ASSESSING CRASH OCCURRENCE ON URBAN FREEWAYS USING STATIC AND DYNAMIC FACTORS BY APPLYING A SYSTEM OF INTERRELATED EQUATIONS

by

RAJASHEKAR PEMMANABOINA
B.E (Hons), Birla Institute of Technology and Science, Pilani, India, 2002

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Civil and Environmental Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2005

ABSTRACT

Traffic crashes have been identified as one of the main causes of death in the US, making road safety a high priority issue that needs urgent attention. Recognizing the fact that more and effective research has to be done in this area, this thesis aims mainly at developing different statistical models related to the road safety. The thesis includes three main sections: 1) overall crash frequency analysis using negative binomial models, 2) seemingly unrelated negative binomial (SUNB) models for different categories of crashes divided based on type of crash, or condition in which they occur, 3) safety models to determine the probability of crash occurrence, including a rainfall index that has been estimated using a logistic regression model. The study corridor is a 36-mile stretch of Interstate 4 in Central Florida. For the first two sections, crash cases from 1999 through 2002 were considered.

Conventionally most of the crash frequency analysis model all crashes, instead of dividing them based on type of crash, peaking conditions, availability of light, severity, or pavement condition, etc. Also researchers traditionally used AADT to represent traffic volumes in their models. These two cases are examples of macroscopic crash frequency modeling. To investigate the microscopic models, and to identify the significant factors related to crash occurrence, a preliminary study (first analysis) explored the use of microscopic traffic volumes related to crash occurrence by comparing AADT/VMT with five to twenty minute volumes immediately preceding the crash. It was found that the volumes just before the time of crash occurrence proved to be a better predictor of crash frequency than AADT. The results also showed that road curvature, median type, number

of lanes, pavement surface type and presence of on/off-ramps are among the significant factors that contribute to crash occurrence.

In the second analysis various possible crash categories were prepared to exactly identify the factors related to them, using various roadway, geometric, and microscopic traffic variables. Five different categories are prepared based on a common platform, e.g. type of crash. They are: 1) Multiple and Single vehicle crashes, 2) Peak and Off-peak crashes, 3) Dry and Wet pavement crashes, 4) Daytime and Dark hour crashes, and 5) Property Damage Only (PDO) and Injury crashes. Each of the above mentioned models in each category are estimated separately. To account for the correlation between the disturbance terms arising from omitted variables between any two models in a category, seemingly unrelated negative binomial (SUNB) regression was used, and then the models in each category were estimated simultaneously. SUNB estimation proved to be advantageous for two categories: Category 1, and Category 4. Road curvature and presence of On-ramps/Off-ramps were found to be the important factors, which can be related to every crash category. AADT was also found to be significant in all the models except for the single vehicle crash model. Median type and pavement surface type were among the other important factors causing crashes. It can be stated that the group of factors found in the model considering all crashes is a superset of the factors that were found in individual crash categories.

The third analysis dealt with the development of a logistic regression model to obtain the weather condition at a given time and location on I-4 in Central Florida so that this information can be used in traffic safety analyses, because of the lack of weather monitoring stations in the study area. To prove the worthiness of the weather information

obtained from the analysis, the same weather information was used in a safety model developed by Abdel-Aty et al., 2004. It was also proved that the inclusion of weather information actually improved the safety model with better prediction accuracy.

ACKNOWLEDGMENTS

I would like to begin by thanking my mother, my father, and my brothers for giving me the courage and strength I needed to complete my thesis. They had always been a constant inspiration to me in every walk of my life. They instilled in me the confidence and drive to pursue my Masters, which molded my career a lot. I would also like to thank my friends, Aparna, Anurag, Jai, Nishanth, Piyush, Ravi, Sai Srinivas, Sandeep, and Vidhya, who encouraged me to do my best. In particular, I will be always indebted to Anurag, for helping me with any topic of my thesis. I would like to specially thank Aparna (very enthusiastic), Nishanth (great friend), Ravi (very patient and helpful), Sai Srinivas (simply great!!) and Sandeep (jovial), for their constant support and encouragement, and making my stay at UCF a memorable one.

I would like to express my sincere thanks to Chilakammari Venkata Srinvasa Ravi Chandra, who helped me in every aspect of my research. He was always there when I was in trouble with his expertise and knowledge. Without his constant support, I would not have completed my thesis. He is simply awesome, and I shall cherish his friendship for many more years to come.

I am very thankful to the members of my thesis committee, Dr. Radwan, and Dr. Nizam Uddin. Finally, my deepest appreciation goes to my thesis advisor and guide, Dr. Abdel-Aty, for his constant encouragement and expertise right from the first semester. He incessantly and persuasively conveyed a spirit of adventure in regard to research. Without his guidance and persistent help this thesis would not have been possible. He was always very patient and made my research experience at UCF worthwhile.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGMENTS v
TABLE OF CONTENTS vi
LIST OF FIGURES ix
LIST OF TABLES x
1 INTRODUCTION 1
1.1 Background 1
1.2 Problem Description and Objectives of Study 2
2 ASSESSING CRASH OCCURRENCE ON URBAN FREEWAYS USING STATIC AND DYNAMIC FACTORS 7
2.1 Introduction 7
2.2 Related Studies 8
2.3 Data Collection 9
2.3.1 Traffic and Crash Data 9
2.4 Geometric and Roadway Characteristics 11
2.5 Data preparation for Different Traffic Volume Measures 13
2.5.1 AADT and VMT 13
2.5.2 Peak Fifteen Minute Volumes 14
2.5.3 Five to Twenty Minute Average Volumes just before the Crash Occurrence 14
2.6 Background for Introducing Disaggregate Volume Measures 15
2.7 Model Framework 18
2.8 Crash Frequency Models 19
2.9 Model Estimation and Results 20
2.10 Discussion of Results 24
2.11 Conclusions 25
3 CRASH FREQUENCY MODELING FOR DIFFERENT CRASH CATEGORIES USING SEEMINGLY UNREALTED NEGATIVE BINOMIAL REGRESSION 28
3.1 Introduction 28
3.2 Data Description 30
3.2.1 Traffic and Crash Data 30
3.2.2 Geometric and Roadway Characteristics 31
3.3 Data Preparation 31
3.4 Preliminary Data Analyses 35
3.4.1 Category 1: Types of Crashes 36
3.4.2 Category 2: Peak and Off-peak Period Crashes 38
3.4.3 Category 3: Crashes based on Pavement Condition 41
3.4.4 Category 4: Crashes based on Availability of Daylight 42
3.4.5 Category 5: Crashes based on Injury Occurrence in a Crash 45

3.5 Categorical Data Analyses 47
3.5.1 Type of Crash and Traffic Condition 48
3.5.2 Type of Crash and Availability of Daylight 50
3.5.3 Crash Type and Injury Involvement 51
3.5.4 Crash Type and Pavement Condition 52
4 MODELING APPROACH FOR SEEMINLGLY UNRELATED NEGATIVE BINOMIAL MODELS 54
4.1 Modeling Approach 54
4.2 Seemingly Unrelated Regression 56
4.3 Estimation via Generalized Least Squares Estimation 57
4.4 Development of SUR Models using aML Software 59
4.5 Model Estimation and Results 60
4.5.1 Category 1 64
4.5.1.1 Individual Multiple Vehicle Crash Model 64
4.5.1.2 Individual Single Vehicle Crash Model 65
4.5.1.3 Seemingly Unrelated Negative Binomial Model for Multiple and Single Vehicle Crashes 66
4.5.1.4 Discussion of Results 69
4.5.2 Category 2 70
4.5.2.1 Individual Peak Period Crash Model 71
4.5.2.2 Individual Off-peak Period Crash Model 72
4.5.2.3 Seemingly Unrelated Negative Binomial Model for Peak and Off-peak Period Crashes 73
4.5.2.4 Discussion of Results 75
4.5.3 Category 3 77
4.5.3.1 Individual Dry Pavement Crash Model 77
4.5.3.2 Individual Wet Pavement Crash Model 79
4.5.3.3 Seemingly Unrelated Negative Binomial Model for Dry and Wet Pavement Crashes 80
4.5.3.4 Discussion of Results 82
4.5.4 Category 4 83
4.5.4.1 Individual Daytime Crash Model 84
4.5.4.2 Individual Dark Hour Crash Model 85
4.5.4.3 Seemingly Unrelated Negative Binomial Model for Day and Dark Hour Crashes 86
4.5.4.4 Discussion of Results 88
4.5.5 Category 5 90
4.5.5.1 Individual PDO Crash Model 91
4.5.5.2 Individual Injury Crash Model 92
4.5.5.3 Seemingly Unrelated Negative Binomial Model for PDO and Injury Crashes 93
4.5.5.4 Discussion of Results 95
4.6 Measurement of Goodness-of-fit 96
4.7 Conclusions 99

5 APPLICATION OF LOGISTIC REGRESSION MODEL TO OBTAIN RAINFALL INFORMATION ON INTERSTATE-4 IN CENTRAL FLORIDA 100
5.1 Introduction 100
5.2 Background and Data Collection 101
5.3 Methodology and Data Preparation 105
5.3.1 Dependent Variable 106
5.3.2 Independent Variables 107
5.4 Model Development 112
5.5 Model Evaluation 118
5.6 Model Application 119
5.6.1 Goodness-of-fit 122
5.6.2 Prediction Accuracy 123
5.7 Adjustment of Estimated Probabilities 125
5.8 Conclusions 126
6 CONCLUSIONS 127
APPENDIX A 132
LIST OF REFERENCES 138

LIST OF FIGURES

Figure 2-1: Map of I-4 in the study area 10
Figure 2-2: Influence segment for each loop detector station 12
Figure 2-3: Average AADT values at different Crash Stations 16
Figure 2-4: Peak Fifteen Minute Volumes at different Crash Stations 17
Figure 2-5: Five Minute Crash Volumes at different Crash Stations 18
Figure 3-1: Frequency distribution of multiple and single vehicle crashes 37
Figure 3-2: Frequency distribution of peak and off-peak period crashes 40
Figure 3-3: Frequency distribution of peak and off-peak period crashes 42
Figure 3-4: Crash frequency of daylight and dark hour crashes at different stations 44
Figure 3-5: Crash frequency of PDO and injury crashes at different stations 46
Figure 3-6: Crash frequency of multiple vehicle crashes by peak and off-peak period ... 49
Figure 3-7: Crash frequency of single vehicle crashes by peak and off-peak period 50
Figure 5-1: Average Injury and Fatal Crashes in Adverse Weather Conditions (Goodwin, 2002) 101
Figure 5-2: Map showing locations of the five weather stations surrounding Interstate 4 in Central Florida. 105
Figure 5-3: Scree Plot from Principal Component Analysis 116

LIST OF TABLES

Table 2-1: Code Sheet for all the variables used in the Model 21
Table 2-2: Parameter estimates of significant factors for the First Main Model 22
Table 2-3: Parameter estimates of significant factors for the Third Main Model 23
Table 2-4: Comparison of Standard Errors between First and Third Main Models 24
Table 3-1: Frequency of different types of crashes 36
Table 3-2: Frequency table by peak and off-peak period 39
Table 3-3: Crash frequency table for dry and wet pavement crashes 41
Table 3-4: Crash frequency table for crashes based on availability of sunlight 43
Table 3-5: Crash frequency table for injury and PDO crashes 45
Table 3-6: Distribution of multiple and single vehicle crashes by peak and off-peak period 48
Table 3-7: Frequency table for multiple and single vehicle crashes by lighting condition 51
Table 3-8: Frequency table for multiple and single vehicle crashes by injury occurrence51
Table 3-9: Frequency table for multiple and single vehicle crashes by pavement condition 52
Table 4-1: Code Sheet for all the variables used in the Model 62
Table 4-2: Estimation results for individual multiple vehicle crash model 65
Table 4-3: Estimation results for individual single vehicle crash model 66
Table 4-4: SUNB model estimation results for multiple vehicle crash model 67
Table 4-5: SUNB model estimation results for single vehicle crash model 67
Table 4-6: Model estimation results contd. 67
Table 4-7: Comparison of standard errors between individual and SUNB multiple vehicle crash models 68
Table 4-8: Comparison of standard errors between individual and SUNB for single vehicle crash models 68
Table 4-9: Estimation results for individual peak period crash model 71
Table 4-10: Estimation results for individual off-peak period crash model 72
Table 4-11: SUNB model estimation results for peak period crash model 73
Table 4-12: SUNB model estimation results for off-peak period crash model 74
Table 4-13: Comparison of standard errors between individual and SUNB models 74
Table 4-14: Estimation results for individual dry pavement crash model 78
Table 4-15: Estimation results for individual dry pavement crash model 79
Table 4-16: SUNB model estimation results for dry pavement crash model 80
Table 4-17: SUNB model estimation results for wet pavement crash model 80
Table 4-18: Estimation results for individual day time crash model 84
Table 4-19: Estimation results for individual dark hour crash model 86
Table 4-20: SUNB model estimation results for day time crash model 87
Table 4-21: SUNB model estimation results for dark hour crash model 87
Table 4-22: SUNB model estimation results contd. 87
Table 4-23: Comparison of standard errors for day time crash model 88

Table 4-24: Comparison of errors for dark hour crash model 88
Table 4-25: Estimation results for individual PDO crash model 91
Table 4-26: Estimation results for individual injury crash model 92
Table 4-27: SUNB model estimation results for PDO crash model 93
Table 4-28: SUNB model estimation results for injury crash model 94
Table 4-29: Comparison of standard errors between individual and SUNB models 94
Table 4-30: Comparison of standard errors between individual and SUNB models 95
Table 4-31: Goodness-of-fit statistics for different crash categories 98
Table 5-1: Number of crashes occurred during rain during 1999 – 2001 on I-4 102
Table 5-2: Sample weather information extracted from the crash database 107
Table 5-3: Sample information with dependent and independent variables used in the model 108
Table 5-4: Geographical co-ordinates of the Crash Stations 109
Table 5-5: Geographical Co-ordinates of the Weather Stations 110
Table 5-6: Order of Weather Stations based on the distance from Crash stations 111
Table 5-7: Rainfall Information at Weather Station Avalon 111
Table 5-8: Chi-Square test of Independence of Variables 113
Table 5-9: Results from Principal Component Analysis 115
Table 5-10: Results from Principal Component Analysis 115
Table 5-11: Logistic Regression Model Results 117
Table 5-12: Logistic Regression Model Results 117
Table 5-13: Quantiles for the rain index values 119
Table 5-14: Classification table for the test data 119
Table 5-15: Model fit statistics for the safety model without “Weather” variable 121
Table 5-16: Parameter estimates of the safety model without “Weather” variable 121
Table 5-17: Model fit statistics for the safety model with “Weather” variable 121
Table 5-18: Parameter estimates of the safety model with “Weather” variable 121
Table 5-19: Measures of association between the predicted probabilities and observed responses for the model without “Weather” 124
Table 5-20: Measures of association between the predicted probabilities and observed responses for the model with “Weather” 124

1 INTRODUCTION

1.1 Background

Road safety has increasingly been a vital topic for discussion, as traffic crashes have been identified as one of the top 10 causes of death in the United States of America (The World Almanac and Book of Facts, 1996). They are climbing up in the list of death causes, from No. 9 in 1999 to an estimated No. 3 in 2020. According to U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA) release on Highway Fatalities in USA, for 2003, an estimated highest number of people were killed in traffic crashes since 1990 (NHSTA, 2004). NHTSA also estimated that highway crashes cost society $230.6 billion per year or an average of $820 for every person. The yearly economic costs of road traffic crashes estimated by NHSTA are provided below.

- $61 billion in lost workplace productivity
- $20.2 billion in lost household productivity
- $59 billion in property damage
- $32.6 billion in medical costs
- $25.6 billion in travel delay costs.

Most part of the burden is on people not directly involved in the crash in the form of travel delay, property damage and medical costs. These figures reveal the seriousness of the problem and the need for immediate road safety measures. At the same time, it cannot be ruled out that safety improvement on roadways is neglected. There has been a substantial growth in the field of road safety research during the last decade.

While the traffic demand is rising by leaps and bounds day after day, there are limited resources available for safety research. Therefore cautious use of existing

resources, while striving for improved research in road safety field is necessary. Having recognized the gravity of the topic, it is very important that the attributes of crash occurrence are well understood. Drivers while traveling interact with other vehicles; pedestrians, roadway, and surrounding environment and these interactions tend to be very intricate. A clash among the drivers and any of the other elements is the cause for crash occurrence. So a careful comprehension of these elements is the key to tackle the problem. Based on the saying, to err is human, it is important to realize that driver behavior always plays an important role in crash occurrence. But the work of other factors like the roadway geometrics, traffic characteristics, and environmental factors cannot be ruled out. Since it is impracticable to control or predict the driver behavior, engineers have to make every effort to improve the factors, which can be controlled. So it is judicious to work on the factors that roadway designers can control, which on improvement might reduce the driver mistakes.

1.2 Problem Description and Objectives of Study

Road traffic safety can include very extensive array of research areas, of which crash frequency modeling is a crucial and vital component. Essentially crash frequency modeling is done to enumerate the relationship between observed crash count and existing geometric, roadway, and traffic conditions at a given stretch of a roadway. Till now many studies have been carried out in modeling crash frequencies for a variety of roads, with different factors associated with crashes such as traffic volumes, geometric characteristics and environmental factors. Garber and Ehrhart (2000) combined traffic

and geometric factors, and developed crash rate models to deal with the inconsistency of results when traffic or geometric factors are considered individually.

Conventionally most of the crash frequency models used all crashes instead of dividing them based on type of crash, peaking conditions, availability of light, severity, or pavement condition etc. Also researchers traditionally used AADT to represent traffic volumes in their models. These two cases are examples of macroscopic crash frequency modeling. Increasingly, researchers are moving towards microscopic crash analysis, which includes splitting the crashes based on type of crash (Persaud and Macsui, 1995), or using hourly volumes as one of the traffic variables. These studies used microscopic measures such as hourly volumes to cope with the uncertainty in the measurement of AADT values and incapability of this aggregate factor in capturing accurate traffic flow variations (Garber and Wu, 2001, Pasupathy et al., 2000).

Apart from identifying the factors directly related to the crash occurrence, research has also been done to identify distinctive factors related to crashes. For instance, Polanis, 1995 has shown that majority of the fatal crashes happen during dark hours. So to accurately determine the factors causing different crashes, there is a need to split the crashes into various possible logical categories. For instance, multiple and single vehicle crashes might have completely different causal factors. But just categorizing the crashes might not be sufficient. There is also a necessity to include microscopic or disaggregated data in crash frequency analysis to better identify the factors related to crash occurrence. The existing crash frequency models include single variable and multivariate deterministic models, stochastic multivariate models, and artificial neural networks (Garber and Wu, 2001). Multiple regression techniques could not explain the discrete

nature of crash occurrence. Researchers, therefore, started applying stochastic modeling methods such as Poisson and Negative Binomial regression techniques to overcome the problems of multiple regression techniques.

This thesis report can be divided into three main analyses dealing with two crash frequency model developments and a logistic regression model for acquiring weather information on the I-4 study corridor. The first study (second chapter in the thesis) describes the crash frequency model developed considering all crashes that happened on a 36-mile stretch of Interstate 4 for the years 1999 through 2002. In the process of this model development, innovative way of using microscopic traffic volume measures was attempted and these microscopic traffic volume measures were compared with AADT and VMT usage while including various roadway and geometric variables. A negative binomial model was developed for this purpose, after assessing the best among Poisson and negative binomial models.

The second study (third and fourth chapters' in the thesis) makes an effort to evaluate another microscopic crash frequency modeling which becomes the next logical extension of the study explained in first chapter. The same data was used for this study. The models developed in this chapter deal with splitting the crashes into various categories and using various microscopic traffic factors including speed factors to gain more efficiency in crash frequency modeling. Although individual negative binomial models can be developed for each crash category, the models tend to lose their efficiency with erroneous parameter estimates as error terms might be correlated across the equations. Therefore, seemingly unrelated negative binomial regression technique was utilized to estimate the models. These models use microscopic traffic measures, which

include statistical outputs of speed, and volume values extracted from the dual loop detectors installed on Interstate 4. Data collection and preparation was an important part in both studies. Archived loop detector data obtained from the University of Central Florida's data warehouse, and crash data from Florida Department of Transportation was used in this endeavor.

The third analysis (fifth chapter in the thesis) mainly deals with the development of a logistic regression model to obtain the weather condition at a given time and location on the I-4 study corridor in Central Florida so that this information can be used in traffic safety analyses. In the study area there are no weather monitoring stations located on I-4, which can provide the exact rainfall information at a desired time and location. Alternatively the Florida crash database provides the exact weather condition at the time of crash on I-4. Many safety studies use only the crash cases in their analysis; some safety analyses use not only the crash cases on a particular roadway, but also crash and non-crash cases in their analysis. Information on such analysis can be obtained from Abdel-Aty et al. (2004). For instance a safety study may use the binary logit model with a response variable containing both crash and non-crash cases. Now the task is to obtain the weather condition for the non-crash cases. Essentially the aim of the third chapter is to obtain weather information at a particular time and location on I-4 other than the time of crash occurrences.

The research objectives of this thesis can be summarized as follows:

- Investigate different traffic volume forms to account for the best form to be used in crash frequency analysis
- Identify the significant factors that affect crash frequencies on freeways
- Address the problem of correlation between the error terms, when the crashes are divided into different logical categories (for instance, single and multiple vehicle crashes)
- Model crash frequencies for different types of crash categories
- Investigate how to account for rain in modeling the probability of crashes, while the analysis includes both crash and non-crash cases

2 ASSESSING CRASH OCCURRENCE ON URBAN FREEWAYS USING STATIC AND DYNAMIC FACTORS

2.1 Introduction

There is a need to make the roads safer, not only to save billions of dollars but most importantly thousands of lives. One of the main ways to reduce crashes, when there is less of a human fault is by identifying high risk locations on roadways. One of the effective ways in identifying a hazardous location and the factors contributing to the occurrence of crashes is by modeling the frequency of crashes at that location. Till now many studies have been carried out in modeling crash frequencies for a variety of roads, with different factors associated with the crashes such as traffic volumes, geometric characteristics and environmental factors. The existing models include single variable and multivariate deterministic models, stochastic multivariate models, and artificial neural networks (Garber and Wu, 2001). Multiple regression techniques could not explain the discrete nature of crash occurrence. Researchers, therefore, started applying stochastic modeling methods such as Poisson and Negative Binomial regression techniques to overcome the problems of multiple regression techniques. Garber and Ehrhart (2000) combined traffic and geometric factors, and developed crash rate models to deal with the inconsistency of results when traffic or geometric factors are considered individually. But there has been limited research conducted into identifying the exact traffic volume measure to be used. Some of the studies used hourly volumes to cope with the uncertainty in the measurement of AADT values and incapability of this aggregate factor in capturing the exact traffic flow variations (Garber and Ehrhart, 2000; Pasupathy et al, 2000). The present study evaluates the use of different traffic volume measures

combined with geometric and roadway characteristics in crash frequency modeling to establish a volume measure that has the ability to capture the exact traffic flow circumstances causing the crashes. The different traffic volume measures studied are 1) AADT and VMT values, 2) peak 15-minute volumes taken for a typical day, and 3) average five to twenty minute traffic volumes obtained just before crashes. Poisson and Negative Binomial regression techniques were tried and the best between these two techniques was used to model the crash frequencies and identify the significant traffic volume measure, geometric and roadway factors that affect the occurrence of crashes on a 36-mile stretch of I-4 in the state of Florida for a study period of four years. This model is referred to as overall model for future references in chapter 4.

