstream reductions in speed and downstream increases in speed
- Changes speed limit by large values (15 mph), although a 10 mph change is also effective.
- Short upstream and downstream distances are sufficient (2 miles each)

While the low-speed case analysis did not show clear improvement, it indicated that VSL would not be advised during congested periods (i.e., peak periods). Different ITS strategies could be effective during congested periods. Based on our experience in analyzing the safety of I-4 (Abdel-Aty et al, 2004; 2005) and based on preliminary simulations, we suggest the use of ramp metering, specifically the ramp just downstream of the location where high crash potential is observed. Also, Borough's (1997) work suggests that lane changing behavior is an important factor in high congestion situations; therefore, lane changing restrictions might be an important factor that needs further exploration. The conclusion of this study provides direction towards finding the best strategies for using Variable Speed Limits as measures to improve freeway safety. A successful field deployment based upon these guidelines would open the possibility of creating and testing numerous strategies using simulations.

CHAPTER 6

CONCLUSIONS

This project investigated key elements that are related to developing a system of real-time crash prediction and mitigation on I-4. The significant traffic parameters collected by loop detector data, the possible effect of real-time weather data, and the freeway geometric elements were all investigated. Variable speed limits were investigated as ITS systems that can positively improve the safety situation on the freeway in real-time.

The analysis showed that the coefficient of variation is speed, average occupancy and the standard deviation of volume in the 5 – 10 minutes before crash occurrence are the most significant variables that could lead to crashes on the freeway.

We have obtained detailed rain fall data from 5 weather stations in Central Florida and developed a rain index based on the archived rain data to investigate whether real-time rain data would be needed for implementation. The analysis showed marginal benefit in using real-time rain data in addition to real-time traffic data to predict traffic crashes. The only constraint is the availability of weather stations in close proximity to I-4 that would provide real-time rain data in the future application.

Investigating the geometric elements that are related with crash occurrence and could be used with association with real-time traffic conditions from loop detector data, indicated that the locations of the ramps are significant.

Finally, VSL were investigated using the micro simulation model (Paramics). This investigation showed that VSL can be used to reduce the crash risk in real-time. However, the analysis showed that this is most effective only in moderate-to-high-speed conditions. Also, we have noticed the possibility that the crash risk would relocate (migrate) to their locations other than the location that we intend to treat. The strategy to use VSL for real-time safety application is still in its infancy and would require more investigation.

REFERENCES

Abdel-Aty M., Pande A., Udding N., Al-Deek H. and Radwan E. "Linking Crash Patterns to ITS-Related Archived Data", Final report submitted to Florida Department of Transportation, Nov. 2004.

Abdel-Aty, M., Uddin, N., and Pande, A. "Split Models for Predicting Multi-Vehicle Crashes during High-Speed and Low-Speed Operating Conditions on Freeways." Presented at the 84th Annual Meeting of the Transportation Research Board, January 2005, Washington D.C.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M., and Hsia, L. "Predicting Freeway Crashes Based on Loop Detector Data Using Matched Case-Control Logistic Regression." *Journal of the Transportation Research Board,* No. 1897, pp. 88-95, 2004.

Abdulhai, B., Shalaby, A., Lee, J., and Georgi, A. "Microsimulation Modeling and Impact Assessment of Streetcar Transit Priority Options: The Toronto Experience," Presented at the Transportation Research Board 81st Annual Meeting, January 2002, Washington D.C.

Bertini, R., Lindgren, R., and Tantiyanugulchai, S. "Application of PARAMICS Simulation at a Diamond Interchange," Portland State University, Transportation Research Group, Research Report PSU-CE-TRG-02-02, April 2002.

Borrough, P. "Variable Speed Limits Reduce Crashes Significantly in the U.K." The Urban Transportation Monitor. March 14, 1997.

Cheu, R., Qi, H., and Lee, D. "Mobile Sensor and Sample-Based Algorithm for Freeway Incident Detection," Presented at the Transportation Research Board 81st Annual Meeting, January 2002, Washington D.C.

Gardes, Y., May, D., Dahlgren, J., Skabardonis, A. "Freeway Calibration and Application of the PARAMICS Model," Presented at the Transportation Research Board 81st Annual Meeting, January 2002, Washington D.C.

Ha, T., Kang, J., and Park, J. "The Effects of Automated Speed Enforcement Systems on Traffic-Flow Characteristics and Crashes in Korea." ITE Journal, February 2003, pp. 28-31.

Lee C., Hellinga B. and Saccomanno F. "Assessing Safety Benefits of Variable Speed Limits." Presented at the 83rd TRB Annual Meeting, 2004. (Paper No. 04-4835).

Lee, C., Hellinga, B., and Saccomanno, F. "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.

Lee, D., Yang, X., and Chandrasekar, P. "Parameter Calibration and Application of the PARAMICS Model." Presented at the Transportation Research Boards 81st Annual Meeting, Washington D.C., 2002.

Lillard, L. A. & Panis, C. W. A. "AML multilevel multiprocess statistical software." version 2.0. Los Angeles: EconWare, 2003.

Pande, A., Abdel-Aty, M. and Hsia, L. "Spatio-temporal variation of risk preceding crash occurrence on freeways." presented at the 84th Annual Meeting of Transportation Research Board, Washington, D.C., 2005.

Peltola, H. "The Role of Seasonal Speed Limits in Speed Management." 15th ICTCT Workshop. 1997.

Pilli-Sivola, Y. "State of the Art in Finland Concerning RWIS and Variable Message Signs." Finnish National Road Administration. 2 February 2000.

Placer, J. "Fuzzy Variable Speed Limit Device Modification and Testing – Phase II." Arizona Department of Transportation. July 2001.

Portaankorva, P. "Road Weather and Traffic Data in Traffic Management." SIRWEC Sapporo, Japan. 2002.

Quadstone Limited, "PARAMICS Modeller Version 4.1 Reference Manual." Edinburgh, U.K., 2002.

Rämä, P. "Effects of Weather-Controlled Variable Speed Limits and Warning Signs on Driver Behavior." In Transportation Research Record 1689, TRB, National Research Council, Washington, D.C., 1999, pp. 53- 59.

Senn, L. "Variable Speed Limit and In-Vehicle Signing Operational Test." http://www.wsdot.wa.gov/biz/atb/html/ta.html. Accessed on April 28, 2004.

Sisiopiku, V. "Variable Speed Control: Technologies and Practices." Transportation Research Board 80th Annual Meeting. 2001.

Smulders, S. "Control of freeway traffic flow by variable speed signs. Transportation Research, Vol. 24B, p. 111-132, 1990.

Stewart, P. "M8 Paramics Ramp Metering Assessment," Executive Summary form Scottish Executive Development Department, 2001, www.scotland.gov.uk/library3/transport/m8paramics.pdf, accessed on February 20, 2004

Trapp, R. "Microscopic Traffic Flow Modeling of Large Urban Networks -Approach and Techniques at the Example of the City of Cologne," Presented at the Transportation Research Board 81st Annual Meeting, January 2002, Washington D.C.