2.2 Related Studies

Previous work in modeling crashes identified various factors causing roadway crashes, their relationship to the crash occurrence, and different modeling methodologies. Okamoto and Koshi (1989) first suggested the stochastic nature of the occurrence of road crashes. Garber and Joshua (1990) developed the Poisson models for large truck crashes. Oh et al. (2001) applied non-parametric Bayesian approach to quantify the measures of crash likelihood using real-time traffic data from inductive loop detectors, which showed the statistical importance of dynamic variables such as traffic volumes. Lee et al. (2002) examined traffic flow characteristics that lead to crashes on urban freeways and referred to them as "crash precursors". They developed a log-linear model relating crash frequency to selected crash precursors and showed that the use of real-time traffic data is promising in predicting crash potential on freeways. Shankar et al. (1995) developed a

Negative Binomial crash frequency model based on roadway geometrics, weather and other seasonal effects. Metcalf et al. (1999) studied the relationship between various measures of traffic speed and crash rate in UK and Bahrain by developing a Poisson model. Recently Kockelman and Kweon (2004) used fixed-effects and random effects Poisson and Negative Binomial regression techniques for modeling fatal crashes, injury crashes and property-damage-only (PDO) crashes. Washington et al. (2004) separated crash data into fatal, injury and property damage types and conducted simultaneous estimation of the models with Negative Binomial regression to account for correlation of the error terms across the models.

2.3 Data Collection

2.3.1 Traffic and Crash Data

The objective of this study is to analyze various traffic, roadway and geometric factors related to crash occurrence on Interstate 4 in Central Florida. The I-4 corridor (shown in the Figure 2-1, Map of I-4), considered for the present study is a 36 mile stretch roadway from US-1792 in the west to Lake Mary in the east. The traffic data used in this study was collected from the dual loop detectors installed on I-4. A total of 69 loop detectors, ranging from 2-71 provide average speed, volume and average occupancy (percent of time a loop detector is occupied by vehicles) for every 30 seconds, throughout the year. These values are measured for each lane on I-4 in both directions, approximately spaced at half a mile apart. Each direction is separate leading to a total of 138 loop detector stations. This data is available through the data warehouse at the University of Central Florida.

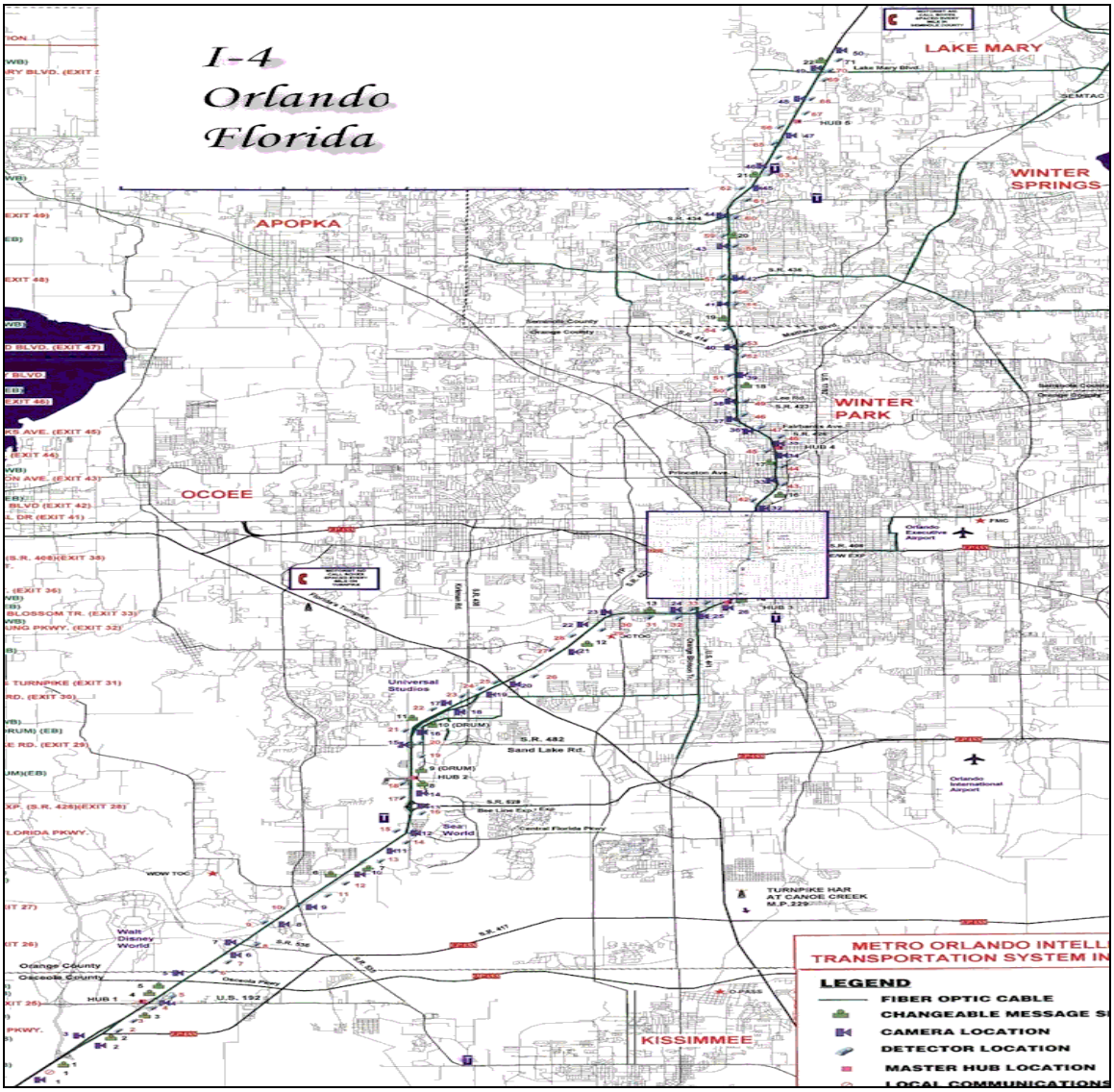

Figure 2-1: Map of I-4 in the study area

The crash data was obtained from Florida Department of Transportation crash database for the same 36-mile stretch of I-4 which includes Orange, Osceola and Seminole counties in Florida. A total of 3146 crashes that happened along this stretch were collected for a span of four years from 1999 through 2002. The FDOT database

provides the milepost for each crash which is generally the distance between the crash location and the starting of the county line. In the same way, mileposts for all the loop detector stations are also established. For the study purposes, the nearest loop detector station to the crash location is considered as the station of the crash. The time of crash occurrence was estimated by a methodology developed using shockwave speed and rule based methods (Abdel-Aty et al., 2004). The methodology estimates the speed of the backward forming shockwave resulting from the crash. The difference between times of shockwave arrival at the two adjacent stations located immediately upstream of the crash location was used. Since the milepost of all loop detectors on I-4 was known accurately, distance between the two detectors could be used to get the shockwave speed. Once the shockwave speed is known it is not difficult to determine the crash time, using the milepost of crash location (also known from the FDOT crash database). Accurate crash time determination was important for the present study as volumes aggregated at different time levels preceding the crash occurrence were used for model fitting. The methodology uses the loop data immediately upstream of the crash station.

2.4 Geometric and Roadway Characteristics

A total of 9 different geometric and roadway factors were considered for all the 138 loop detector stations in both directions. They include radius of the freeway section, number of lanes, median type, median width, pavement index, 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 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 2-2.

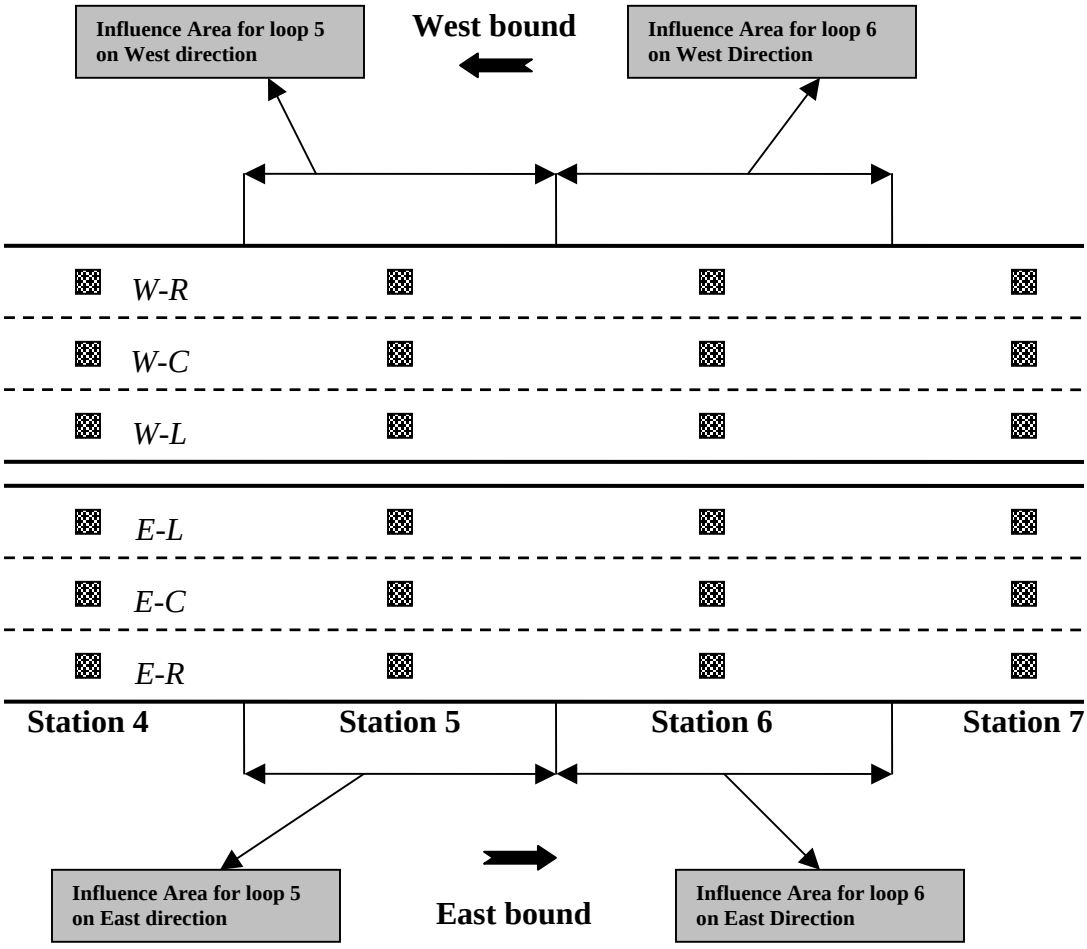

Figure 2-2: Influence segment for each loop detector station

Figure 2-2 also provides a visual representation of instrumented Interstate 4 with installed dual loop detectors under the freeway in 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 selected section of I-4.

2.5 Data preparation for Different Traffic Volume Measures

Various traffic volume measures were evaluated in the study. They include, static/aggregate measures such as the AADT and VMT values, and disaggregate measures such as peak fifteen minute volumes and five to twenty minute average volumes just before the crash occurrence.

2.5.1 AADT and VMT

Two types of static/aggregate measures, i.e. AADT and VMT were used in the study. The AADT values at each loop were obtained from the "Florida Traffic Information" database for the years 1999 through 2002. For modeling purposes, the average AADT values for these four years were taken. As explained earlier, the loop detectors in the 36-mile stretch are spaced at approximately half a mile (the distance between loops is not consistent). The influence area for each loop is defined as the sum of half the distances between that loop and the loops on each side. The influence area for each loop signifies the fact that the model considers the crash occurrence in the influence area of the loops, taking crash station or loop station as the center point. Figure 2-2 provides a sample view of the influence segment for a loop station. The Vehicle Miles Traveled (VMT) values were determined by multiplying the AADT values for each loop with the corresponding influence area's length.

2.5.2 Peak Fifteen Minute Volumes

A traffic volume measure, which would represent the intensity of traffic flow at each loop detector station, was considered. For this reason, the peak fifteen-minute volumes at each station were obtained from the archived loop detectors' data for a typical day in a month representing general traffic conditions. Volumes at 30-second level for each crash station were taken for all Wednesdays in the month of February 2002 and the 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. 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.

2.5.3 Five to Twenty Minute Average Volumes just before the Crash Occurrence

As explained in data collection section, once the time and location of crashes were identified, aggregate five to twenty minute volumes before each crash occurrence were prepared for the study period. After the crash frequencies were determined for each station in the four years study period, an average volume measure at each crash station was calculated. For example, to obtain a five-minute average volume for each station, the total of five-minute aggregate volumes before all crashes that occurred within the influence area of a particular loop detector station is divided by the total number of crashes for the four years. Similarly the same approach was repeated for ten, fifteen and twenty minute increments. The five, ten, fifteen and twenty minute average volumes will be referred to as "crash volumes" for future reference in this paper. Crash volumes

ranging from five to twenty minutes were considered for the study, to test the crash volume duration accurately representing the traffic flow.

2.6 Background for Introducing Disaggregate Volume Measures

The purpose for introducing disaggregate or microscopic measures of volume can be linked to the fact that AADT or VMT values signify a static or aggregate level of traffic volume form. At such an aggregate level, there is a high chance of accuracy being compromised. Two freeway sections having the same AADT values can have different crash frequencies. The change in crash frequency cannot be attributed only to the roadway geometrics. This fact was reinforced by previous traffic safety studies (Garber and Ehrhart, 2000; Shankar et al., 1995). Figure 2-3 shows the plot of AADT values at different crash stations in the East and West bound directions.

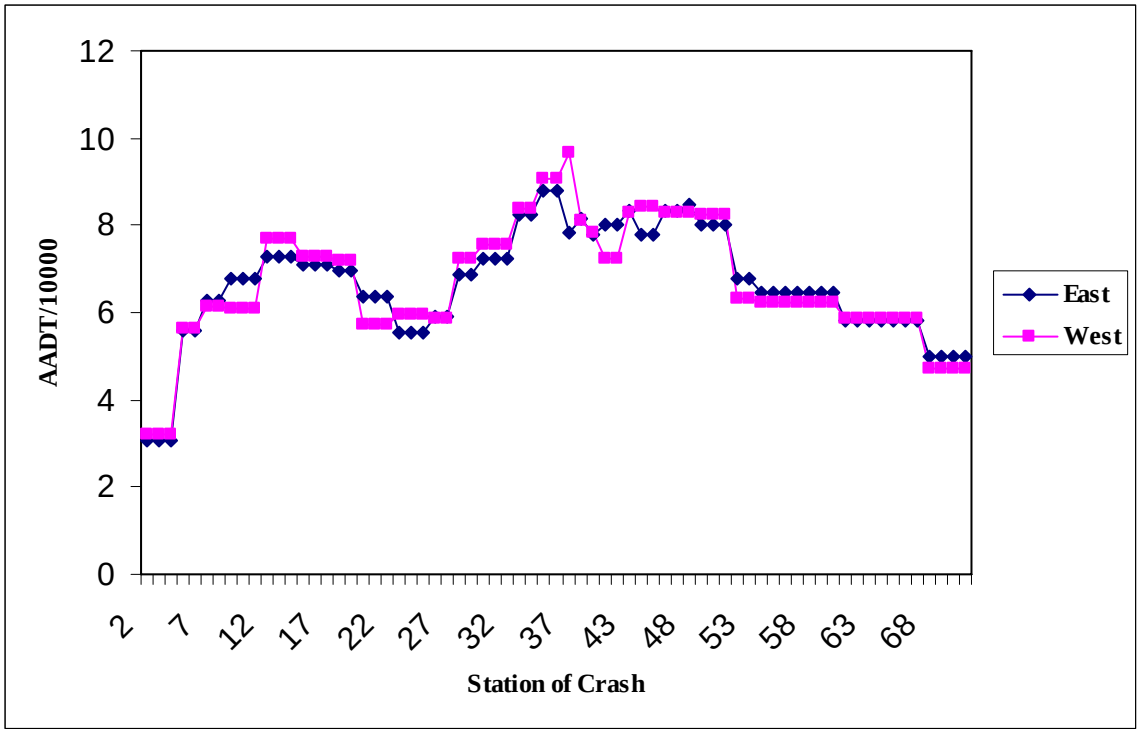

Figure 2-3: Average AADT values at different Crash Stations

As shown in the plot, there is no variation in AADT values for many stations in both directions. A typical traffic volume measure is essential to capture the traffic variations and exact contribution of other factors causing crashes. For this purpose, peak fifteen minute volumes and crash volumes were considered for the present study. The peak fifteen minute volumes at each station in the East and West directions are shown in Figure 2-4.

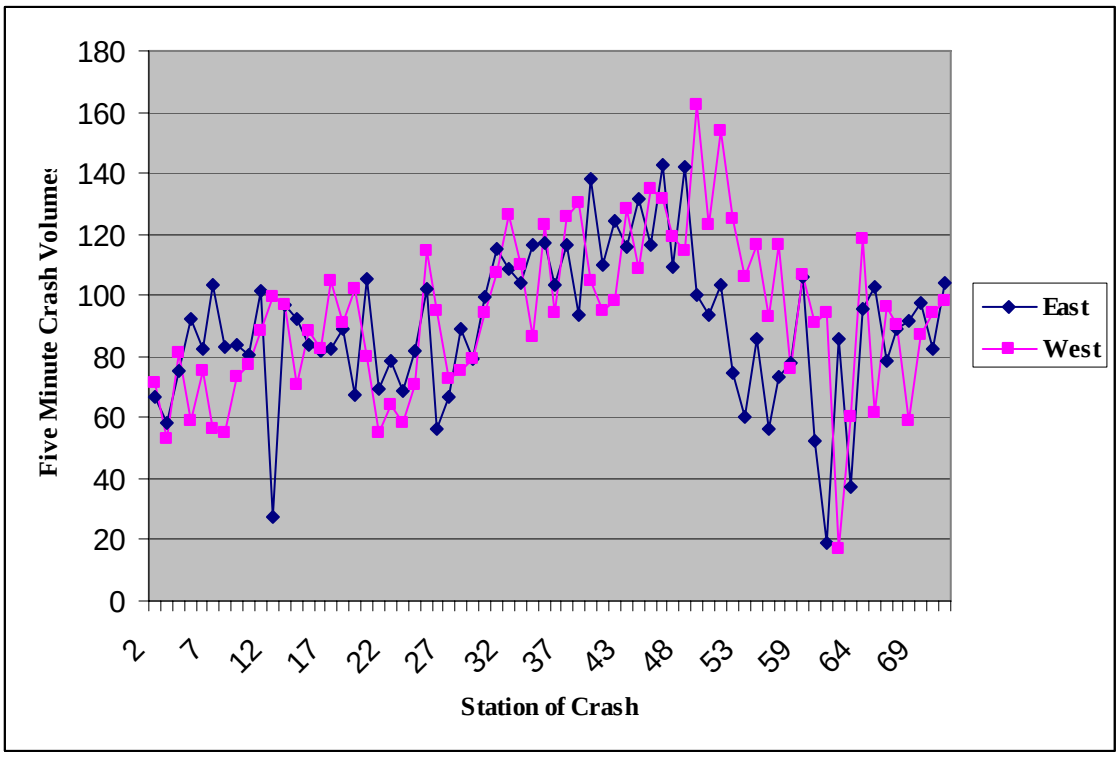

Figure 2-4: Peak Fifteen Minute Volumes at different Crash Stations

The plot has two peaks in the East direction and one peak in West direction. Different peaking profiles at each of the crash stations compared to almost invariant AADT profile might explain the occurrence of crashes better than the AADT. Unlike the peak 15 minute volumes, five minute crash volumes shown in Figure 2-5 do not exhibit two peaks. The maximum crash volume occurs at station 51 in the East direction and at station 50 in West direction. A number of factors lead to the occurrence of crashes and it is difficult to evaluate every factor. By taking the volumes preceding crash occurrences, it is possible to capture the true effects of the specific traffic volume (and also other factors) that are causing crashes.

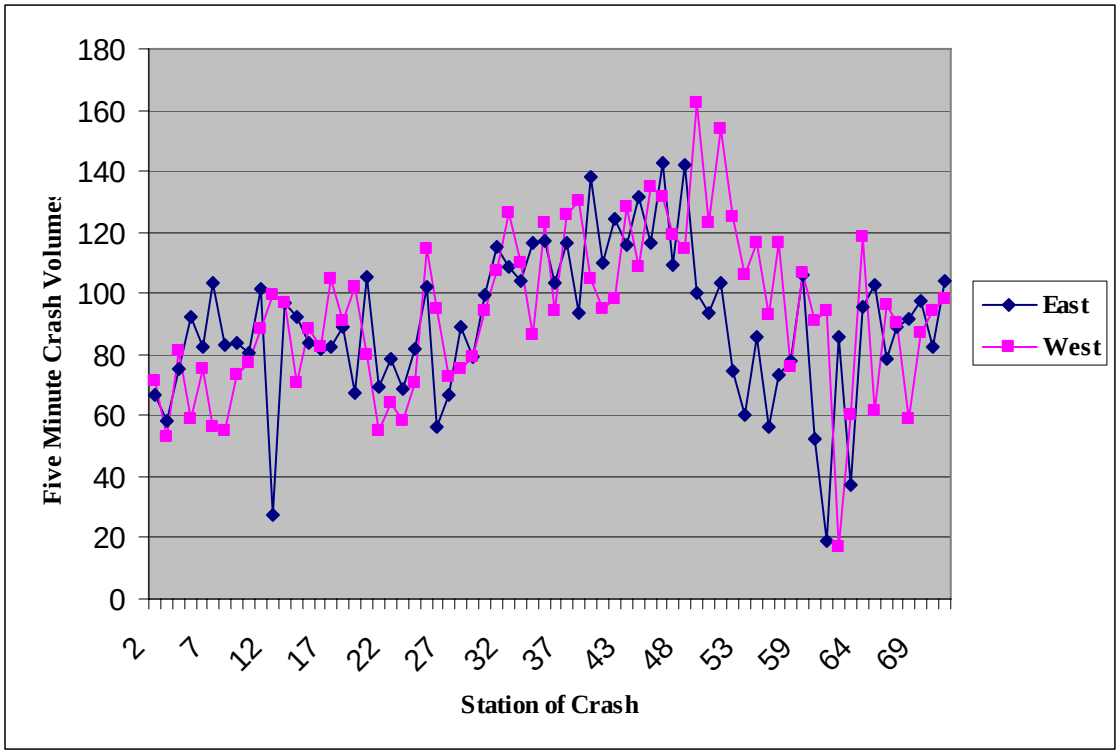

Figure 2-5: Five Minute Crash Volumes at different Crash Stations

To conclude, microscopic volume forms can accurately represent the crash location in both spatial and temporal contexts, since the traffic volume changes by station and time of day. To support these initial ideas an appropriate statistical analysis method was conducted to decide the best among these different volume forms.

2.7 Model Framework

This study estimates crash frequency for influence area segments defined for 138 loop detector stations on I-4 as a function of traffic, geometric and roadway characteristics. In general the model takes the following form:

Crash Frequency = *f (Traffic volume, geometric and roadway characteristics)*

The traffic volume measure is included in different forms 1) five, ten, fifteen or twenty minute crash volumes provided by the loop detector stations, 2) The peak fifteen minute volumes, and 3) AADT and VMT values.

2.8 Crash Frequency Models

Although a Poisson distribution can account for the discrete nature of crash occurrence, this distribution assumes that mean equals the variance for dependent variable. Previous work in the field of road safety research has shown that this assumption is mostly not true (Shankar et al., 1995; Abdel-Aty and Radwan, 2000). In such case, as an alternative, a Negative Binomial regression technique can be used. It allows for an unequal mean and variance, both when estimated mean exceeds the variance (over-dispersion), and estimated mean is less than variance (under-dispersion) of the distribution.

The Negative Binomial model has the following form (Washington et al., 2003):

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

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.

With the variables explained in the previous section, first a Poisson model was estimated. The model was checked for over-dispersion or under-dispersion, which is common in Poisson models. Based on results from the Poisson model, it was decided to fit a Negative Binomial model to accommodate for over-dispersion. For the best model selection Akaike Information Criteria (AIC) was applied. 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.

The best model is decided by the lowest value of AIC.

The Negative Binomial model can be shown as:

Expected number of crashes:

$$(\lambda_i) = \exp(\text{Intercept} + \beta * \text{traffic volume} + \gamma * \text{geometric characteristics} + \delta * \text{roadway characteristics})$$

Where β, γ, and δ are the vectors of corresponding estimable parameters.

2.9 Model Estimation and Results

This section contrasts the models developed using different traffic volumes while trying the same geometric and roadway features in each case. A total of three different main models were developed for this purpose. Table 2-1 provides a code sheet for all variables used in the frequency model.

Table 2-1: Code Sheet for all the variables used in the Model

Variable	Type	Code	Explanation of variables
Frequency of crashes a each loop detector station	Response	Freq	
Natural log of Average Five minute Volumes before the crash	Quantitative	LogFIVE	
Natural log of Average Ten minute Volume before the crash	Quantitative	LogTEN	
Natural log of Average Fifteen minute Volume before the crash	Quantitative	LogFIFT	
Natural log of Average Twenty minute Volume before the crash	Quantitative	LogTWEN	
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 Index Category	Qualitative	Pindcat	0 - High Asphalt 1 – Concrete
Pavement Surface Type Category	Qualitative	Psurcat	0 – Sheet Asphalt, Asph. Conc., BI. 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
Natural log of Average Annual Daily Traffic Volumes	Quantitative	LogAADT	
Natural log of Vehicle Miles Traveled values	Quantitative	LogVMT	
Natural log of Peak Fifteen minute volumes	Quantitative	LogPEAKFIFT	

Originally the database provided radius of curve at a crash station in feet. To identify the best possible cut-off value to separate between presence and absence of curve at a crash station that would significantly affect the crash occurrence, various values were attempted. For different models that were tried in the present study, a cut-off value of 3000 ft for presence of curve was found to be significant in the occurrence of crashes. The future reference for all the variables will be based on Table 2-1

Different logical transformations were tried for each of the volume variables and the logarithmic transformation was taken based on a better model fit. Also to be sure that no important factors were neglected, a 90% confidence level was chosen for independent variable selection. First, separate Negative Binomial models were run to identify the best aggregate traffic factor between *LogAADT* and *LogVMT* to obtain the first main model. Two-way interactions between any two possible independent variables were also tested. The *LogVMT* factor was found to be highly insignificant at 90% confidence level. The *LogAADT* factor was found to be significant and the model had the least AIC value of 1055.63 compared to the similar model using *LogVMT*. This was taken as the first main model. The parameter estimates for the first main model are provided in Table 2-2.

Table 2-2: Parameter estimates of significant factors for the First Main Model

Parameter	**Estimate**	**Pr > |t|**
Intercept	-6.6998	0.025
LogAADT	0.8403	0.002
Radcat	0.3598	0.035
Mtypcat	-0.3415	0.012
Psurcat	0.6929	<.0001
Offrcat	0.4555	<.0001
Onrcat	0.3298	0.002

The second main model was developed with the peak fifteen-minute volumes as the traffic volume measure. The *LogPEAKFIFT* was found to be insignificant at 90% confidence level. So this main model was not considered in developing the overall final model.

To obtain the third main model, the best among each of the models using crash volumes were determined. For this, separate models were fitted using *LogFIVE, LogTEN, LogFIFT and LogTWEN* covariates. Here also two-way interactions between any two possible independent variables were tested. All the crash volumes were highly significant. The model with the *LogFIVE* variable was found to be the best with an AIC value of 1026.20. The model results with the *LogFIVE* variable is provided in Table 2-3.

Table 2-3: Parameter estimates of significant factors for the Third Main Model

Parameter	**Estimate**	**Pr > |t|**
Intercept	0.0255	0.958
LogFIVE	0.9441	<.0001
Radcat	0.3767	0.0155
Lanes	0.1681	0.0606
Mtypcat	-0.288	0.0244
Psurcat	0.5774	0.0004
Offrcat	0.4679	<.0001
Onrcat	0.319	0.0015

To reach the overall best model among the three main models, a comparison between the model with the *LogFIVE* and the model with *LogAADT* was done. As the model with five-minute crash volumes had the least AIC value, the third main model was taken as the overall best model. Table 2-4 provides a comparison of standard errors between the first and the third main model. The standard errors, a measure of goodness-of-fit of the model, were comparatively lower for the third main model.

Table 2-4: Comparison of Standard Errors between First and Third Main Models

Common Variable	Standard Errors for the First Main Model	Standard Errors for the Third Main Model
Radcat	0.171	0.155
Mtypcat	0.137	0.128
Psurcat	0.175	0.162
Offrcat	0.112	0.105
Onrcat	0.110	0.100

As shown in the results, both main models have the same significant variables, except for the variable *Lanes* which was significant only in the third main model. The mathematical form of the overall final model can be written as:

Expected number of crashes for the four year period at freeway section *i*:

λ_i=exp(0.0255 + 0.9441*LogFIVE+ 0.3767*Radcat + 0.1681*Lanes - 0.288*Mtypcat + 0.5774*Psurcat + 0.4679*Offrcat + 0.319* Onrcat)

The variables used in the above equation are defined in Table 2-1.

2.10 Discussion of Results

Other than the volume factor, the significant factors in the overall model were road curvature, number of lanes, median type, pavement surface type and presence of on-ramps/off-ramps. As crash volume increases, crashes are more likely to occur. As the number of vehicles before a crash increases, the possibility of a crash among vehicles, also increase. Crashes are more likely to occur on sections with relatively sharp curves, as it would be difficult for easy maneuverability on such sections. Shankar et al. (1995) found similar results. The present study found that a radius of freeway section less than 3000 ft plays a significant role in the occurrence of crashes.

It is also likely to experience more crashes at locations having on-ramps or off-ramps in their vicinity due to the conflict among the vehicles around merge and diverge areas of the freeway. This result is consistent with the study by Lee at al. (2002).

With an increase in the number of lanes, the number of crashes was found to increase (Abdel-Aty and Radwan, 2000). This is expected as the increase in number of lanes increases the number of possible maneuvers undertaken by the vehicles, which in turn increases the chances for conflicts and potential crashes.

Freeway sections having medians with no barriers were found to have a higher number of crashes, which confirms the findings of Souleyrette et al. (2001). The presence of concrete pavement surface type is found to cause more crashes than the combination of Sheet Asphalt, Asphaltic Concrete and Bituminous surface. This can be attributed to the inherent smoothness of the asphalt pavements, a key to maintain vehicle's stability (Brock, 2002). Other possible reasons for Asphalt surfaces being involved in lesser number of crashes could be due to better visibility on asphalt surfaces, as asphalt reflects lesser light than the concrete counterpart. Also asphalt pavement might have other advantages including better drainage and less noise. However, this study points to further research that need to be conducted into the effects of pavement surface types on crash occurrence.

2.11 Conclusions

Crash frequency modeling is an important part of road safety analysis, as it helps in identifying hazardous locations on roadways. Poisson and Negative Binomial models have been widely used for modeling frequency of crash occurrence. Negative Binomial

models can represent the crash frequency data by accounting for unequal mean and variance of the dependent variable. Although many previous studies have used the Negative Binomial technique for modeling crash frequency, there has been always a debate regarding the representation of traffic volume for a crash location. The results of this study suggest a new way for accounting for the effect of traffic volume. The results show that road curvature, median type, number of lanes, pavement surface type and presence of on/off-ramps are among the significant factors that contribute to crash occurrence. Since crash rate is defined as crash frequency divided by AADT or VMT (a value that can be used as an index for comparing different locations thereby identifying the most crash prone locations). This study showed that AADT or VMT might not be good measures of traffic volume. For future research, more precise traffic related characteristics of a location may need to be determined, which can provide improved models by accounting for more precise traffic volumes. This is now achievable with the use of loop detectors on urban freeways. To conclude, this study points that the traffic volumes collected from loop detectors before crashes are a better form of traffic volume than peak volumes also obtained from loop detectors or more aggregate measures such as AADT or VMT. While the implementation of models developed in this study was not one of the objectives of this paper, the results illustrated the significant factors that influence crash occurrence, and the need for the development of better measures to account for traffic volume, of which some are suggested here. Using the model 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 resources. A possible extension 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.

3 CRASH FREQUENCY MODELING FOR DIFFERENT CRASH CATEGORIES USING SEEMINGLY UNREALTED NEGATIVE BINOMIAL REGRESSION

3.1 Introduction

Typical crash frequency models use all crashes instead of dividing them based on type of crash, peaking conditions, availability of light, severity, or pavement condition etc. Moreover researchers traditionally used AADT as one of the traffic variable. These two cases are examples of macroscopic crash frequency modeling. Apart from identifying the factors directly related to the crash occurrence, research has also been done to identify distinctive factors related to crashes. For instance, Polanis, (1995) has shown that majority of the fatal crashes happen during dark hours. Hence splitting the crashes into various possible logical categories would help in identifying more accurate causes of crash occurrence. For instance, injury and property damage only crashes might have different causal factors. But just categorizing the crashes might not be sufficient. There is also a necessity to include microscopic or disaggregated data in crash frequency analysis. The study described in this chapter aims at developing microscopic crash frequency models, both by splitting the crashes into various categories, and using various microscopic traffic factors while including static factors like roadway and geometric factors.

Some transportation data are best modeled by a system of interrelated equations (Washington et al., 2004). Some examples include the interrelation of utilization of vehicles, if there is more than one vehicle in a household, interrelation among traffic variables on a multilane roadway, etc. For instance single and multiple vehicle crashes

may have some factors, which are omitted from both models and appear in the error terms. As expected these error terms might be correlated. The standard estimation techniques work poorly for such interrelated equations. The present study which models different crash categories has the same problem and seemingly unrelated regression (SUR) technique was used to solve this problem. Essentially the present study models various crash categories using seemingly unrelated Negative binomial techniques. The discrete nature of crash is approximated by negative binomial process while the correlation between the error terms across equations is tackled by seemingly unrelated regression technique.

The previous chapter dealt with crash frequency modeling considering all crashes that happened during the study period of 4 years (overall model in Table 2-3). The model compared different forms of traffic volumes, both microscopic and macroscopic ones. This model had the benefit of using average traffic volumes immediately preceding the crash occurrence because at least one crash was reported at every crash station in the study corridor. Once the crashes are split into various categories based on different criteria, there are some crash stations that recorded zero crashes. Hence average volumes immediately preceding the crash aggregated to 5, 10, and 15 minute intervals could not be used in the split models. To make the models more efficient, traffic factors were introduced in a different way. For this purpose, the study incorporated different traffic variables such as 5, 10, and 15 minute volume and speed factors taken during normal traffic days during the year 2002 for different categories of crashes in the same way, overall model used peak fifteen minute volumes. The process used to produce these microscopic factors is explained in detail in the data preparation part of this chapter.

Essentially the models developed in this chapter used peak fifteen minute volumes taken for normal traffic days during the year 2002 and compared them with AADT/VMT usage to assess the best among the two factors in all SUR models while including geometric, roadway and other microscopic traffic factors.

3.2 Data Description

The data analyzed in the study, consists of crash data, traffic data, and geometric and roadway characteristics for a 36-mile stretch of Interstate 4 in Central Florida which includes Osceola, Orange and Seminole counties.

3.2.1 Traffic and Crash Data

The traffic data used in this study was collected from dual loop detectors (LD) installed on Interstate 4 for a 36-mile stretch. There are a total of 69 loop detectors in each direction, numbered from 2-71 installed on the freeway. Each direction is separate leading to a total of 138 loop detector stations. These dual loop detectors provide average speed, volume and average occupancy (percent of time a loop detector is occupied by vehicles) for every 30 seconds, 24 hours a day and 7 days a week throughout the year. These values are measured for each lane on I-4 in both directions, approximately spaced at half a mile apart. This data is available through the archived loop data at the University of Central Florida. The crash data was obtained from Florida Department of Transportation crash database for the same 36-mile stretch of I-4 for the years 1999 through 2002. The FDOT database provides the milepost for each crash which is generally the distance between the crash location and the starting of the county line. In

the same way, the mileposts for all the loop detector stations are also established. For the study purposes, the nearest loop detector station to the crash location is considered as the station of the crash. A total of 3146 crashes were used in the study.

In addition to the traffic data obtained from loop detectors, AADT was also used in the study. The AADT values were obtained from the AADT stations located along the selected corridor. This particular data was obtained from "Florida Traffic Information" database for the years 1999 through 2002. The main disadvantage of using AADT volumes is that there exists less number of AADT stations when compared to number of loop detector stations. Therefore, several consecutive loop stations would have the same AADT volumes.

3.2.2 Geometric and Roadway Characteristics

A total of 8 different geometric and roadway factors were considered for all the 138 loop detector stations in both directions. They include radius of the freeway section, number of lanes, median type, median width, pavement surface type, and the presence of off or on-ramps within the influence area of each crash station (the sum of half the distances between that loop and the loops on each side). Other factors such as the shoulder width, shoulder type, etc was not considered, as there was no variability in these factors along the selected section of I-4.

3.3 Data Preparation

Data preparation is an important part of analysis because of the data requirement of the study. Before the loop data was utilized in the study, it had to be cleaned for

unreasonable or unexpected values of vehicle speeds and volumes. Dual loop detectors installed under the freeway frequently report erroneous data due to sporadic hardware problems and other random errors. Hence a simple set of rules were created to clean the erroneous loop data by which the majority of these false values can be corrected. The rules which were used to eliminate these unrealistic values from 30 second raw data are as follows (Al-Deek and Chilakamarri, 2004):

- Occupancy > 100, (situation where the loop detector is 100% occupied)
- Speed = 0 or > 100 mile/hour,
- Flow > 25 /30 second,
- Flow = 0 with speed > 0, and
- Speed = 0 with Flow > 0

Although these rules cannot entirely remove the unrealistic values, yet they help in obtaining a reasonable amount of accurate data. Now this cleaned loop data was used for further analysis. Data was prepared mainly for the microscopic traffic factors used in the analysis. The microscopic traffic factors were obtained from vehicle speeds and volumes obtained from loop detectors. To represent the normal traffic characteristics of a crash station, various statistical measures like average, standard deviation and coefficient of variation of speeds and volumes for each crash station were taken. The usage of these statistical measures can be linked to recent studies made by Abdel-Aty et al. (2004), which showed the influence of these factors on crash occurrence. Peak fifteen volumes for each crash station were also included in the study to decide the best between AADT/VMT and peak volumes. The extraction of peak fifteen minute volumes is almost same as that explained in chapter 2. The only difference is that the days during which the

volumes were taken have been extended by two more days. The raw 30 second volumes were aggregated for 15 minutes and maximum value was taken for each crash station. This was considered as the peak volume or a measure of capacity at each crash station. Fifteen-minute peak volumes were considered to capture the effect of actual peaking condition on crash occurrence.

As indicated previously loop detector data was available for an entire year from 1999 through 2002. But 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. There is a chance that during these days, crashes might have happened. And indeed, crashes occurred during these days. To remove the abnormal traffic pattern caused by these crashes 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 loop detector stations was for 30 second interval. As 30 seconds 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 for 5, 10, 15 minute intervals. Data was aggregated to a maximum of 15 minute interval to keep the aggregation level as microscopic as possible. The following factors were considered for the study, which used the loop data as explained in the above paragraphs.

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 other lanes, average across all the lanes was taken to avoid the correlation. 75^{th} percentile of average speed values at every station is taken as the variable for consideration in the model. 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 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 75^{th} percentile in the analysis.

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 to cope with the correlation problem. 75^{th} percentile of standard deviation values at every station is taken as the variable for consideration in the model.

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 the one after getting the coefficient of variation first, and then obtaining its 75th percentile.

3.4 Preliminary Data Analyses

This section describes various crash categories, and interesting conclusions made based on preliminary analyses of the crash data.

The crashes that occurred on I-4 for the years 1999 through 2002 were broadly divided into the following categories:

1) By type of crash which include single and multiple vehicle crashes as the main categories,
2) By peak and off-peak period of the day,
3) By dry and wet pavement condition,
4) By availability of daylight, and
5) By severity of crash occurrence categorized into Property damage only (PDO) crashes, and crashes involving injuries/fatalities.

3.4.1 Category 1: Types of Crashes

This category has “multiple”, “single” and “other type” crashes as the main splits.

1) Multiple vehicle crashes which include more than one vehicle like rear-end, angle, side-swipe, head-on etc.
2) Single vehicle crashes which include vehicle hitting a sign-post, vehicle overturning, etc.
3) The crashes which cannot be classified exactly into multiple or single vehicle crashes were put into the “other” category. Some examples are crashes placed under this category are “crash with fire or explosion”, collision with animal, “unknown”, etc.

Table 3-1 provides a frequency table for this category.

Table 3-1: Frequency of different types of crashes

Direction	Type of Crash			Total
	Multiple	Single	Other	
Eastbound	1188	258	130	1576
	37.77%	8.2%	4.13%	50.11%
Westbound	1181	265	123	1569
	37.55%	8.43%	3.91%	49.89%
Total	2369	523	253	3145
	75.33%	16.63%	8.04%	100%

Out of 3145 crashes, 2369 were multiple vehicle crashes, 523 were single vehicle crashes and rest were other types of crashes. Majority of the crashes (75.33%) were multiple crashes and a large part of the multiple vehicle crashes were rear-end collisions (71.38%). Freeways in general experience rear-end collisions, as the traffic in the opposite direction is completely separated and there is less likelihood for head-on

collisions. Also this study deals with only the two main types: multiple and single vehicle crash types, the causes of which are relatively different and fairly known.

Figure 3-1 provides a distribution of multiple and single vehicle crashes for different loop stations ranging from 2 -71 in the Eastbound direction for all the 4 years.

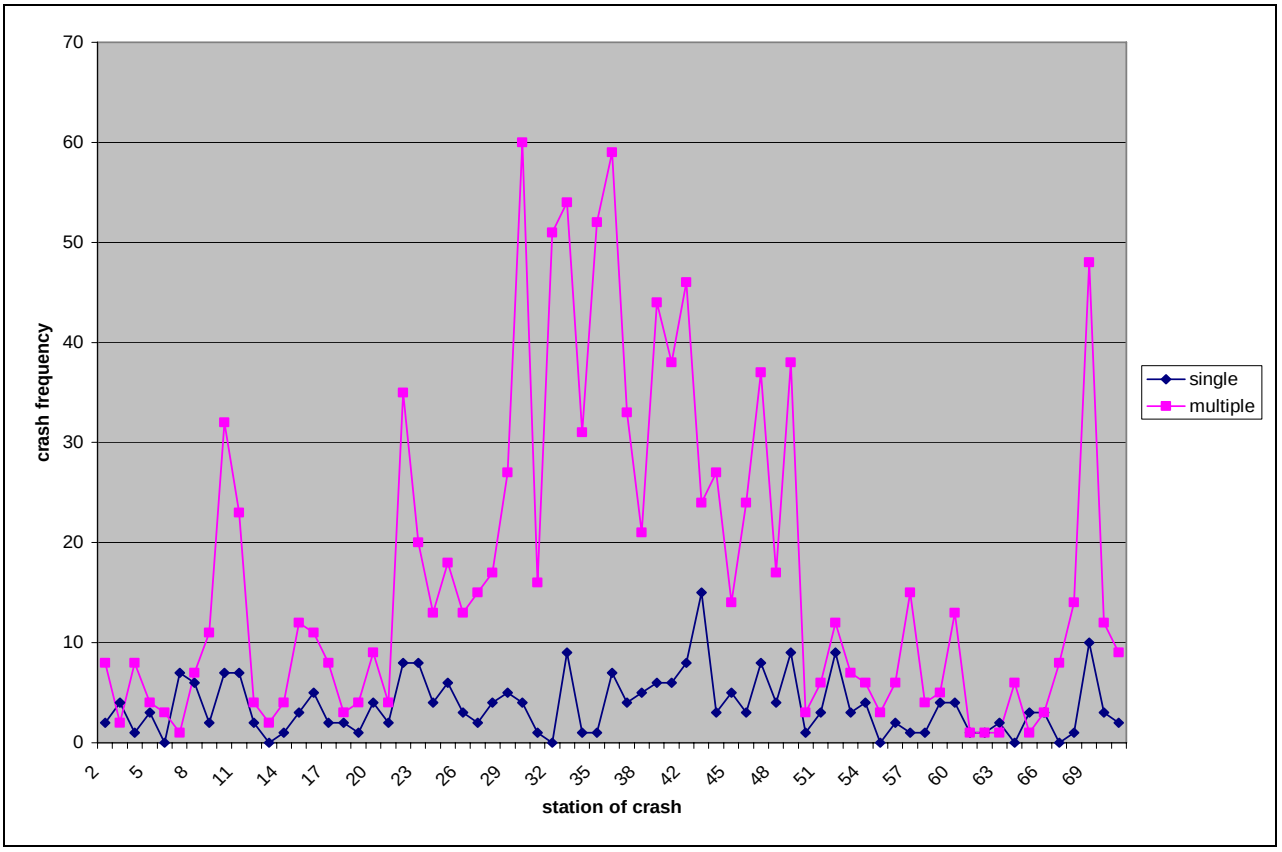

Figure 3-1: Frequency distribution of multiple and single vehicle crashes

The crash frequency for multiple vehicle crashes is highest for station 30. In general more multiple vehicle crashes occurred along stations 30 through 42. The reason could be the presence of higher number of ramps connecting the Orlando downtown which in turn generate heavy traffic around these stations. The crash frequency approximately increases from station 2 to around station 30, then again decreases from

station 42 to station 71. But stations 10 and 69 experienced comparatively large number of multiple crashes than the stations around them most probably due to reduction in number of lanes compared to the upstream station. In the case of single vehicle crashes, there is no trend as observed for the multiple vehicle crashes. Some stations experienced more or comparatively the same number of single vehicle crashes. Some examples are station 7, station 52 and station 59. Although the exact reason behind these crash occurrences cannot be concluded without a model development, it can always be attributed to some visible trends. After various characteristics of these stations were examined, it was observed that these locations allow free flow traffic. Hence vehicles tend to travel at high speeds, which can be related to more single vehicle crashes. Station 43 experienced the maximum number of single vehicle crashes with 15 of them. Station 69 has comparatively more number of crashes than the stations surrounding it, probably due the reduction in number of lanes when compared to the upstream station.

3.4.2 Category 2: Peak and Off-peak Period Crashes

This category comprises the morning and evening peak crashes in one group, day-time off-peak period crashes in second group, and crashes that happened during the remaining period in the third group. The third group can be called as night time period.

1) Morning peak between 6:30 A.M. to 9:00 A.M. and evening peak from 4:00 P.M. to 7:00 P.M.
2) Day-time off-peak from 9:00 A.M. to 4:00 P.M.
3) Night-time period between 7:00 P.M. to 6:30 A.M.

The frequency table for this category is provided in Table 3-2.

Table 3-2: Frequency table by peak and off-peak period

Direction	Crashes by peak, off-peak and night time period			Total
	Peak	Off-peak	Night-time	
Eastbound	531(96.54)	570(81.44)	475(41.30)	1576
Westbound	551(100.18)	546(78.00)	472(41.04)	1569
Total	1082(196.72)	1116(159.42)	947(82.34)	3145

The values provided in brackets in Table 3-2 denote normalized values of crash numbers by number of hours for each category. The peak period has higher percentage of crashes than the off-peak period, but the difference is not large. There are a sizeable number of crashes during the night-time period. The reason might be that the time range for both the peak and off-peak periods is approximately equal (5.5 and 7 hours respectively), whereas the night time period covers a considerably large time range of 11 hours.

A plot providing the number of crashes at different stations for peak and off-peak periods in eastbound direction is given in Figure 3-2.

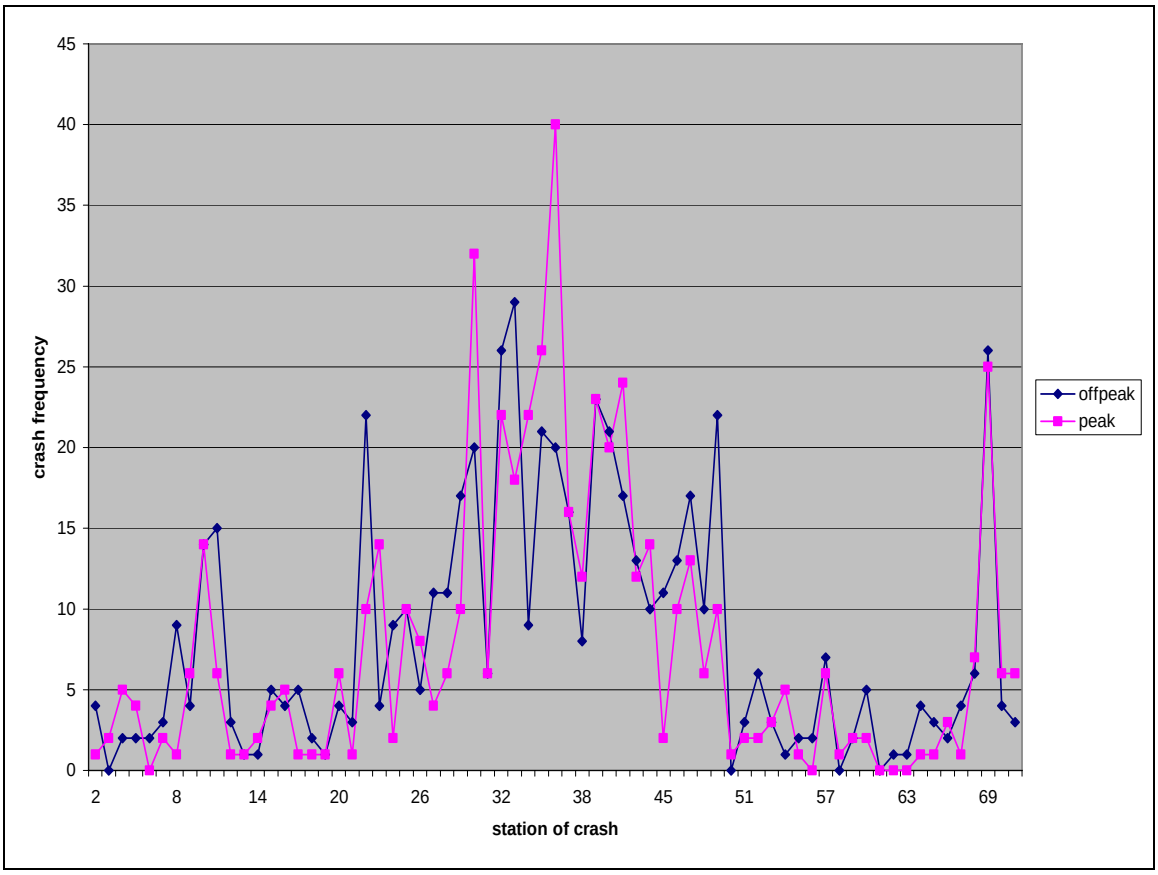

Figure 3-2: Frequency distribution of peak and off-peak period crashes

Most of the off-peak crashes occurred at station 33, and maximum number of peak period crashes occurred at station 40. Station 69 has comparatively more number of crashes than the stations surrounding it. Roughly there is an increasing trend till stations 30-32 and then the crash frequency follows a decreasing trend in case of both peak and off-peak period crashes.

3.4.3 Category 3: Crashes based on Pavement Condition

Based on the condition of the pavement during the crash occurrence, the crashes were divided into two groups, based on whether the pavement was wet or dry.

The frequency table is provided in Table 3-3.

Table 3-3: Crash frequency table for dry and wet pavement crashes

Direction	Crashes based on pavement condition			Total
	Dry	Wet	Other	
Eastbound	1300(82.50%)	262(16.62%)	14(0.88%)	1576(100%)
Westbound	1276(81.35%)	285(18.50%)	8(0.51%)	1569(100%)
Total	2576(82.00%)	547(17.40%)	22(0.70%)	3145(100%)

Most of the crashes happened on dry pavement (82.00%), and a significant amount of crashes (17.40%) happened on wet pavement. Crash occurrence during "other" pavement condition is negligible. The wet pavement crashes were an indication of rain occurrence during the crash occurrence. Although it cannot be certainly determined whether there was rain occurrence during the crash, but drivers generally experience reduced controllability of the vehicle on a wet pavement. A plot providing the number of crashes at different stations in the Eastbound direction for dry and wet pavement conditions is given in Figure 3-3.

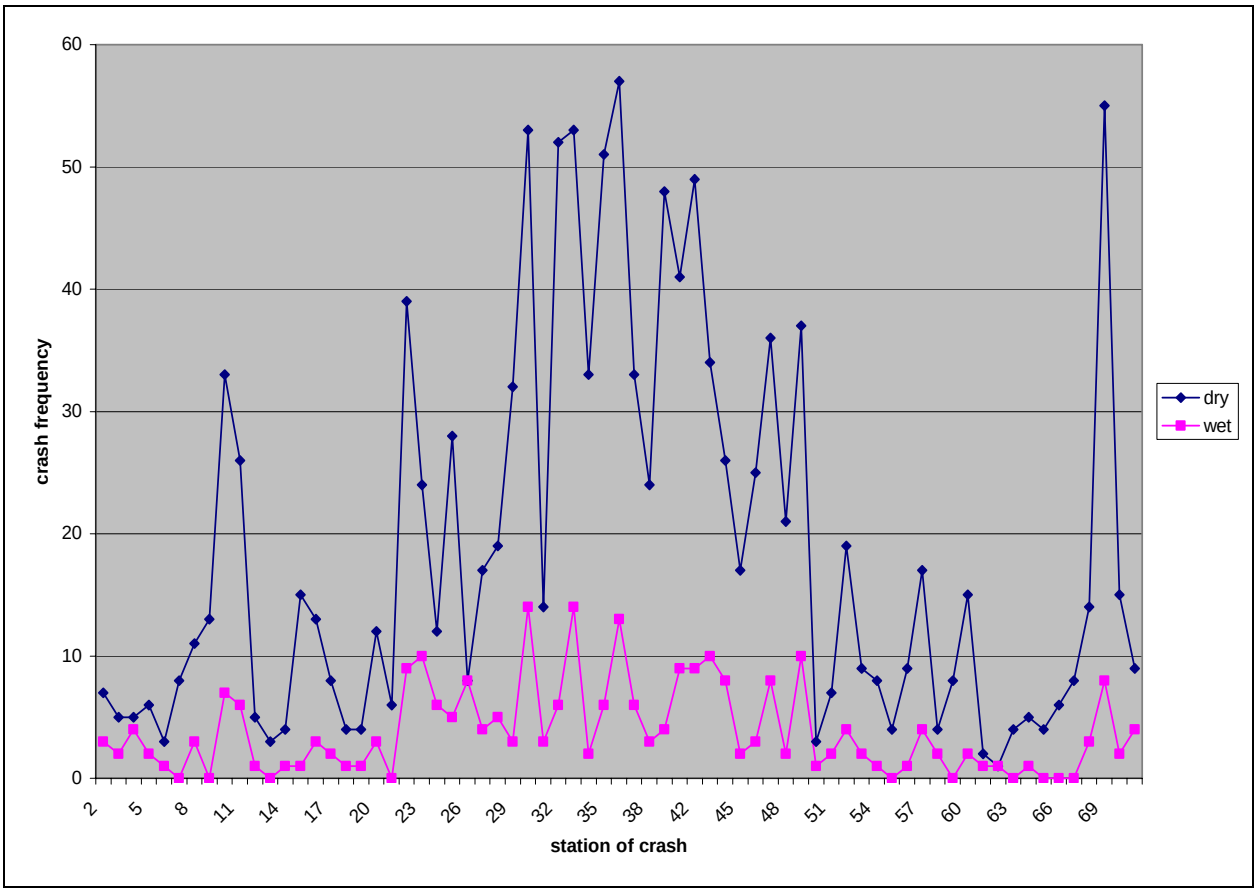

Figure 3-3: Frequency distribution of peak and off-peak period crashes

Station 36 experienced most of the dry pavement crashes. The stations from 30 through 42 had a higher number of crashes when compared with other stations. Station 69 had higher number of dry pavement crashes when compared with the stations surrounding it.

3.4.4 Category 4: Crashes based on Availability of Daylight

Based on availability of daylight, the crashes were divided into two main groups. The crashes happened during the dawn and dusk hours were not considered in the study because few crashes occurred during these hours. The hour in which a crash has occurred

was obtained from the Florida crash database. Group 1 has crashes that happened during day light and group 2 has crashes happened during dark hours, i.e. after the dusk and before dawn. The dawn and dusk hour crashes were taken as group 3.

The frequency table for this category is provided in Table 3-4.

Table 3-4: Crash frequency table for crashes based on availability of sunlight

Direction	Crashes based on availability of sunlight			Total
	Day light	Dark	Dusk & Dawn	
Eastbound	1074(68.15%)	435(27.60%)	67(4.25%)	1576(100%)
Westbound	1084(69.10%)	397(25.30%)	88(5.61%)	1569(100%)
Total	2158(68.00%)	832(27.00%)	155(5.00%)	3145(100%)

A high percentage of crashes (68%) occurred during day-light and a significant number of crashes happened during dark hours (27%). Crashes during dusk and dawn hours combined were around 5% of the total. A plot providing the number of day and dark time crashes at different stations in the Eastbound direction, is given in Figure 3-4.

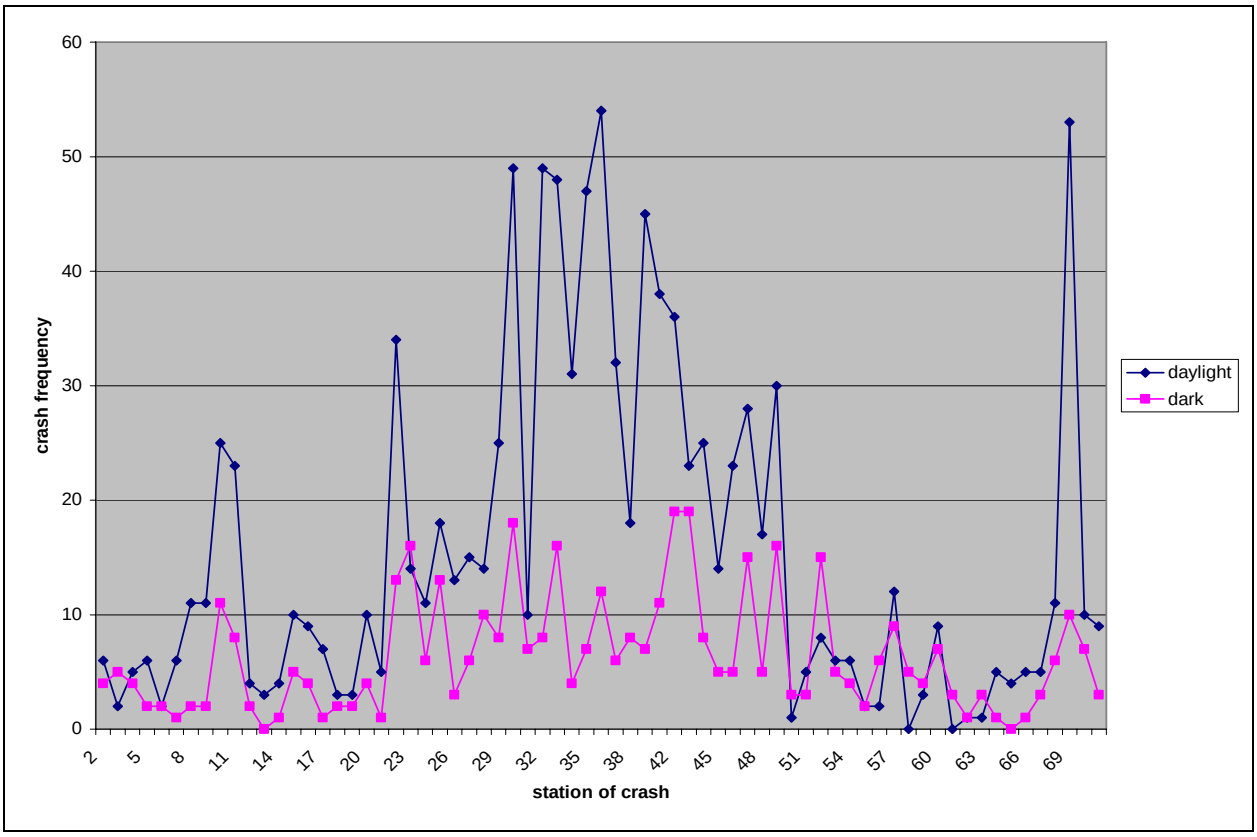

Figure 3-4: Crash frequency of daylight and dark hour crashes at different stations

The daylight and dark time crashes have approximately the same trend along all the crash stations. The highest number of daylight crashes occurred at station 36. Also station 69 has experienced comparatively a higher number of crashes (53 crashes) than the stations surrounding it. For example, station 68 had 11 crashes and station 70 recorded 10 crashes. In the case of crashes happened during dark hours, station 42 has the highest number. Although most of the stations have more daylight crashes, some stations like 50, 51, 52, 63, etc. had more dark hour crashes when compared with the daylight crashes.

3.4.5 Category 5: Crashes based on Injury Occurrence in a Crash

Based on the injury occurrence during a crash, this category has two main groups. Property damage only crashes, i.e., crashes without injuries or fatalities were placed in group1 and crashes with injuries or fatalities were placed in group 2. Since negligible number of fatalities was reported during the crash occurrences, it was decided to combine fatal crashes along with the injury crashes. The frequency table for this category is provided in Table 3-5.

Table 3-5: Crash frequency table for injury and PDO crashes

Direction	**Crash severity**		**Total**
	PDO	**Injury/Fatal**	
Eastbound	599(38.00%)	977(62.00%)	1576(100%)
Westbound	592(37.73%)	977(62.27%)	1569(100%)
Total	1191(37.87%)	1954(62.13%)	3145(100%)

There were around 62% injury and fatal crashes and 38% property damage only crashes, which imply that the majority of the crash occurrences involve injuries. A plot providing the number of PDO and injury/fatal crashes at different stations in East direction is given in Figure 3-5.

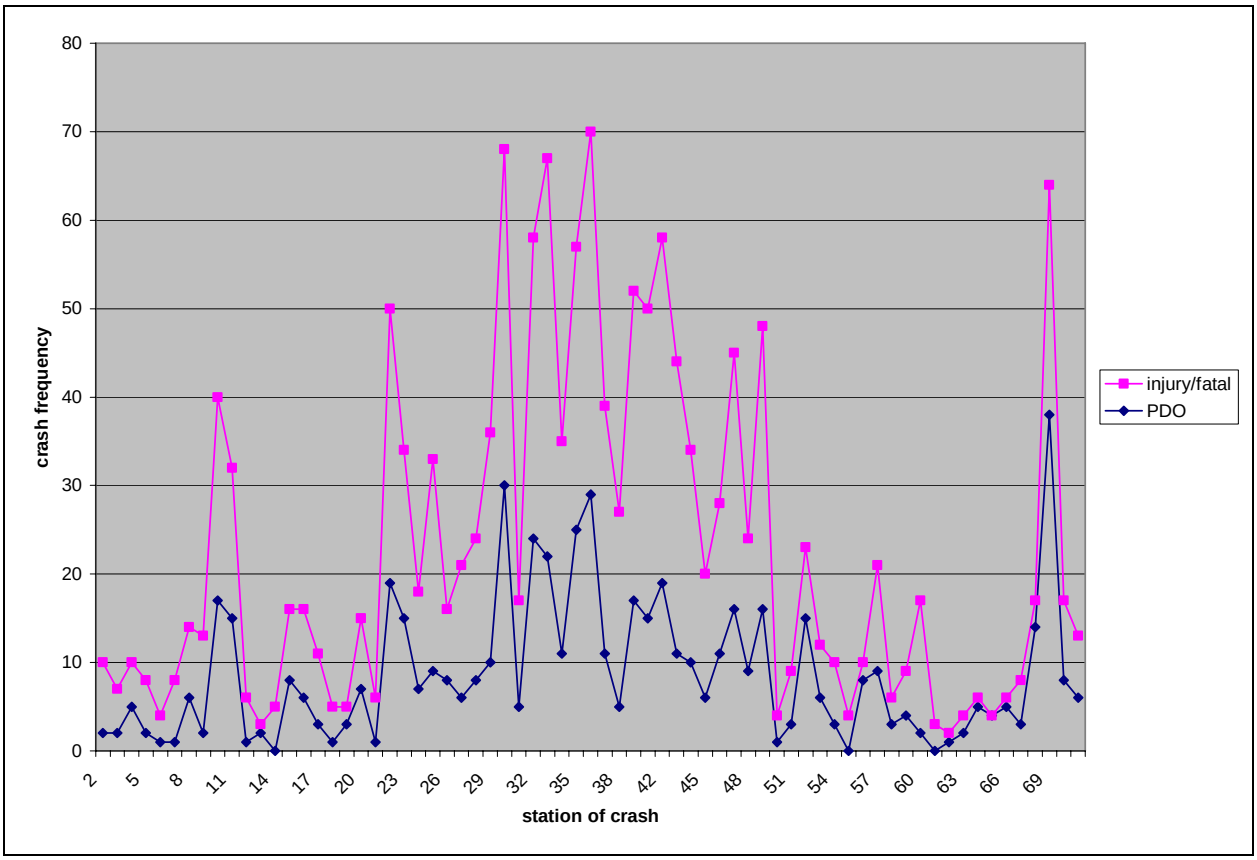

Figure 3-5: Crash frequency of PDO and injury crashes at different stations

Both the PDO and injury crashes nearly follow the same trend across all the stations. Roughly there is an increasing trend till stations 30-32 and then the crash frequency follows a decreasing trend. Station 69 experienced, relatively a higher number of both the injury and PDO crashes than the stations surrounding it.

3.5 Categorical Data Analyses

In this section, the association between type of crash (single and multiple) and several factors related to the traffic condition, and environment of the crash are explored. The main aim is to find the situations in which these two crash types experience high frequency. The crash types, multiple and single vehicle crashes were analysed using contingency tables. The relationship between type of crash and different conditions like the pavement condition, injury severity, etc was investigated using the conditional probabilities. Considerable association was determined by rejecting the null hypothesis for the Chi-square test of independence. The hypothesis of independence between the type of crash and the variable of interest is rejected based on a confidence level of 95%, corresponding to a p-value of 0.05. The chi-square value and contingency coefficient are provided under each distribution table of multiple and single vehicle crashes in different conditions. Contingency coefficient is provided to indicate the strength of the association between the variables of interest. Contingency coefficient can be defined as:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Where C is the contingency coefficient, χ^2 is the chi-square value, and N is the total number of observations. A value of C close to 1 indicates a strong association. More about this topic can be found in Abdel-Aty et. al (1999). The following sections present the statistically significant results, i.e., the hypothesis of independence between crash type and other variables.

3.5.1 Type of Crash and Traffic Condition

To observe the pattern of multiple and single vehicle crashes during peak and off-peak hours a contingency table was prepared as shown in Table 3-6. Pearson chi-square test was conducted to test the hypothesis of independence. The probability for the test statistic was found to be 0.0067, which is less than 0.05. This indicates that the rows and columns of this contingency table are dependent. So it can be concluded that more percentage of multiple vehicle crashes occur during the peak period. This can be justified given that freeways in general experience more rear-end collisions during peak period.

Table 3-6: Distribution of multiple and single vehicle crashes by peak and off-peak period

Peaking conditions	**Type of crash**		
	Multiple	**Single**	**Total**
Peak	895(89.05%)	110(10.95%)	1005(100%)
Off-peak	868(84.93%)	154(15.07%)	1022(100%)
Total	1763(86.98%)	264(13.02%)	2027(100%)

Chi-square = 7.6048, DF = 1, P-value = 0.05, C = 0.06

A plot providing the number of peak and off-peak period multiple vehicle crashes at different stations is given in Figure 3-6.

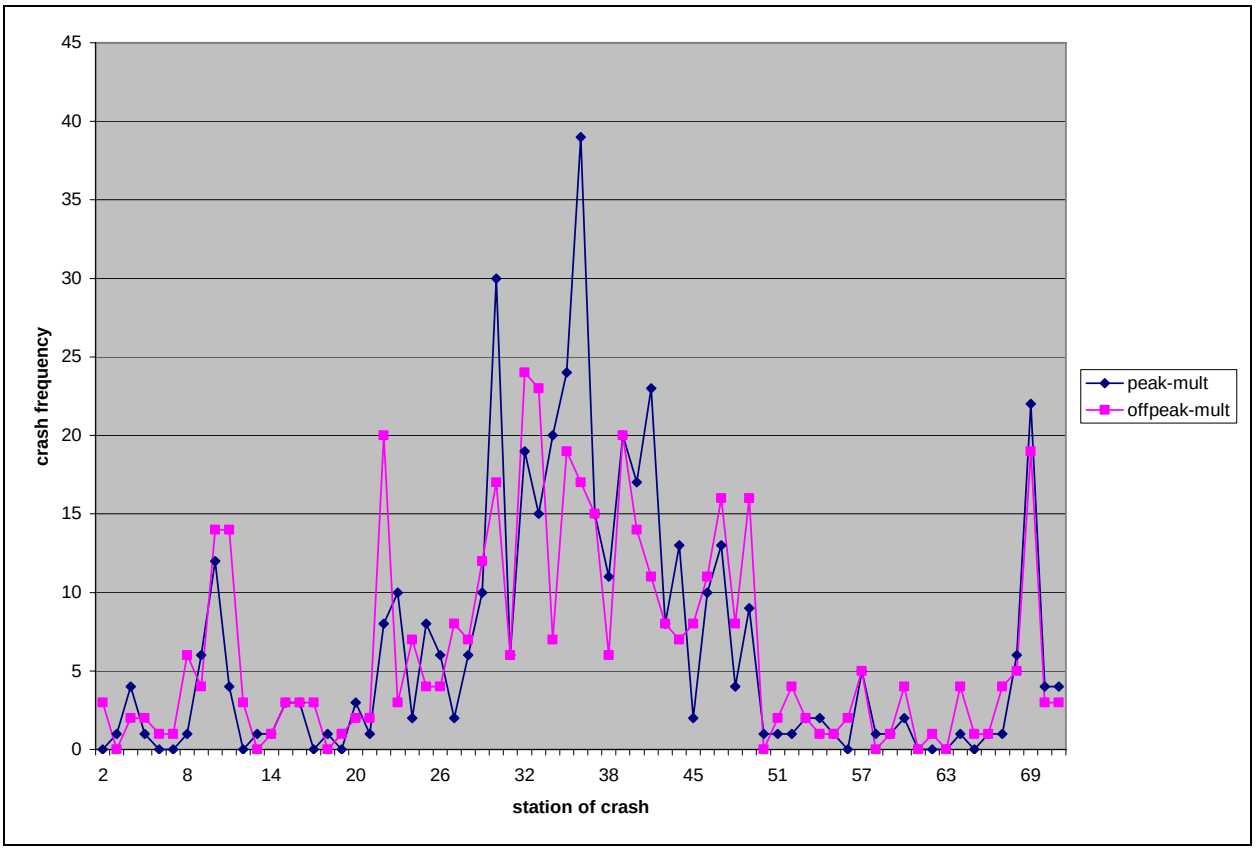

Figure 3-6: Crash frequency of multiple vehicle crashes by peak and off-peak period

In general peak and off-peak multiple vehicle crashes roughly have the same frequency at most of the stations. But contrary to this, stations around 30-42 have more peak multiple vehicle crashes when compared to off-peak multiple vehicle crashes. Analysis of various factors would help in understanding the reason behind this pattern. The stations 30-42, which recorded more peak hour multiple vehicle crashes, experienced more off-peak hour single vehicle crashes. A plot providing the number of peak and off-peak hour single vehicle crashes at different stations is given in Figure 3-7.

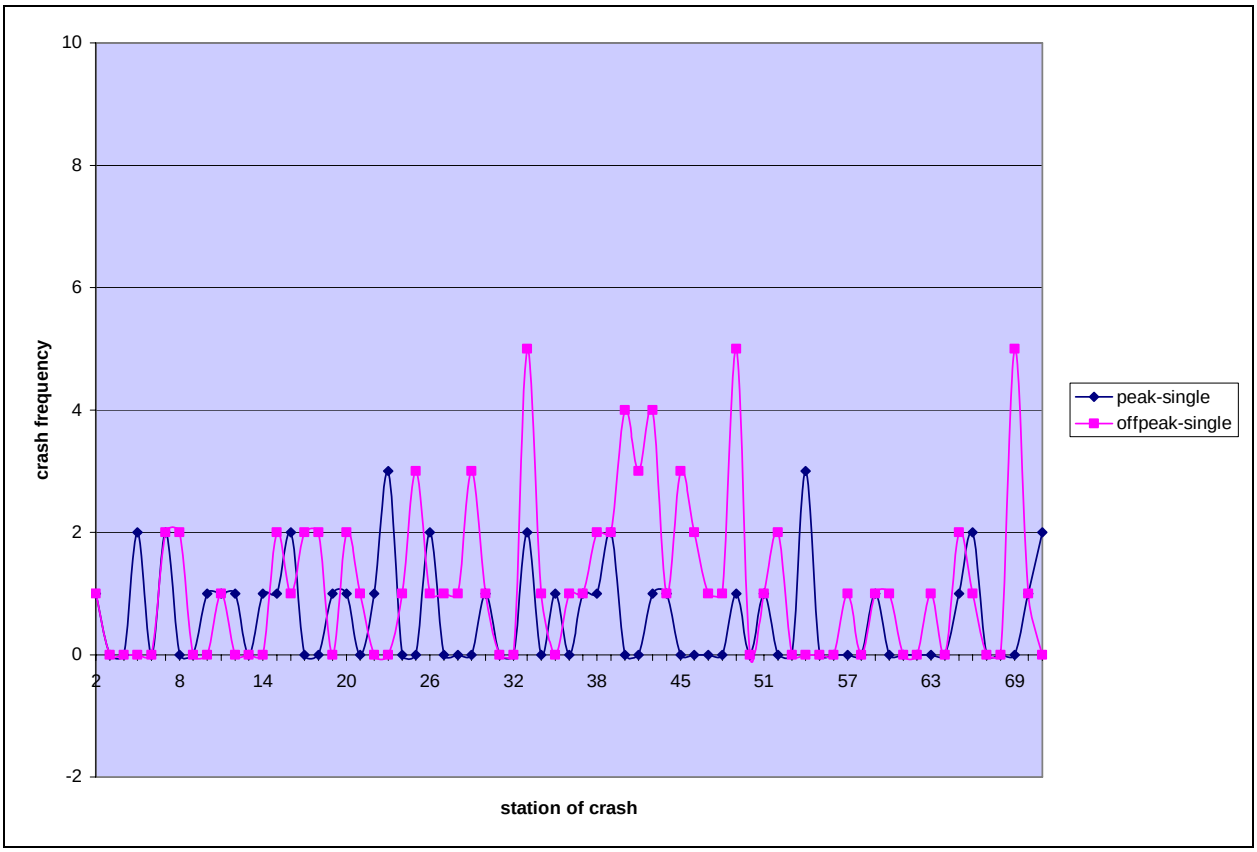

Figure 3-7: Crash frequency of single vehicle crashes by peak and off-peak period

The trend of crash frequency for peak and off-peak hour, in general, is nearly the same for most of the crash stations

3.5.2 Type of Crash and Availability of Daylight

Now to notice the distribution of multiple and single vehicle crashes, during daylight and dark hours, the following contingency table was prepared as shown in Table 3-7. Pearson chi-square test was conducted to test the hypothesis of independence. The probability for the test statistic was found to be less than 0.0001, which is less than 0.05. This indicates that the rows and columns of this contingency table are dependent.

Table 3-7: Frequency table for multiple and single vehicle crashes by lighting condition

Lighting Condition	Type of crash		
	Multiple	**Single**	**Total**
Daylight	1724(86.72%)	264(13.28%)	1988(100%)
Dark	533(69.67%)	232(30.33%)	765(100%)
Total	2257(81.98%)	496(18.02%)	2753(100%)

Chi-square = 108.68, DF = 1, P-value = 0.05, C = 0.20

So it can be said that more percentage of multiple vehicle crashes occur during the daylight and there is more probability for single vehicle crashes during dark hours. This can be justified given that freeways in general experience more multiple vehicle crashes during peak period which falls in daytime. Also there is more possibility for single vehicle collisions during dark hours, owing to the fact that there exists less traffic flow and drivers tend to drive at high speeds during nighttime.

3.5.3 Crash Type and Injury Involvement

A frequency table is prepared to see how many multiple and single vehicle crashes caused injuries/fatalities. It is provided in Table 3-8.

Table 3-8: Frequency table for multiple and single vehicle crashes by injury occurrence

Type of crash	Severity of crash		Total
	No injuries	**Injuries**	
Multiple	876(36.96%)	1494(63.04%)	2370(100%)
Single	192(36.71%)	331(63.29%)	523(100%)
Total	1068(63.08%)	1825(36.92%)	2893(100%)

Chi-square = 0.0116, DF = 1, P-value = 0.05, C = 0.01

Pearson chi-square test was conducted to test the hypothesis of independence. The probability for the test statistic was found to be 0.9204, which is higher than 0.05. This indicates that the rows and columns of this contingency table are independent. Therefore no particular conclusions can be drawn from the cells in the contingency table as crash type is similarly distributed across the different levels of injury involvement.

3.5.4 Crash Type and Pavement Condition

To observe the distribution of single and multiple vehicle crashes on dry and wet pavements, a contingency table is prepared as shown in Table 3-9.

Table 3-9: Frequency table for multiple and single vehicle crashes by pavement condition

Type of Crash	Pavement Condition		
	Dry	**Wet**	**Total**
Multiple	1970(83.51%)	389(16.49%)	2359(100%)
Single	411(79.04%)	109(20.96%)	520(100%)
Total	2381(82.70%)	498(17.30%)	2879(100%)

Chi-square = 5.99, DF = 1, P-value = 0.05, C = 0.041

Pearson chi-square test was conducted to test the hypothesis of independence. The probability for the test statistic was found to be 0.0152, which is less than 0.05. This indicates that the rows and columns of this contingency table are dependent. Therefore meaningful conclusions can be drawn from the contingency table. More percentage of multiple vehicle crashes happened on dry pavements (83.51%). Also more percentage single vehicle crashes happened on wet pavement (21%). Freeways in general experience multiple vehicle crashes and most of the times pavement condition is dry. Hence it is

obvious that most of the multiple vehicle crashes happen on dry pavements. It is a known fact that drivers tend to lose vehicle control on wet pavements due to reduced friction between the tires and pavement. As wet pavement condition arise mostly due to rain occurrence, drivers also experience reduced visibility which in turn can cause a vehicle to collide with sign pots or any other immovable objects on the road. This might be one of the causes of the more number of single vehicle crashes on wet pavements. Not every possible crash category or sub-category was explained. Nevertheless, effort was made to identify any visible patterns in different types or ways of crash occurrence through preliminary data analyses.

4 MODELING APPROACH FOR SEEMINLGLY UNRELATED NEGATIVE BINOMIAL MODELS

4.1 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 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 (Washington et al., 2003):

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

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 things (Greene, 1997). 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 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 equations 4-1 and 4-2.

$$\lambda_i = Exp(\beta_i + \varepsilon_i + \phi_i) \quad (4\text{-}1)$$

$$\lambda_i = Exp(\beta_i + \chi_i) \quad (4\text{-}2)$$

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) magnitude and sign of coefficient of estimated variable is in agreement with 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 Seemingly Unrelated Regression

Simultaneous models have seen very few applications in transportation engineering. But some transportation data is best explained by simultaneous models. Some examples are: interrelation among traffic variables (speed or volume) in one lane and other lanes, and interrelation of vehicle utilization in a household with more than vehicle. These examples create a set of equations where the target variable in one equation becomes independent variable in another. Interrelated sets of equations have to be estimated in a different way as the standard ordinary least squares(OLS) estimation does not take into consideration the correlation between regressors and error terms. If a set of equations are not simultaneous, since no dependent variable is used as independent variable in another equation, seemingly unrelated regression (SUR) can be used to estimate systems of equations with correlated disturbances. Some set of equations, for instance single and multiple vehicle crashes in the present study may appear unrelated. Nevertheless these equations may be related by the fact that some coefficients are the same, the disturbances are correlated across equations, and a subset of right hand side

variables are the same. If these systems of equations are solved by OLS estimation, coefficients of the estimated variables might be consistent but not efficient. For an efficient parameter estimates, contemporaneous correlation of error terms or disturbances have to be taken into account. Estimation of SUR models is accomplished by generalized least squares.

4.3 Estimation via Generalized Least Squares Estimation

Ordinary Least Squares estimation assumes that disturbance terms have equal variances and are not correlated. Generalized least squares (GLS) estimation is often utilized to relax these assumptions (Washington et al, 2003).

Under ordinary least squares assumptions, we have in matrix notation

$$E(\varepsilon\varepsilon^T) = \sigma^2 \mathrm{I}$$

Where $E(.)$ denotes expected value, ε is an N x 1 column vector of equation disturbance terms (where N is the total number of observations in the data), ε^T is the 1 x N transpose of ε, σ^2 is the disturbance term variance, and I is the N x N identity matrix,

$$\mathrm{I} = \begin{bmatrix} 10.....0 \\ 01.....0 \\ ..... \\ 00.....1 \end{bmatrix}$$

If heteroscedasticity is present, $E(\varepsilon\varepsilon^T) = \Upsilon$, where Υ is n x n matrix,

$$\Upsilon = \begin{bmatrix} \sigma_1^2 0 \ldots\ldots\ldots\ldots\ldots 0 \\ 0\sigma_2^2 \ldots\ldots\ldots\ldots\ldots 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 00\ldots\ldots\ldots\ldots\ldots\sigma_n^2 \end{bmatrix}$$

For disturbance-term correlation, (for auto-regressive models) $E(\varepsilon\varepsilon^T) = \Upsilon$, where Υ is n x n matrix

$$\Upsilon = \begin{bmatrix} 1\rho\ldots\ldots\ldots\ldots\ldots\rho^{N-1} \\ \rho 1\ldots\ldots\ldots\ldots\ldots\rho^{N-2} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \rho^{N-1}\rho^{N-2}\ldots\ldots\rho \end{bmatrix}$$

In ordinary least squares, parameters are estimated from

$$\hat{\beta} = (X^T X)^{-1} X^T Y ,$$

Where $\hat{\beta}$ is a p x 1 column vector (where p is number of parameters), X is an n x p matrix of data, X^T is the transpose of X, and Y is an n x 1 column vector. In General least squares, $\hat{\beta}$ can be written as,

$$\hat{\beta} = (X^T \Upsilon^{-1} X)^{-1} X^T \Upsilon^{-1} Y$$

The important and difficult part of GLS estimation is evaluating the Υ matrix. In seemingly unrelated regression Υ is estimated from initial OLS estimates of individual equations.

4.4 Development of SUR Models using aML Software

As explained in earlier paragraphs, negative binomial regression is the most suitable statistical approach to model crash frequency models. So aML Software was used which had the capability to solve the seemingly unrelated negative binomial models. aML uses the iterative process, Gauss-Newton (Judge et al., 1988) 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 (aML reference manual, 2003). The approach used to solve the SUR models in aML needs some explanation. 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. 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.

4.5 Model Estimation and Results

As mentioned previously there are five main crash categories, based on 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.

A sheet providing explanation of various variables included in model development is provided in Table 4-1. 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 for

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, day 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.

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	

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 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, 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 macroscopic variable while PEAKFIFT indicates 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(various statistical measures) like average speed, standard deviation of speed/volume, and coefficient of variation of speed/volume were found to notably influence the crash occurrence at 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 model estimation. Hence these factors were used separately in the models and the best among various models was selected based on lowest AIC value.

4.5.1 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: Multiple and single vehicle crashes. The right-hand side (independent) variables consisted of traffic, roadway and geometric factors.

4.5.1.1 Individual Multiple Vehicle Crash Model

Before arriving at the final model, two models were tried, 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 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 statistical 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.

Table 4-2: 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

The individual multiple vehicle crash model consists of presence of curve, median type, pavement surface type, presence of off-ramp/on-ramp, and AADT as the independent variables. All variables have reasonable sign and found significant at 90% confidence level. These variables are almost the same as found in Table 2-3 except for AADT. The dispersion factor ALPHA 1 is considerably different from zero, which confirms the appropriateness of negative binomial model. No other microscopic traffic variable was found to be significant in the model.

4.5.1.2 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.

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

The individual single vehicle crash model consists of presence of curve, median type, presence of off-ramp/on-ramp as the independent variables. All variables have plausible sign and found significant at 90% confidence level. The significant variables in this model can be found in Table 2-3. The dispersion factor ALPHA 2 is considerably different from zero, which confirms the appropriateness of negative binomial model.

4.5.1.3 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, 4-5, and 4-6.

Table 4-4: 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 4-5: 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 4-6: 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

As shown in Table 4-6, the correlation between the disturbance terms is substantially high with a value of 0.75. This implies that the omitted variables are allocated across the model disturbances for multiple and single vehicle crashes. Therefore the use of SUNB 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. Through the estimation of SUNB models, the errors were decreased

for some of the variables. Although the standard errors were not improved for every variable in the models, SUNB estimation can be always be justified as the correlation coefficient is highly significant (Washington et al., 2004). A comparison of standard errors between individual models and SUNB models would help in evaluating the situation. This comparison table is provided in Table 4-7 and 4-8.

Table 4-7: Comparison of standard errors between individual and SUNB multiple vehicle crash models

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.194975	0.192185
MTYCAT	0.172878	0.170384
PSURCAT	0.227456	0.247369
OFFRCAT	0.122057	0.128966
ONRCAT	0.125682	0.122117
AADT	0.086238	0.093646

Table 4-8: Comparison of standard errors between individual and SUNB for single vehicle crash models

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.201705	0.199444
MTYPCAT	0.1753	0.176071
OFFRCAT	0.131351	0.132039
ONRCAT	0.126161	0.125955

As shown in Tables 4-7 and 4-8, for half of the variables (highlighted rows) the error terms are less for SUNB models.

4.5.1.4 Discussion of Results

The variables which significantly affected the occurrence of crashes in the overall model (when all crashes were combined) were found to affect both multiple and single vehicle crashes. AADT was included in the model as crash volumes just before the crash were not available. No microscopic traffic variables were significant at 90% confidence level in both of the models. The significant factors in multiple vehicle crash model were road curvature, median type, pavement surface type and presence of on-ramps/off-ramps and AADT. In the case of single vehicle crash model, the significant factors were road curvature, median type, and presence of on-ramps/off-ramps. Thus, 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.

Both multiple and single vehicle crashes are more likely to occur on sections with relatively sharp curves as vehicles tend to lose control on such sections, although there is more possibility for single vehicle crashes. The present study found that a radius of freeway section less than 3000 ft plays a significant role in the occurrence of crashes.

It was found that more multiple vehicle crashes occur at locations having on-ramps or off-ramps in their vicinity due to the conflict among the vehicles around merge and diverge areas of the freeway. In general there would be more of multiple vehicle crashes around such locations, but there is likelihood for single vehicle crashes also. Freeway sections having medians with no barriers were found to have a higher number of both single and multiple vehicle crashes, which confirms the findings of Souleyrette et al.,

2001. The presence of concrete pavement surface type is found to cause more multiple vehicle crashes than the combination of Sheet Asphalt, Asphaltic Concrete and Bituminous surface. This can be attributed to the inherent smoothness of the asphalt pavements, a key to maintain vehicle's stability (Brock, 2002). Other possible reasons for Asphalt surfaces being involved in lesser number of crashes could be due to better visibility on asphalt surfaces, as asphalt reflects lesser light than the concrete counterpart. Also asphalt pavement might have other advantages including better drainage and less noise. However, this study points to further research that need to be conducted into the effects of pavement surface types on crash occurrence. As AADT increases, multiple vehicle crashes are more likely to occur. As the number of vehicles on a given stretch of highway increase, chances of collision among vehicles increase leading to multiple vehicle crashes. No effect of volume in the single vehicle model since these crashes tend to be caused by speeding and usually at night and involving alcohol.

4.5.2 Category 2

This category comprises two sub categories, of which one has peak period crashes and the other has off-peak period crashes. As mentioned earlier, the peak period consists both A.M. and P.M. peak period. The off-peak period consists of time period between morning peak and evening peak. A SUNB model was estimated for this category. Two left-hand side (dependent) variables were considered: 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. To deal with this problem and to evaluate whether the parameter estimates were improving reflected in terms of reduced standard errors, the correlation coefficient was set to one and then the models were estimated. Although there is no considerable literature support for this kind of estimation, nevertheless the parameter estimates were improved and hence the SUNB model for these two crash models was included in the thesis. The following sub-sections include first the development of individual models followed by the SUNB model.

4.5.2.1 Individual Peak Period Crash Model

Before arriving at the final model, two models were tried, 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 peak period 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 statistical measures. It has a log-likelihood value of -913.71. The estimation results are provided in Table 4-9.

Table 4-9: Estimation results for individual peak period crash model

Parameter	Estimate	Standard Error
CONSTANT	-0.22122	1.390376
RADCAT	0.446799	0.293079
PSURCAT	0.740952	0.324715
OFFRCAT	0.55751	0.164875
ONRACT	0.584691	0.164205
AADT	0.156247	0.113952
CVS_15	0.522615	0.292743
ALPHA_1	0.157366	0.049701

Log Likelihood: -913.71 No. of observations: 552

The individual peak period crash model consists of presence of curve, pavement surface type, presence of off-ramp/on-ramp, AADT, and coefficient of variation in speed during peak period aggregated at 15 minute intervals as the independent variables. All variables have reasonable sign and found significant at 90% confidence level. These variables are almost the same as found in overall model (Table 2-3) except for AADT and coefficient of variation in speed at 15 minute aggregation level. The dispersion factor ALPHA_1 is considerably different from zero, which confirms the appropriateness of negative binomial model. No other microscopic traffic variable was found to be significant in the model.

4.5.2.2 Individual Off-peak Period Crash Model

As explained in peak period crash model estimation, before arriving at the final model, two models were tried, one with AADT and another with PEAKFIFT (the maximum or peak fifteen minute volumes taken during the off-peak period) keeping all other variables same. AADT was found to be significant in the model. Also there were no significant microscopic traffic variables. It has a log-likelihood value of -948.2768. The estimation results are provided in Table 4-10.

Table 4-10: Estimation results for individual off-peak period crash model

Parameter	Estimate	Standard Error
CONSTANT	-1.55322	1.084202
RADCAT	0.284532	0.207281
MTYPCAT	-0.32504	0.1867
PSURCAT	0.855105	0.259004
OFFRCAT	0.278306	0.136422
ONRCAT	0.222948	0.130703
AADT	0.224538	0.099779
ALPHA_2	0.16986	0.045012

Log Likelihood: -948.27 No. of observations: 552

The individual off-peak period crash model consists of the presence of curve, median type, pavement surface type, presence of off-ramp/on-ramp, and AADT as the independent variables. All variables have plausible sign and found significant at 90% confidence level. The significant variables in this model can be found in overall model (Table 2-3) except for AADT. The dispersion factor ALPHA_2 is considerably different from zero, which confirms the appropriateness of negative binomial model.

4.5.2.3 Seemingly Unrelated Negative Binomial Model for Peak and Off-peak Period Crashes

SUNB estimation was performed for peak and off-peak period crash models. Dispersion parameters (ALPHA 1 and ALPHA 2) were evaluated. As the correlation coefficient (RHO_U1U2) is set at 1 and then the SUNB is estimated, there is no bivariate distribution for the standard deviations (SIGMA_U1 and SIGMA_U2). In this case there exists only a univariate distribution, although with different scales in peak and off-period crash models. This can be called as a single factor model. The estimation results are provided in Table 4-11, and 4-12.

Table 4-11: SUNB model estimation results for peak period crash model

Parameter	Estimate	Standard Error
CONSTANT	-0.98038	1.281598
RADCAT	0.553719	0.180192
PSURCAT	0.921018	0.271271
OFFRCAT	0.530614	0.237522
ONRCAT	0.569478	0.16036
AADT	0.191448	0.154015
CVS_15	0.32665	0.217164
ALPHA 1	0.126569	0.046609

Log Likelihood: -1823.65 No. of observations: 552

Table 4-12: SUNB model estimation results for off-peak period crash model

Parameter	Estimate	Standard Error
CONSTANT	-1.61685	1.098659
RADCAT	0.311454	0.160585
MTYPCAT	-0.33349	0.177407
PSURCAT	0.941511	0.23906
OFFRCAT	0.282931	0.197701
ONRCAT	0.232357	0.13611
AADT	0.231086	0.133695
ALPHA 2	0.151542	0.040848

Log Likelihood: -1823.65 No. of observations: 552

Since the correlation between the disturbance terms is substantially high and fixed at a value of 1, it implies that the model disturbances arising from the omitted variables have the same distribution for peak and off-peak crashes. Through the estimation of SUNB models for peak and off-peak crashes, the errors were decreased for some of the variables The reliability of the models increase through smaller standard errors. A comparison of standard errors between individual models and SUNB models would help in understanding the efficiency gained. But it has to be remembered that more research has to done into this type of estimation where the correlation coefficient is extremely high, to actually prove the validity of the simultaneous model. The case in which the efficiency is gained is highlighted in the tables. This comparison table is provided in Table 4-13 and 4-14.

Table 4-13: Comparison of standard errors between individual and SUNB models

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.293079	0.180192
PSURCAT	0.324715	0.271271
OFFRCAT	0.164875	0.237522
ONRACT	0.164205	0.16036
AADT	0.113952	0.154015
CVS_15	0.292743	0.217164

Table 4-14: Comparison of standard errors between individual and SUNB models

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.207281	0.160585
MTYPCAT	0.1867	0.177407
PSURCAT	0.259004	0.23906
OFFRCAT	0.136422	0.197701
ONRACT	0.130703	0.13611
AADT	0.099779	0.133695

As shown in Tables 4-13 and 4-14, for most of the variables the error terms are less for SUNB models in case of peak period crash models and in case of off-peak crash model parameter estimates improved for three out of six.

4.5.2.4 Discussion of Results

Most of the variables which were found significant in overall model (Table 2-3) were also found to affect both peak and off-peak period crashes. 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. So the common factors influencing both these crashes were road curvature, pavement surface type, presence of on-ramps/off-ramps, and AADT. Both peak and off-peak period crashes are more likely to occur on sections with relatively sharp curves as vehicles find it difficult to maintain control on such corridors. The present study found that a radius of freeway section less than 3000 ft plays a significant role in the occurrence of crashes. Freeway sections having medians with no

barriers were found to have a higher number of off-peak period crashes, which confirms the findings of Souleyrette et al., 2001. The presence of concrete pavement surface type is found to cause more peak and off-peak period crashes than the combination of asphalt surface. This can be attributed to the inherent smoothness of the asphalt pavements, a key to maintain vehicle's stability (Brock, 2002). Other possible reasons for Asphalt surfaces being involved in lesser number of crashes could be due to better visibility on asphalt surfaces, as asphalt reflects lesser light than the concrete counterpart. Also asphalt pavement might have other advantages including better drainage and less noise. However, this study points to more research that needs to be conducted into the influence of pavement surface types on crash occurrence. It was found that more peak and off-peak period crashes occur at locations having on-ramps or off-ramps in their vicinity since there will be conflict among the vehicles around merge and diverge areas of the freeway during both day and dark hour time periods. As AADT increases, both peak and off-peak period crashes are more likely to occur. As the number of vehicles on a given stretch of highway increase, chances of collision among vehicles increase leading to both peak and off-peak period crashes. Higher coefficient of variation in speed aggregated for 15 minute intervals at crash stations was found to cause more peak period crashes. In general crashes are associated with higher coefficient of variation in speeds which is supported by studies conducted by Abdel-Aty et al., 2004, and Lee et al., 2003. Coefficient of variation is defined as standard deviation divided by mean for a given data set. So for high coefficient of variation of speed, the denominator (mean) will be low. And it is expected that whenever speeds vary highly from the mean speed at a given stretch, there will be high likelihood for a crash occurrence. The low speeds can be seen as an

indication of peak period during day time during which there is high likelihood of crashes, in particular rear-end crashes.

4.5.3 Category 3

This category also comprises two sub categories, dry pavement crashes and wet pavement crashes. A SUNB model was estimated for this category. Two left-hand side (dependent) variables were considered: 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. To deal with this problem and to evaluate whether the parameter estimates were improving reflected in terms of reduced standard errors, the correlation coefficient was set to one and then the models were estimated. Some of the parameter estimates were improved and hence the SUNB model for these two crash models was included in the thesis. The following sub-sections include the development of individual models followed by the SUNB model.

4.5.3.1 Individual Dry Pavement Crash Model

Before arriving at the final model, two models were tried, 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 dry pavement 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 statistical measures. The dry pavement crash model was selected based on the criteria illustrated in modeling approach chapter. It has a log-likelihood value of -1289.30. The estimation results are provided in Table 4-15.

Table 4-15: Estimation results for individual dry pavement crash model

Parameter	Estimate	Standard Error
CONSTANT	0.182832	0.929752
RADCAT	0.311586	0.173398
MTYPCAT	-0.39893	0.155185
PSURCAT	0.708082	0.217268
OFFRCAT	0.433264	0.115608
ONRACT	0.328742	0.112156
AADT	0.233139	0.081307
ALPHA_1	0.131774	0.025775

Log Likelihood: -1289.30 No. of observations: 552

The individual dry pavement crash model consists of presence of curve, median type, pavement surface type, presence of off-ramp/on-ramp, and AADT, as the independent variables. All variables have reasonable sign and found significant at 90% confidence level. These variables are almost the same as found in overall model (Table 2-3) except for AADT. The dispersion factor ALPHA_1 is considerably different from zero, which confirms the aptness of negative binomial model. No microscopic traffic variable was found to be significant in the model.

4.5.3.2 Individual Wet Pavement Crash Model

As explained in dry pavement 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. AADT was found to be significant in the model. Also there were no significant microscopic traffic variables. It has a log-likelihood value of -709.67. The estimation results are provided in Table 4-16.

Table 4-16: Estimation results for individual dry pavement crash model

Parameter	Estimate	Standard Error
CONSTANT	-1.16808	1.186761
RADCAT	0.385446	0.215415
OFFRCAT	0.507188	0.146006
ONRCAT	0.575273	0.139426
AADT	0.214892	0.105738
ALPHA_2	0.282005	0.09881

Log Likelihood: -709.67 No. of observations: 552

The individual dry pavement crash model consists of presence of curve, presence of off-ramp/on-ramp, and AADT as the independent variables. All variables have plausible sign and found significant at 90% confidence level. The significant variables in this model can be found in overall model (Table 2-3) except for AADT. The dispersion factor ALPHA_2 is considerably different from zero, which confirms the appropriateness of negative binomial model.

4.5.3.3 Seemingly Unrelated Negative Binomial Model for Dry and Wet Pavement Crashes

SUNB estimation was performed for dry and wet pavement crash models. Dispersion parameters (ALPHA 1 and ALPHA 2) were evaluated. As the correlation coefficient (RHO_U1U2) is set at 1 and then the SUNB is estimated, there is no bivariate distribution for the standard deviations (SIGMA_U1 and SIGMA_U2). In this case there exists only a univariate distribution, although with different scales in dry and wet pavement crash models. This can be called as a single factor model. The estimation results are provided in Table 4-17, and 4-18.

Table 4-17: SUNB model estimation results for dry pavement crash model

Parameter	Estimate	Standard Error
CONSTANT	0.413626	0.954578
RADCAT	0.330581	0.169984
MTYPCAT	-0.41797	0.159693
PSURCAT	0.675354	0.230946
OFFRCAT	0.434135	0.105689
ONRACT	0.365302	0.134585
AADT	0.239668	0.074348
ALPHA_1	0.413626	0.954578

Log likelihood: -1963.514 No. of observations: 552

Table 4-18: SUNB model estimation results for wet pavement crash model

Parameter	**Estimate**	**Standard Error**
CONSTANT	-0.86231	1.29415
RADCAT	0.440508	0.220841
OFFRCAT	0.478528	0.146127
ONRCAT	0.644333	0.171139
AADT	0.235953	0.101984
ALPHA_2	0.200808	0.078951

Log Likelihood: -1963.514 No. of observations: 552

Since the correlation between the disturbance terms is substantially high and fixed at a value of 1, it implies that the model disturbances arising from the omitted variables have the same distribution for dry and wet pavement crashes. Through the estimation of SUNB models for dry and wet pavement crashes, some of the errors were decreased. The reliability of the models increases through smaller standard errors. A comparison of standard errors between individual models and SUNB models would help in understanding the efficiency gained. But it has to be remembered that more research has to be done into this type of estimation where the correlation coefficient is extremely high, to actually prove the validity of the simultaneous model. The case in which the efficiency is gained is highlighted in the tables. This comparison table is provided in Table 4-19 and 4-20.

Table 4-19: Comparison of standard errors between individual and SUNB models

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.173398	0.169984
MTYPCAT	0.155185	0.159693
PSURCAT	0.217268	0.230946
OFFRCAT	0.115608	0.105689
ONRACT	0.112156	0.134585
AADT	0.081307	0.074348

Table 4-20: Comparison of standard errors between individual and SUNB models

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.215415	0.220841
OFFRCAT	0.146006	0.146127
ONRCAT	0.139426	0.171139
AADT	0.105738	0.101984

As shown in Tables 4-19 and 4-20, for most of the variables the error terms are less for SUNB model in case of dry pavement crash model. The parameter estimates improved for three out of six. In case of wet pavement model, there is no much improvement in most of the variables.

4.5.3.4 Discussion of Results

Most of the variables that were found significant in overall model (Table 2-3) were also found to affect both dry and wet pavement crashes. The significant factors in dry pavement crash model were road curvature, median type, pavement surface type, presence of on-ramps/off-ramps, and AADT. In the case of wet pavement crash model, the significant factors were road curvature, presence of on-ramps/off-ramps and AADT. Thus, the common factors influencing both these crashes were road curvature, presence of on-ramps/off-ramps, and AADT. Both dry and wet pavement crashes are more likely to occur on sections with relatively sharp curves as vehicles find it difficult to maintain control on such corridors. The present study found that a radius of freeway section less than 3000 ft plays a significant role in the occurrence of crashes. Freeway sections having medians with no barriers were found to have a higher number of dry pavement crashes, confirming the findings of Souleyrette et al., 2001, in which the study found medians without barriers cause more crashes in general. The presence of concrete pavement surface type is found to cause more dry pavement crashes than the combination of asphalt surface. This can be attributed to the inherent smoothness of the asphalt pavements, a key to maintain vehicle's stability. Other possible reasons for Asphalt surfaces being involved in lesser number of crashes could be due to better visibility on asphalt surfaces, as asphalt

reflects lesser light than the concrete counterpart. Also asphalt pavement might have other advantages including better drainage and less noise. However, this study points to more research that needs to be conducted into the influence of pavement surface types on crash occurrence. Pavement type was not significant in wet pavement crashes, making both types of pavements i.e., concrete and asphalt pavements behave in a similar manner. It was found that more dry and wet pavement crashes occur at locations having on-ramps or off-ramps in their vicinity since there will be conflict among the vehicles around merge and diverge areas of the freeway during both day and dark hour time periods. It can be seen that presence of on-ramp has more effect on wet pavement crashes when compared to off-ramp presence. For dry pavement crashes the effect is almost the same.

As AADT increases, both dry and wet pavement crashes are more likely to occur. As the number of vehicles on a given stretch of highway increase, chances of collision among vehicles increase leading to both dry and wet pavement crashes.

4.5.4 Category 4

Based on availability of daylight, there are two sub-categories, day and dark hour crashes. A SUNB model was estimated for this category. Two left-hand side (dependent) variables were considered: day and dark hour crashes. The right-hand side (independent) variables consisted of traffic, roadway and geometric factors. The microscopic traffic factors included in these models were obtained separately for day 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 day is

counted from 5:30 A.M. to 7:00 P.M. during summer and 6:30 A.M. to 5:30 P.M. during winter.

4.5.4.1 Individual Daytime Crash Model

Before arriving at the final model, two models were tried, 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 day time 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 statistical measures. The variables were selected based on the criteria illustrated in modeling approach chapter. It has a log-likelihood value of -1211.34. The estimation results are provided in Table 4-21.

Table 4-21: Estimation results for individual daytime 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

The individual daytime crash model consists of presence of curve, median type, pavement surface type, presence of off-ramp/on-ramp, AADT, and coefficient of variation in speed during daytime aggregated at 15-minute intervals, as the independent variables. All variables have reasonable sign and found to be significant at 90% confidence level. These variables are almost the same as found in overall model (Table 2-3) except for AADT and coefficient of variation in speed at 15 minute aggregation level. The dispersion factor ALPHA 1 is considerably different from zero, which confirms the appropriateness of negative binomial model. No other microscopic traffic variable was found to be significant in the model.

4.5.4.2 Individual Dark Hour Crash Model

As explained in day time 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. PEAKFIFT was not found to be significant and AADT was found to be significant. Hence the model with AADT was chosen as the dark hour crash model. Also there were no significant microscopic traffic variables which were extracted from loop detector data. It has a log-likelihood value of -859.649. The estimation results are provided in Table 4-22.

Table 4-22: 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

The individual dark hour crash model consists of presence of curve, median type, pavement surface type, presence of off-ramp/on-ramp as the independent variables. All variables have reasonable sign and found to be significant at 90% confidence level. The significant variables in this model can be found in overall model (Table 2-3) except for AADT. The dispersion factor ALPHA 2 is considerably different from zero, which proves the correctness of negative binomial model.

4.5.4.3 Seemingly Unrelated Negative Binomial Model for Day and Dark Hour Crashes

As mentioned previously, SUNB estimation was performed for day 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 4-23, 4-24, and 4-25.

Table 4-23: 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 4-24: 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 4-25: 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

As shown in Table 4-25, 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 reliability of the models was increased which is shown by smaller standard errors (highlighted rows). A comparison of standard errors between individual models and SUNB models would help in understanding the efficiency gained. This comparison table is provided in Tables 4-26 and 4-27.

Table 4-26: 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 4-27: 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

As observed from Tables 4-26 and 4-27, most of parameter coefficients in SUNB models have smaller standard errors.

4.5.4.4 Discussion of Results

Most of the variables which were found significant in overall model (Table 2-3) were also found to affect both day and dark hour crashes. AADT was included in the

model as crash volumes immediately preceding the crash were not available. The significant factors in 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, presence of on-ramps/off-ramps and AADT. So the common factors influencing both these crashes were road curvature, median type, pavement surface type, presence of on-ramps/off-ramps, and AADT. Both day and dark hour vehicle crashes are more likely to occur on sections with relatively sharp curves as vehicles find it difficult to maintain control on such corridors. The present study found that a radius of freeway section less than 3000 ft plays a significant role in the occurrence of crashes. Freeway sections having medians with no barriers were found to have a higher number of both day and dark hour vehicle crashes, which confirms the findings of Souleyrette et al., 2001 based on crashes in general, without categorizing them. The presence of concrete pavement surface type is found to cause more day and dark hour crashes than the combination of asphalt surface. This can be attributed to the inherent smoothness of the asphalt pavements, a key to maintain vehicle's stability. Other possible reasons for Asphalt surfaces being involved in lesser number of crashes could be due to better visibility on asphalt surfaces, as asphalt reflects lesser light than the concrete counterpart. Also asphalt pavement might have other advantages including better drainage and less noise. Nevertheless this study points to additional research that needs to be conducted into the influence of pavement surface types on crash occurrence. It was found that more day and dark hour crashes occur at locations having on-ramps or off-ramps in their vicinity since there will be conflict

among the vehicles around merge and diverge areas of the freeway during both day and dark hour time periods. As AADT increases, both day and dark hour crashes are more likely to occur. As the number of vehicles on a given stretch of highway increase, chances of collision among vehicles increase leading to both day and dark hour crashes. Higher coefficient of variation in speed aggregated for 15 minute intervals at crash stations was found to cause day time crashes. In general crashes are associated with higher coefficient of variation in speeds which is supported by studies conducted by Abdel-Aty et al., 2004, and Lee et al., 2003. Coefficient of variation is defined as standard deviation divided by mean for a given data set. So for high coefficient of variation of speed, the denominator (mean) will be low. And it is expected that whenever speeds vary highly from the mean speed at a given stretch, there will be high likelihood for a crash occurrence. The low speeds can be seen as an indication of peak period during day time during which there is high likelihood of crashes, in particular rear-end crashes.

4.5.5 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 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 PDO and injury crashes simultaneously. To handle this problem and to evaluate whether the models are improving reflected in terms of reduced standard errors and goodness-of-fit, the correlation coefficient was set to one and then the models

were estimated. The following sub-sections include first the development of individual models followed by the SUNB model.

4.5.5.1 Individual PDO Crash Model

Before arriving at the final model, two models were tried, 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 dry pavement 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 statistical measures. The PDO crash model was selected based on the criteria illustrated in modeling approach chapter. It has a log-likelihood value of -1003.55. The estimation results are provided in Table 4-28.

Table 4-28: Estimation results for individual PDO crash model

Parameter	Estimate	Standard Error
CONSTANT	0.486784	1.058839
RADCAT	0.367864	0.201882
MTYPCAT	-0.49183	0.175645
PSURCAT	0.666785	0.250662
OFFRCAT	0.503047	0.131273
ONRACT	0.344526	0.125107
AADT	0.120023	0.094137
ALPHA_1	0.207102	0.048315

Log Likelihood: -1003.55 No. of observations: 552

The individual PDO crash model consists of presence of curve, median type, pavement surface type, presence of off-ramp/on-ramp, and AADT, as the independent variables. All variables have reasonable sign and found significant at 90% confidence

level. These variables are almost the same as found in overall model (Table 2-3) except for AADT. The dispersion factor ALPHA_1 is considerably different from zero, which confirms the appropriateness of negative binomial model. No microscopic traffic variable was found to be significant in the model.

4.5.5.2 Individual Injury Crash Model

As explained in PDO 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. Only AADT was found to be significant in the model. Also there were no significant microscopic traffic variables. It has a log-likelihood value of -1166.0827. The estimation results are provided in Table 4-29.

Table 4-29: Estimation results for individual injury crash model

Parameter	Estimate	Standard Error
CONSTANT	-0.51724	0.939364
RADCAT	0.28291	0.183739
PSURCAT	0.613131	0.216604
OFFRCAT	0.391574	0.115474
ONRCAT	0.402771	0.113789
AADT	0.30742	0.081933
ALPHA_2	0.125468	0.030096

Log Likelihood: -1166.08 No. of observations: 552

The individual injury crash model consists of presence of curve, pavement surface type, presence of off-ramp/on-ramp, and AADT as the independent variables. All variables have plausible sign and found significant at 90% confidence level. The significant variables in this model can be found in overall model (Table 2-3) except for

AADT. The dispersion factor ALPHA_2 is considerably different from zero, which confirms the suitability of negative binomial model.

4.5.5.3 Seemingly Unrelated Negative Binomial Model for PDO and Injury Crashes

SUNB estimation was performed for PDO and injury crash models. Dispersion parameters (ALPHA 1 and ALPHA 2) were evaluated. As the correlation coefficient (RHO_U1U2) is set at 1 and then the SUNB is estimated, there is no bivariate distribution for the standard deviations (SIGMA_U1 and SIGMA_U2). In this case there exists only a univariate distribution, although with different scales in PDO and injury crash models. This can be called as a single factor model. The estimation results are provided in Table 4-30, and 4-31.

Table 4-30: SUNB model estimation results for PDO crash model

Parameter	Estimate	Standard Error
CONSTANT	0.681702	1.08623
RADCAT	0.356391	0.192706
MTYPCAT	-0.57902	0.208886
PSURCAT	0.695876	0.283399
OFFRCAT	0.507043	0.128499
ONRCAT	0.416046	0.179167
AADT	0.167241	0.089807
ALPHA_1	0.173509	0.041809

Log Likelihood: -2120.684 No. of observations: 552

Table 4-31: SUNB model estimation results for injury crash model

Parameter	Estimate	Standard Error
CONSTANT	-0.27939	0.976893
RADCAT	0.284644	0.174296
PSURCAT	0.582664	0.255877
OFFRCAT	0.409808	0.114811
ONRCAT	0.474826	0.166651
AADT	0.316334	0.080536
ALPHA_2	0.103297	0.026983

Log Likelihood: -2120.684 No. of observations: 552

Since the correlation between the disturbance terms is substantially high and fixed at a value of 1, it suggests that the model disturbances developing from the omitted variables have the same distribution for both PDO and injury crashes. Through the estimation of SUNB models, the errors were decreased for some of the variables. A comparison of standard errors between individual models and SUNB models would help in understanding the efficiency gained. But it has to be remembered that more research has to done into this type of estimation where the correlation coefficient is extremely high, to actually prove the soundness of the simultaneous model. The case in which the efficiency is gained is highlighted in the tables. This comparison table is provided in Table 4-32 and 4-33.

Table 4-32: Comparison of standard errors between individual and SUNB models

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.201882	0.192706
MTYPCAT	0.175645	0.208886
PSURCAT	0.250662	0.283399
OFFRCAT	0.131273	0.128499
ONRACT	0.125107	0.179167
AADT	0.094137	0.089807

Table 4-33: Comparison of standard errors between individual and SUNB models

Parameter	Std error for individual model	Std error for SUNB model
RADCAT	0.183739	0.174296
PSURCAT	0.216604	0.255877
OFFRCAT	0.115474	0.114811
ONRACT	0.113789	0.166651
AADT	0.081933	0.080536

As shown in Tables 4-32 and 4-33, there has been improvement for three variables in case of PDO crash model and for three out of five in injury crash model.

4.5.5.4 Discussion of Results

Most of the variables which were found significant in the overall model (Table 2-3) were also found to affect both PDO and injury crashes. The significant factors in PDO crash model were road curvature, median type, pavement surface type, presence of on-ramps/off-ramps, and AADT. In the case of injury crash model, the significant factors were road curvature, pavement surface type, presence of on-ramps/off-ramps and AADT. So the common factors influencing both these crashes were road curvature, pavement surface type, presence of on-ramps/off-ramps, and AADT. Both PDO and injury crashes are more likely to occur on sections with relatively sharp curves as vehicles find it difficult to maintain stability on such corridors. The present study found that a radius of freeway section less than 3000 ft plays a significant role in the occurrence of crashes. Freeway sections having medians with no barriers were found to have a higher number of PDO crashes, which confirms the findings of Souleyrette et al., 2001 based on crashes in general, without categorizing them. It is believed that drivers tend to be less cautious while driving around medians without barriers. The presence of concrete pavement surface type is found to cause more of PDO and injury crashes than the combination of

asphalt surface. This can be attributed to the inherent smoothness of the asphalt pavements, a key to maintain vehicle's stability. Other possible reasons for Asphalt surfaces being involved in lesser number of crashes could be due to better visibility on asphalt surfaces, as asphalt reflects lesser light than the concrete counterpart. Also asphalt pavement might have other advantages including better drainage and less noise. However, this study points to more research that needs to be conducted into the influence of pavement surface types on crash occurrence. It was found that more PDO and injury crashes occur at locations having on-ramps or off-ramps in their vicinity since there will be conflict among the vehicles around merge and diverge areas of the freeway. As AADT increases, both PDO and injury crashes, are more likely to occur. As the number of vehicles on a given stretch of highway increase, chances of collision among vehicles increase leading to both PDO and injury crashes.

4.6 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 (Greene, 1997), 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 derive at the model with better goodness-of-fit. The following table provides the details of the G^2 statistic for various individual and SUNB models for the five categories.

Table 4-34: 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 3	Dry Pavement	**5388.35**	Dry Pavement	**5376.65**
	Wet Pavement	**1447.85**	Wet Pavement	**2630.6**
Category 4	Daytime	**3257.90**	Daytime	**5826.67**
	Dark Hour	**4217.21**	Dark Hour	**5388.35**
Category 5	PDO	**2838.88**	PDO	**6239.56**
	Injury	**4496.995**	Injury	**7706.75**

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 off-peak period crash model
- Dry pavement crash model improved with SUNB estimation, while there was no improvement in wet pavement crash model
- There was no improvement in both daytime and dark hour crash models with SUNB estimation
- There was no improvement in both PDO and injury crash models with SUNB estimation

Even though goodness-of-fit statistics does 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.

4.7 Conclusions

The analyses in this chapter proved that microscopic crash frequency modeling in terms of both using microscopic traffic factors and categorizing the crashes resulted in improved crash frequency models. The research has also revealed that simultaneous estimation of these different categories using seemingly unrelated negative binomial regression produced enhanced models in terms of better parameter estimates and better goodness-of-fit. Investigating the techniques to handle correlation between the disturbance terms was an important part of the research endeavor. With the help SUNB estimation for different categories of crashes, this research study paved the way for better identification of factors related to crash occurrence, occurring in different environments, different traffic conditions, and different styles. Future work could be to add more independent variables in the models to avoid the difficulties in estimating SUNB models with high correlation between the error terms. Also it is suggested that more work has to be done regarding SUNB estimation for models with high correlation coefficient.

5 APPLICATION OF LOGISTIC REGRESSION MODEL TO OBTAIN RAINFALL INFORMATION ON INTERSTATE-4 IN CENTRAL FLORIDA

5.1 Introduction

Weather related information is considered as one of the important factors in road safety analyses. Adverse weather conditions contribute to crashes by impairing visibility, reducing stability and decreasing controllability. According to a report on crashes on U.S. highways, over 22% of the total crashes in 2001 were weather-related (Goodwin, 2003). Figure 5-1 shows the nationwide average number of injury and fatal crashes that occurred during adverse weather conditions between 1995 and 2001 (Goodwin, 2002). It is clear from the figure that most of the accidents occur during rain. Drivers experience low visibility and reduced control of the vehicle in rain. Rain decreases the friction between pavement and tires and thereby making it difficult for the vehicle to stop at a desired distance. Several studies, in fact, concluded that crashes increase during rainfall by 100 percent or more (Brodsky and Hakkert, 1988; Bertness, 1980; NTSB, 1980), while others find more moderate (but still statistically significant) increases (Andreescu and Frost, 1998; Fridstrom et al., 1995; Andrey and Olley, 1990). Of two studies that focus specifically on *fatal* traffic crashes, one finds an increase in the crash rate of over 100 percent during rainy conditions (Brodsky and Hakkert, 1988), and the other finds an increase in one country (Denmark) and no significant change in two other countries (Norway and Sweden) (Fridstrom et al, 1995). Keeping in mind the above facts there is a need to determine the significance of rainfall on crash occurrence on I-4 which directs us towards finding the rainfall information that can be used in safety analyses. The present

report is an effort to obtain the rainfall information on a 36 mile stretch of I-4 in Central Florida.

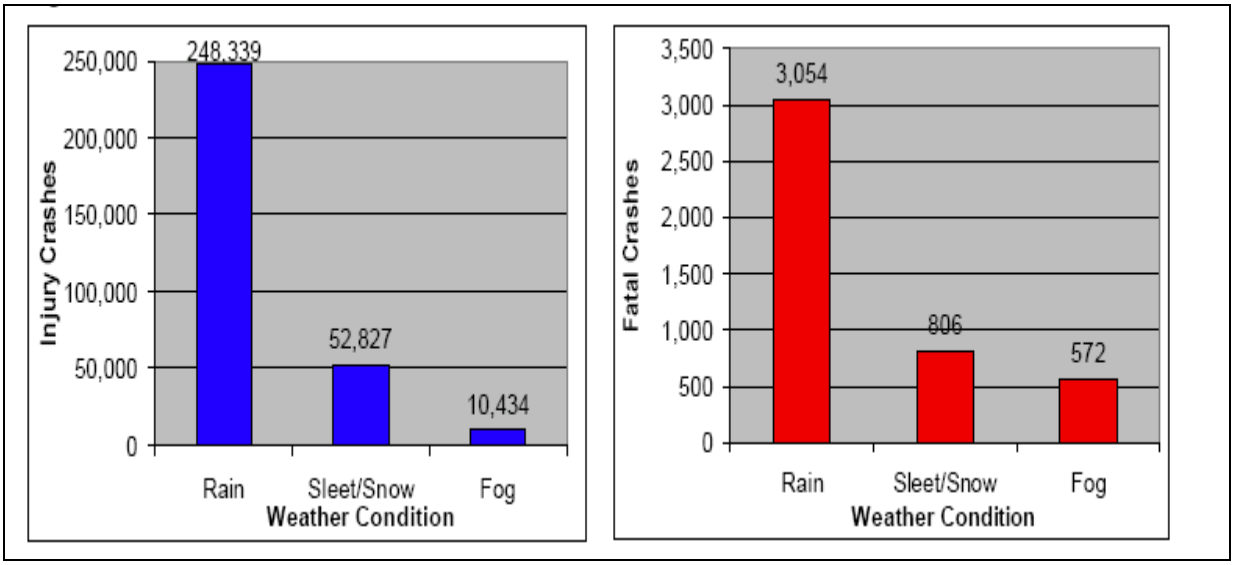

Figure 5-1: Average Injury and Fatal Crashes in Adverse Weather Conditions (Goodwin, 2002)

5.2 Background and Data Collection

The objective of this effort is to obtain the weather condition ("rain" or "no rain") at a given time and location on I-4 in Central Florida so that this information can be used in traffic safety analyses. In the study area there are no weather monitoring stations located on I-4, which can provide the exact rainfall information at a desired time and location. Alternatively the Florida crash database provides the exact weather condition at the time of crash on I-4. Also many safety studies use only the crash cases in their analysis. Then a question may arise as to what is the need to obtain the weather condition at a time other than the time of crash. The answer lies in the fact that not all safety analyses use only the crash cases on a particular roadway, but some use both the crash and non-crash cases in their analysis. A non-crash case can be defined as when there is no

crash occurrence at a particular time and location on a given roadway. For instance a safety study may use the binary logit model with a response variable containing both crash and non-crash cases. Now the task is to obtain the weather condition for the non-crash cases. Essentially the aim is to obtain weather 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 5-1. Out of 1964 crash cases that happened during 1999 through 2001, 217 of them occurred during rain, which comes to 11 percent of the total number of crashes. This is a significant percentage of rainfall occurrence which explains the need to develop a model to obtain the rainfall condition for crash and non-crash cases which in turn helps to see the effect of rainfall on crash occurrence.

Table 5-1: Number of crashes occurred during rain during 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

Initially to start with the process, various agencies were contacted to obtain any kind of rainfall information. The main aim was to obtain rainfall information for I-4 at a desired time and location. The agencies contacted to obtain the information are listed below.

- Municipalities
- Saint Johns River Water Management District
- Federal Aviation Administration

- Melbourne office of the Climate Record Center
- Earthinfo
- South Eastern Climate Record Center
- Florida Automated Weather Network (FAWN)
- Airports
- National Oceanic and Atmospheric Administration (NOAA)

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, at address 2725 Binion Road, Apopka, FL and 17498 McKinney Rd, Winter Garden, 34787. *NOAA* provided access to their database that consisted of 15 minute and hourly rainfall totals. The database is a more complete set than the Southeastern Climate Record Center as it is the National Climate Center. The only fifteen minute site around Orlando is in Lake County, well to the north and west of the City Beautiful. Finally the hourly rainfall information for the weather stations located at Orlando International Airport, Executive Airport and Sanford Airport were obtained from NOAA. To summarize the whole rainfall data collection process, there was no rainfall information available for I-4. But 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.

The weather stations on the western side of I-4 are located at:

1) Apopka
2) Avalon

The weather stations on the eastern side of I-4 are located at:

1) Orlando International Airport
2) Executive Airport
3) Sanford Airport

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 report deals with the development of this logistic regression model. The model details are explained at full length in the later chapters of this report. A map showing the locations of the five weather stations surrounding I-4 is provided in Figure 5-2.

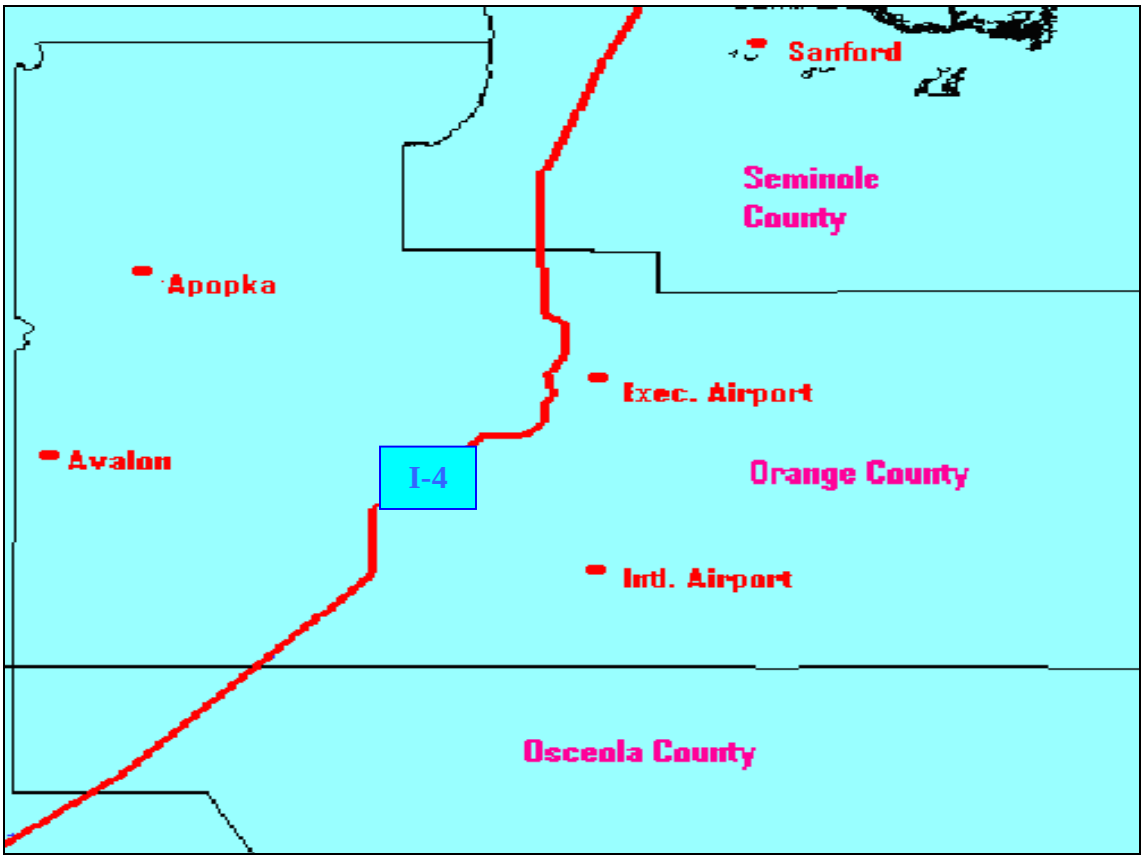

Figure 5-2: Map showing locations of the five weather stations surrounding Interstate 4 in Central Florida.

5.3 Methodology and Data Preparation

Now the question is how to get rainfall information at a given time and at a specified location on I-4 in Central Florida using the information from the five weather stations surrounding the freeway. The study uses a simple logistic regression model for this purpose. 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 (Washington et al., 2003). The dependent variable in the case of logistic regression is the probability (P) that the resulting outcome is equal to 1.

Thus, 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,……,n for a set of n independent variables.

So P_i can be written as

$P_i = \text{Exp}(\beta_0 + \beta_i X_i) / 1 + \text{Exp}(\beta_0 + \beta_i X_i)$

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.

5.3.1 Dependent Variable

In the study area, a total of 1964 crashes 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 build 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-mile stretch numbered from 2 to 71. For each crash case, the nearest loop station is identified as the crash location. A sample of the information prepared as explained in above paragraph is provided in Table 5-2.

Table 5-2: Sample weather information extracted from the crash database

Sl No	Time of Crash	Station/ Location of 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	20:52:00	9	4/2/1999	CLEAR
9	18:16:00	44	4/2/1999	CLEAR

So in Table 5-2, the weather condition is 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.

5.3.2 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 crash station and a weather station. This particular order is explained with the help of Table 5-3. Table 5-3 provides a sample of independent variables entered in the model.

Table 5-3: 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
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
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

* The units for rain_1 – rain_5 are in inches/hour

In Table 5-3, 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. The source of the response variable is the crash database since the police officer identifies the rain condition at the time and location of the crash. The source of the independent variable is the rain data at the weather stations. Rain_1 contains the rain information at the first 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 forth. For instance, the first independent variable for the crash that happened on 10/19/2001 at time 18:41:00 and at station 4 has 0.02 inches of rainfall and is the first 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. To get the rainfall information for each of the independent variables, the following procedure was followed.

1) At first the geographical co-ordinates for all the 69 crash stations and the five weather stations were obtained. The geographical (x, y) co-ordinates for the 69 crash stations and the five weather stations are provided in Table 5-4 and Table 5-5 respectively.

Table 5-4: Geographical co-ordinates of the Crash Stations

Crash Station	X Co-ordinate	Y Co-ordinate
2	445269.79	3132580.78
3	446038.86	3133549.48
4	446548.53	3134154.82
5	447129.48	3134846.29
6	447679.62	3135504.61
7	448323.89	3136312.44
8	449253.22	3137470.71
9	449739.74	3138078.07
10	450299.02	3138780.05
11	450889.97	3139503.99
12	451470.38	3140235.8
13	452267.09	3141236.99
14	452854.87	3141974.58
15	453431.63	3142710.26
16	453567.45	3143541.53
17	453568.26	3144552.32
18	453569.81	3145418.11
19	453570.1	3146286.4
20	453573.58	3147114.3
21	453676.13	3148027.62
22	454191.61	3148569.44
23	454694.76	3148948.38
24	455513.74	3149832.1
25	455989.87	3150278.79
26	456616.79	3150860.65
27	457227.56	3151445.97
28	457793.42	3152174.51
29	458278.63	3152798.39
30	459058	3153616.62
31	459935.32	3153687.21
32	460598.26	3153686.99
33	461349.9	3153898.32
34	461860.8	3154286.94
35	462224.81	3154971.63
36	462217	3155790.97
37	462508.13	3156521.89
38	462645.71	3157067.97
40	462596.96	3157857.15

41	462641.53	3158703.39
42	462961.69	3159486.35
43	463263.12	3160176.06
44	463273.69	3160987.2
45	463324.18	3161800.55
46	463013.92	3162606.49
47	462548.63	3163059.81
48	462241.75	3163697.25
49	462223.89	3164419.38
50	462219.38	3165088.78
51	462220.98	3165817.46
52	462211.45	3166676.57
53	462118.48	3167393.41
54	462098.19	3168230.88
55	462047.75	3169032.35
56	462044.1	3169908.5
57	462043.33	3170620.22
58	462037.63	3171806.41
59	462041.16	3172657.68
60	462035.98	3173574.8
61	462229.97	3174430.19
62	462721.79	3175278.54
63	462906.33	3175840.04
64	463173.25	3176690.67
65	463397.48	3177400.98
66	463719.6	3178416.2
67	463902.28	3178997.2
68	464126.03	3179704.84
69	464356.37	3180435.71
70	464530.92	3180990.49
71	464926.58	3182244.32

* All co-ordinates are in meters

Table 5-5: Geographical Co-ordinates of the Weather Stations

Weather Station	**X Co-ordinate**	**Y Co-ordinate**
Orlando International Airport	469053.8839	3144755.935
Orlando Sanford Airport	476818.1376	3183369.139
Orlando Executive Airport	467434.6421	3157677.817
Apopka	446319.4681	3168472.523
Avalon	436507.9272	3149777.686

* All co-ordinates are in meters

2) Based on these co-ordinates, the distance between a crash 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 first nearest weather station is put first, the second nearest is put second and so on. Table 5-6 provides this order for some of the stations.

Table 5-6: Order of Weather Stations based on the distance from Crash stations

Crash Station	First	Second	Third	Fourth	Fifth
2	5*	1*	3*	4*	2*
3	5	1	3	4	2
4	5	1	3	4	2
5	5	1	3	4	2
6	5	1	3	4	2
7	5	1	3	4	2
8	5	1	3	4	2
9	5	1	3	4	2
10	5	1	3	4	2
11	5	1	3	4	2
12	5	1	3	4	2

*5 – Avalon, 4 – Apopka, 3 - Orlando Executive Airport, 2 – Sanford Airport, 1 – Orlando International Airport

3) Also tables were prepared for each of the five weather stations separately for each year (1999 – 2001), consisting of rainfall information. A sample table for Avalon station for the year 1999 is provided in Table 5-7.

Table 5-7: Rainfall Information at Weather Station Avalon

Time	Rainfall
1/1/1999 12:15	1.05
1/1/1999 12:30	0.08
1/1/1999 12:45	0.02
1/1/1999 15:00	0
1/1/1999 15:15	0.02

*rainfall is in inches/hour

Now using the information in Tables 5-2, 5-6 and 5-7, the rain values are entered in Table 5-3 using a program developed in Visual Basic. For example, let us take a crash that happened on 10/19/2001 at time 18:41:00 and at station 4. We first go to Table 5-6 and take the order of the weather stations located from the crash station 4. So the first 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 19th October at time 18:41:00 and put this value in Table 5-3 for the rain_1.

5.4 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 are provided in Table 5-8. The test has a null hypothesis that the variables are independent against the alternate hypothesis with the case that the variables are not independent. The test was conducted at a 95% confidence level. The test statistic and the p-value are provided in Table 5-8. Seeing at the p-value which is very less than 0.05, the null hypothesis was rejected. That means the variables cannot be considered as independent. More information on this test can be obtained from Rencher (2002).

Table 5-8: Chi-Square test of Independence of Variables

Test-statistic	P-value
84.325778	7.083E-14

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. A note on Principal Component Analysis would be useful to better understand the process.

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. Rencher (2002) can be referred for more information on Principal component analysis.

Now 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. The rules are:

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.

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 5-9 & 5-10 and Figure 5-3. Table 5-9 provides the covariance matrix of the independent variables from which the eigen values and eigen vectors are calculated. Table 5-10 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 5-10, 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 5-10, the average of Eigen values is 0.00830027 and only 2 Eigen values exceed this value. So two principal components can be retained.

Table 5-9: Results from 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 5-10: Results 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

For rule 3, looking at Figure 5-3 and retaining the eigen values on the steep part of the curve, four principal components can be retained. To conclude, two out of three rules say that four principal components can be retained. So the first four principal components are used as the independent variables in the logistic regression model.

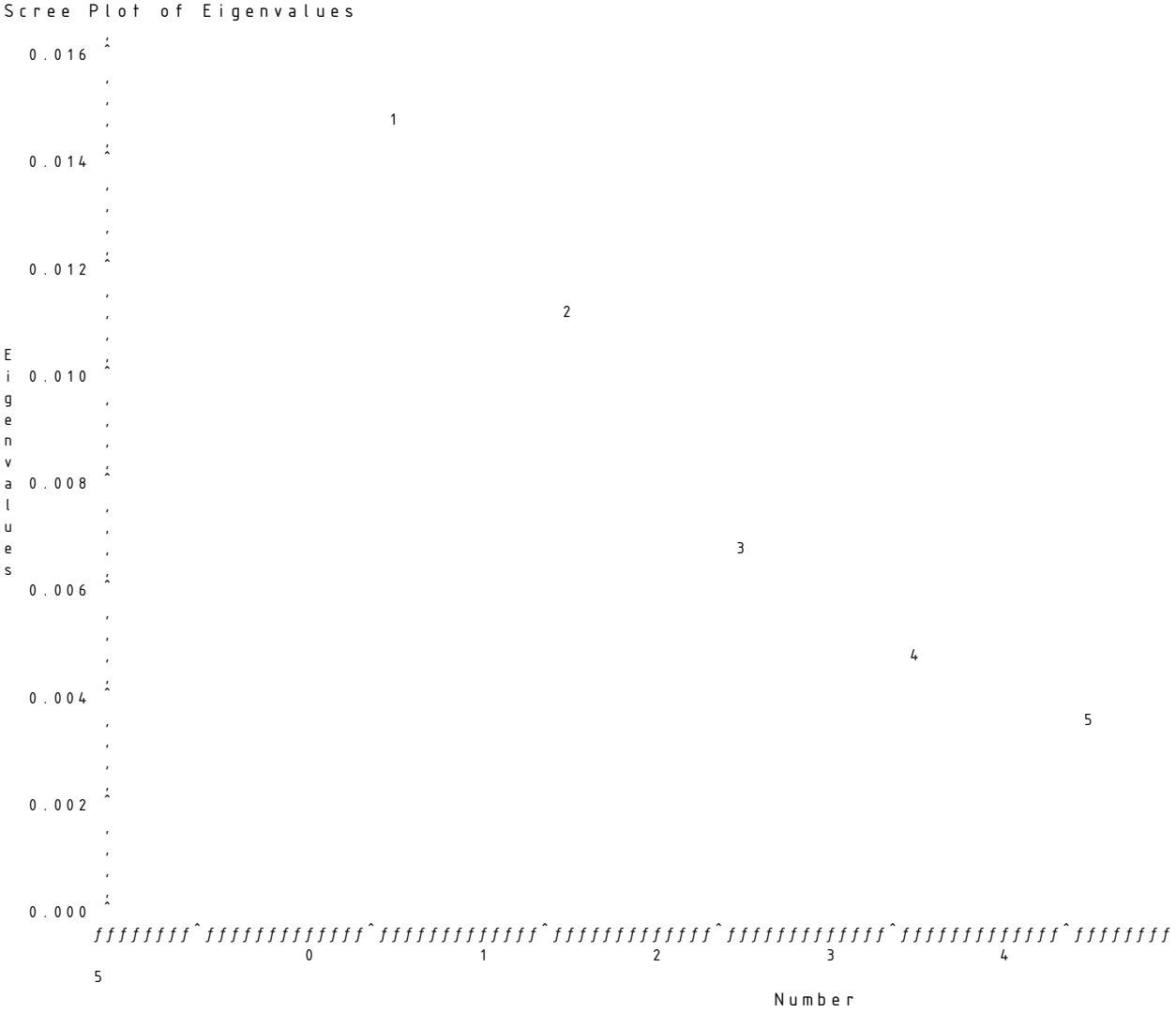

Figure 5-3: Scree Plot from Principal Component Analysis

With the four retained principal components of the variables rain_1 through rain_5, a simple logistic regression model was estimated. The results for the logistic regression model obtained are provided in Table 5-11 and Table 5-12. Table 5-11 provides 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.

Table 5-11: Logistic Regression Model Results

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	910.324	864.188
-2 Log L	908.324	854.188

Table 5-12 provides the parameter estimates of the four principal components used in the model.

Table 5-12: Logistic Regression Model Results

Parameter	DF	Estimate	Standard Error	Pr > ChiSq
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

As shown in Table 5-12, all 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 = e (-2.1444 + 3.3260*Principal component 1 + 1.2645* Principal component 2 + 1.5221* Principal component 3 + 2.3441* Principal component 4) / 1 + e (-2.1444 + 3.3260*Principal component 1 + 1.2645* Principal component 2 + 1.5221* Principal component 3 + 2.3441* Principal component 4)

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

5.5 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 as "rain index" value in the study. So 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 5-13.

Table 5-13: 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

Then a classification table is prepared for the actual and predicted weather conditions for the validation data set to get the prediction accuracy of the model. The classification table is provided in Table 5-14.

Table 5-14: Classification table for the test data

Actual Weather Vs Predicted Weather			
Actual	**Predicted**		
	0	**1**	**Total**
0	532	64	596
1	16	56	72
Total	548	120	668

So the overall prediction accuracy is (532 + 56)/668 = 88.02%. The prediction accuracy for the cases with "rain" is 56/72 = 77.78% and the prediction accuracy for the cases with "no rain" is 532/596 = 89.26%.

5.6 Model Application

The overall prediction accuracy of the model is high, and therefore the model can be used for obtaining the rain index values for a desired time, date and station on 36-mile

stretch of Interstate 4 in central Florida for the non-crash cases. The logistic regression model was applied to a safety database to obtain the rain index values for around 3000 crash and 53000 non-crash cases. The non-crash cases were randomly selected. The safety model which used this safety database, is a logistic regression model, with "crash = 1" and "non-crash = 0" as the dependent variable. The independent variables are average occupancy "AOG2", standard deviation of volume "SVG2" and coefficient of variation of speed "LOGCVSF2". To assess the effect of rainfall on crash occurrence, the rain index values were used as an indication of weather condition and introduced as one of the independent variable under the name "Weather". 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. 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 vise-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 situation. To see if the addition of "Weather" actually enhances the logistic regression model by improving its classification accuracy and goodness-of-fit, we fit two models, one without "Weather" and one with "Weather". The results of the model estimated with the variables discussed in the previous paragraph only, are provided in Table 5-15 and Table 5-16. Table 5-17 and Table 5-18 provide results for the model estimated with the variables discussed in the previous paragraph and the "Weather" variable.

Table 5-15: Model fit statistics for the safety model without "Weather" variable

AIC	11994.87
-2LogL	11986.87
R-square	0.0033

Table 5-16: Parameter estimates of the safety model without "Weather" variable

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

Table 5-15 provides model fit statistics, AIC (Akaike Information Criteria), Log-likelihood value, and R-square value for the model without "Weather" variable. Table 5-16 gives the parameter estimates, standard errors and probability values for the variables used in the model. The probability values suggest a high significance for these variables 95% confidence level.

Table 5-17: Model fit statistics for the safety model with "Weather" variable

AIC	**11951.37**
-2LogL	**11941.37**
R-square	**0.0038**

Table 5-18: Parameter estimates of the safety model with "Weather" variable

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
Weather (rain index)	1.3924	0.267	<.0001

Table 5-17 provides model fit statistics, AIC, Log-likelihood value, and R-square value for the model with the "Weather" variable. Table 5-18 gives the parameter estimates, standard errors and probability values for the variables used in the model. Here also, the probability values imply a high significance for these variables 95% confidence level.

5.6.1 Goodness-of-fit

To ensure that the model fit has improved by adding the "Weather" variable, the following tests were conducted.

- Comparing the AIC values

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 "Weather" variable has an AIC of 11951.37 which is lower than the AIC value of 11994.87 for the model without "Weather" variable.

- Log-likelihood Ratio Test

The Log-likelihood ratio test is based on hypothesis testing. The statistic for this test is the chi-square value determined by the Log-likelihood of the model. The null hypothesis is:

Ho: The model without "Weather" variable is better than the model with "Weather" variable.

The alternative hypothesis is:

Ha: The model with "Weather" variable is not better than the model without "Weather" variable.

The test statistic can be written as (let us assume it as G)

G = 2{[Log-likelihood of model with "Weather"] – [Log-likelihood of model without "Weather"]}

So G = -11941.37 – (-11986.87) = 45.5

=> P-value = Pr (χ_1^2 > G) = << 0.001

The null hypothesis can be rejected, suggesting that the model with "Weather" is a better model.

- Comparing the R-square values

R-square value which is referred to as the "coefficient of determination" in ordinary linear regression is one of the widely accepted measure of good-of-fit. Comparing the R-square values for these two models, the model with "Weather" variable has high value of R-square, indicating a better goodness-of-fit.

5.6.2 Prediction Accuracy

As earlier noted, the model provides only the estimated probability of the event (in this case "crash" or "no crash"). It does not provide the direct outcome. So to get a prediction table or a classification table, a cut-off value for the estimated probability has to be set. When comparing the prediction accuracy of two models, it would not be practical to set the same cut-off value for each of the two models as the estimated probability changes with the input variables. Also there would be a scale problem, as the estimated probability is measured on continuum whereas the outcome is binary. Keeping in mind the above facts, another way of comparing the prediction accuracy was followed.

ROC curve: 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 5-19 for model without "Weather" and in Table 5-20 for model with "Weather".

Table 5-19: Measures of association between the predicted probabilities and observed responses for the model without "Weather"

Somers' D	0.163
Gamma	0.173
Tau-a	0.013
C	0.581

Table 5-20: Measures of association between the predicted probabilities and observed responses for the model with "Weather"

Somers' D	0.178
Gamma	0.189
Tau-a	0.014
C	0.589

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 5-19 and 5-20, the higher "C" value for the model with "Weather" indicates better prediction accuracy for this model.

5.7 Adjustment of Estimated Probabilities

In the present model, the crash to non-crash case ratio is 3000/53000, which is equal to 0.057. But in real life, the ratio would be quite different, because of different number of non-crash cases. The actual number of non-crash cases can be calculated as follows:

The crash cases were analyzed for a period of 4 years from January 1999 to December 2002, for a total of 365+366+365+365 = 1461 days.

For a 24-hour period, each day has 288, five minute intervals (as the safety model uses 5 minute interval traffic data immediately preceding the crash)

There are 69 loop stations in each direction, making a total of 138 stations

Therefore the actual non-crash number would be 1461*288*138-3000 = 58062984

Now the crash to non-crash ratio is 3000/58062984= 0.00005166

This implies that the proportion of crash cases in real life is P1 = 0.00005166, while the proportion of non-crash cases is P2 = 1-0.00005166 = 0.999948. According to the formula developed for adjusting estimated probabilities (Greene, 1997), we have the adjusted probability as follows:

$$AP = \frac{\text{P1*EP}}{\text{P1*EP+P2*(1-EP)}}$$

Where AP is the adjusted probability and EP is actual estimated probability.

Using the above formula, the adjusted probability for the crash occurrence can be calculated. For instance, if the actual estimated probability is 0.05 for a crash occurrence, then the adjusted probability would be:

AP = 0.00005166*0.05/0.00005166*0.05+0.999948* (1-0.05) = 0.0000027

Therefore the adjusted probability would be 0.0000056 for the crash occurrence.

5.8 Conclusions

The study corridor considered in the analysis does not have weather monitoring stations, which can provide the exact rainfall information at a desired time and location. Although the Florida crash database provides the rainfall information for every crash case, it is required to obtain rainfall information for the non-crash cases. Because some safety studies use both crash and non-crash cases in their analysis. Effectively the aim of this research is to obtain weather information at a particular time and location on I-4 other than the time of crash occurrences. Once the model was developed, rain index values were used in the safety model, instead of setting a cut-off value for rainfall occurrence, based on the argument that it is convenient to use the rain index values. Also it was showed that inclusion of rainfall information actually improved the safety model with better prediction accuracy and goodness-of-fit.

6 CONCLUSIONS

This thesis attempts to describe three different efforts related mainly to safety analysis on a 36-mile stretch on Interstate 4 in Central Florida. The research investigated different traffic volume forms to account for the best form to be used in crash frequency analysis and identified the significant factors that affect crash frequencies on freeways, when all crash types were aggregated. In this case, there was at least one crash occurrence at each crash station and allowed for the use of five to fifteen minute volumes immediately preceding the crash occurrence. The results of this study suggested a new and better way for accounting for the effect of traffic volume which is the use of traffic volumes just before a crash when compared to other macroscopic traffic factors, e.g. AADT and VMT. The results showed that road curvature, median type, number of lanes, pavement surface type and presence of on/off-ramps are among the significant factors that contribute to crash occurrence.

The research also investigated the technique to address the problem of correlation between the error terms, when the crashes are divided into different logical categories (for 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 different crash categories.

The first category which 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 multiple vehicle crash model were road curvature, median type, pavement surface type and presence of on-ramps/off-ramps and AADT. In the case of single vehicle crash model, the significant factors were road 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 found that increase in AADT caused more multiple vehicle crashes, while AADT had no effect on single vehicle crashes. This can be justified because 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 behavior.

In category 2 (peak and off-peak period crashes) goodness-of-fit for the SUNB peak period crash model substantially increased 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 (Abdel-Aty et al., 2004).

The third category consisting of dry and wet pavement crashes had a goodness-of-fit improvement in case of dry pavement crashes, while the wet pavement crash model did not improve. The significant factors in dry pavement crash model were road curvature, median type, pavement surface type, presence of on-ramps/off-ramps, and AADT. In the case of wet pavement crash model, the significant factors were road curvature, presence of on-ramps/off-ramps and AADT. Thus, the common factors influencing both these crashes were road curvature, presence of on-ramps/off-ramps, and AADT. Pavement surface type appeared only in the dry pavement crash model. This particular result can be justified based on the explanation that it is quite difficult to differentiate among pavement surface types when the pavement is wet.

The fourth category had the SUNB daytime and dark hour crash models. The significant factors in 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.

The fifth category consisted of PDO and injury crash models. The significant factors in PDO crash model were road curvature, median type, pavement surface type, presence of on-ramps/off-ramps, and AADT. In the case of injury crash model, the significant factors were road curvature, 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.

To summarize, radius category, presence of on-ramps, and presence of off-ramps appeared in all the ten models. AADT was also found to influence all the crash categories except for single vehicle crashes. In case of median type, it 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., 2001 cause more crashes, but wet pavement and injury crashes might be strongly associated with other factors so that median type is not significant in such crashes. Pavement surface type was found 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.

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 resources. A possible extension 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 be to add more independent variables in the models to avoid the difficulties in estimating SUNB models with high correlation between the error terms. Also it is suggested that more work has to be done regarding SUNB estimation for models with high correlation coefficient.

Finally the research developed a logistic regression model to obtain the rainfall information on the same study corridor, so that this information can be used in safety analyses which include both crash and non-crash cases. The research was initiated as the study corridor does not have weather monitoring stations, which can provide the exact rainfall information at a desired time and location. Once the model was developed, rain index values were used in the safety model. Also it was shown that inclusion of rainfall information improved the safety model with better prediction accuracy and goodness-of-fit.

APPENDIX A

VISUAL BASIC CODE USED FOR THE DEVELOPMENT OF WEATHER MODEL

```
Dim rs As New ADODB.Recordset
Dim cn As New ADODB.Connection
Dim rs2 As New ADODB.Recordset
Dim rs3 As New ADODB.Recordset
'Dim nowTime As Date
Private Sub Command1_Click()
Dim nowTime As Variant
Dim nowHour As Variant
Dim rs3_RValues(1) As Double
cn.Open "gen"
mySql = "select * from Cl_svoc_45yars where date >= #1/1/2002#  and cdate <= #31/12/2002# " & "order by date, stationofcrash"
rs.Open mySql, cn, adOpenKeyset, adLockOptimistic
'MsgBox (rs.RecordCount)
rs.MoveFirst
While rsCount < rs.RecordCount
rs_Station = rs.Fields(5)
rs_time = rs.Fields(10)
rs_Minute = Minute(rs.Fields(10))
rs_Prev15 = Int(rs_Minute / 15) * 15
rs_Next15 = (Int(rs_Minute / 15) + 1) * 15
rs_PrevTime = Format(Hour(rs_time) & ":00", "hh:mm")
If Hour(rs_time) <> 23 Then
rs_NextTime = Format(Hour(rs_time) + 1 & ":00", "hh:mm")
Else
rs_NextTime = Format("00:00:00", "hh:mm")
End If
nowDate = CDate(Format(rs.Fields(9), "mm/dd/yyyy"))
If (Hour(rs_time) <> 23) Then
rs_Date = rs.Fields(9)
nextDate = nowDate
Else
currDay = Day(rs.Fields(9))
rs_newDay = currDay + 1
rs_Date = Format(Month(rs.Fields(9)) & "/" & rs_newDay & "/" & Year(rs.Fields(9)), "mm/dd/yyyy")
nextDate = nowDate + 1
nextDate = CDate(Format(nextDate, "mm/dd/yyyy"))
End If
rs_Date = Format(rs_Date, "mm/dd/yyyy")
'MsgBox (rs_Date & " " & rs_PrevTime & " " & rs_NextTime)
prevTime = nowDate & " " & rs_PrevTime
```

```
nextTime = nextDate & " " & rs_NextTime
mySql2 = "select * from nearest where station = " & rs_Station
'MsgBox (mySql2)
rs2.Open mySql2, cn, adOpenKeyset, adLockOptimistic
'MsgBox (rs2.RecordCount)
i = 1
While (rs2.Fields(i + 1) <> 4)
i = i + 1
Wend
rs2.Close
'mysql3 = "select * from Avalon2000 where date > #" & prevTime & "# and date <= #"
& nextTime & "# order by date "
'mySql3 = "SELECT sum(rain) FROM Avalon1999 where date > #" & prevTime & "#
and date<=#" & nextTime & "#"
'mySql3 = "SELECT sum(rain) FROM Avalon2000 where date > #" & prevTime & "#
and date<=#" & nextTime & "#"
'mySql3 = "SELECT sum(rain) FROM Ava2001 where date > #" & prevTime & "# and
date<=#" & nextTime & "#"
'mySql3 = "SELECT sum(rain) FROM Avalon2002 where date > #" & prevTime & "#
and date<=#" & nextTime & "#"
'mySql3 = "SELECT sum(rain) FROM apop1999 where date > #" & prevTime & "# and
date<=#" & nextTime & "#"
'mySql3 = "SELECT sum(rain) FROM apop2000 where date > #" & prevTime & "# and
date<=#" & nextTime & "#"
'mySql3 = "SELECT sum(rain) FROM Apop2001 where date > #" & prevTime & "# and
date<=#" & nextTime & "#"
mySql3 = "SELECT sum(rain) FROM Apop2002 where date > #" & prevTime & "# and
date<=#" & nextTime & "#"
rs3.Open mySql3, cn, adOpenKeyset, adLockOptimistic
'MsgBox (mySql3 & "   " & rs3.Fields(0))
If rs3.RecordCount > 0 Then
rs3.MoveFirst
rs3_RainValue = rs3.Fields(0)
'If (rs3_RValues(2) > 0) Then
'MsgBox (rs3_RValues(2))
'rs3_RainValue = rs3_RValues(2)
'Else
'rs3_RainValue = 0 'rs3_RValues(1)
'End If
Else
rs3_RainValue = 10000
End If
rs3.Close
rs.Fields(10 + i) = CDbl(rs3_RainValue)
rs.Update
```

```
flag = 1
rsCount = rsCount + 1
'Label1.Caption = "Processing Row No." & rsCount & "of " & rs.RecordCount
rs.MoveNext
Wend
MsgBox ("Yahooo")
rs.Close
cn.Close
End Sub
'mySql2 = "select * from 0200 where time >= #" & Format(nowHour & ":00:00", "hh:mm:ss AM/PM") _
'& "# and time <= #" & Format(nowHour + 1 & ":00:00", "hh:mm:ss AM/PM") & "#" & " and date = #" & nowDate & "#"
Private Sub Command2_Click()
Dim nowTime As Variant
Dim nowHour As Variant
Dim rs3_RValues(1) As Double
cn.Open "gen"
mySql = "select * from Cl_svoc_45yars where date >= #1/1/2002#  and date <= #12/31/2002# " & "order by date, stationofcrash"
rs.Open mySql, cn, adOpenKeyset, adLockOptimistic
'MsgBox (rs.RecordCount)
rs.MoveFirst
While rsCount < rs.RecordCount
rs_Station = rs.Fields(5)
rs_time = rs.Fields(10)
rs_Minute = Minute(rs.Fields(10))
rs_PrevTime = Format(Hour(rs_time) & ":00:00", "hh:mm")
If Hour(rs_time) <> 23 Then
rs_NextTime = Format(Hour(rs_time) + 1 & ":00:00", "hh:mm")
Else
rs_NextTime = Format("00" & ":00:00", "hh:mm")
End If
nowDate = CDate(Format(rs.Fields(9), "mm/dd/yyyy"))
If (Hour(rs_time) <> 23) Then
nextDate = nowDate
nextDate = CDate(Format(nextDate, "mm/dd/yyyy"))
Else
nextDate = nowDate + 1
nextDate = CDate(Format(nextDate, "mm/dd/yyyy"))
End If
'MsgBox (rs_Date & " " & rs_PrevTime & " " & rs_NextTime)
prevTime = Format(nowDate & " " & rs_PrevTime, "mm/dd/yyyy hh:mm")
nextTime = Format(nextDate & " " & rs_NextTime, "mm/dd/yyyy hh:mm")
mySql2 = "select * from nearest where station = " & rs_Station
```

```
'MsgBox (mySql2)
rs2.Open mySql2, cn, adOpenKeyset, adLockOptimistic
'MsgBox (rs2.RecordCount)
i = 1
While (rs2.Fields(i + 1) <> 1)
i = i + 1
Wend
rs2.Close
nowHour = rs_Date
'If nowHour = 23 Then
'nowDate = nowDate + 1
'End If
'mySql3 = "select * FROM Orlexec1999 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM Orlexec2000 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM Orlexec2001 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM Orlexec2002 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM OrlSan1999 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM Orlsan2000 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM Orlsan2001 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM Orlsan2002 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM OrlIntl1999 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM Orlintl2000 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'mySql3 = "select * FROM Orlintl2001 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
mySql3 = "select * FROM Orlintl2002 where date >= #" & prevTime & "# and date <=
#" & nextTime & "# order by date "
'MsgBox (mysql3)
rs3.Open mySql3, cn, adOpenKeyset, adLockOptimistic
If rs3.RecordCount > 0 Then
rs3.MoveFirst
flag = 1
'If rs3.Date >= #4/25/1999# And rs3.Date < #4/26/1999# Then
'InputBox ("What value do you weant?")
'End If
While flag <= rs3.RecordCount
```

```
rs3_time = rs3.Fields(2)
rs3_RValues(flag - 1) = rs3.Fields(3)
flag = flag + 1
rs3.MoveNext
Wend
rs3_RainValue = rs3_RValues(1)
'If (rs3_RValues(2) > 0) Then
'MsgBox (rs3_RValues(2))
'rs3_RainValue = rs3_RValues(2)
'Else
'rs3_RainValue = 0 'rs3_RValues(1)
'End If
Else
rs3_RainValue = 10000
End If
'MsgBox (rs3_RainValue)
rs3.Close
rs.Fields(10 + i) = CDbl(rs3_RValues(1))
rs.Update
flag = 1
rsCount = rsCount + 1
rs.MoveNext
Wend
MsgBox ("Yahooo")
rs.Close
cn.Close
End Sub
```

LIST OF REFERENCES

Abdel-Aty M., Pande A., Hsia L. and Abdalla F. (2004) The Potential for Real-Time Traffic Crash Prediction, *ITE Journal* (forthcoming).

Abdel-Aty, M., and Radwan, A. (2000) Modeling Traffic Accident Occurrence and Involvement. Accident Analysis and Prevention, Vol.32, 633-642.

Al-Deek, H. M., and Chilakamarri, V.S.R.C. (2004) New algorithms for filtering and imputation of real-time and archived dual-loop data in data warehouse. Transportation research record 1867, pp. 116-126.

aML Version 2: User's guide and reference manual, Econware. Los Angeles, California, 2003.

Andreescu, M., Frost, D.B., 1998. Weather and traffic accidents in Montreal, Canada. Climate Research 9, 225-230.

Andrey, J., Olley, R, 1990. Relationships between weather and road safety, past and future directions. Climatological Bulletin 24(3), 123-137.

Bertness, J., 1980. Rain-related impact on selected transportation activities and utility services in the Chicago area. Journal of Applied Meteorology 19, 545–556.

Brock, J., (2002) High Performance Asphalt. The New Generation of Pavement. http://www.asphaltalliance.com/upload/High-Performance_Asphalt.pdf., Accessed 31st May, 2004.

Brodsky, H., Hakkert, A.S., 1988. Risk of a road accident in rainy weather. Crash Analysis and Prevention 20(2), 161-176.

Crashes cost drivers and society billions. http://www.drivers.com/article/547/., Accessed 31st November 2004.

Fridstrom L., Ifver J., Ingebrigtsen S., Kulmala R., Thomsen L., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Crash Analysis & Prevention 27(1), 1-20.

Garber, N. and Wu, L., (2001) Stochastic Models Relating Crash Probabilities with Geometric and Corresponding Traffic Characteristics Data. Center for Transportation Studies at the University of Virginia, Research Report No.: UVACTS-5-15-74.

Garber, N., and Ehrhart, A., (2000) Effect of Speed, Flow and Geometric characteristics on Crash Frequency for Two lane Highways. Transportation Research Record 1717.

Garber, N., and Joshua, S., (1990) Traffic and Geometric characteristics affecting the involvement of large trucks in accidents. VDOT Project No.: 9242-062-940, Virginia Transportation Research Council, University Station, Charlottesville, Virginia.

Goodwin, L., 2002. Analysis of Weather Related Crashes on U.S. Highways. Mitretek Systems, Inc. U.S.A.

Goodwin, L., 2003. Weather Related Crashes on U.S. Highways in 2001. Mitretek Systems, Inc. U.S.A.

Greene, W. Econometric Analysis. Prentice-Hall, Inc. Upper Saddle River, New Jersey. 1997.

Judge, George G., Griffiths, William E., Carter, Hill R., Lütkepohl, H., and Lee, T. The Theory and Practice of Econometrics, 2nd Edition, Wiley Series in Probability and Statistics, 1988.

Kockleman, K., and Kweon, Y., (2004) Spatially Disaggregate Models of Crash and Injury Counts: The Effects of Speed Limits and Design. Presented at the 83rd Annual meeting of Transportation Research Board, Washington D.C.

Lee, C., Saccomanno, F., and Hellinga, B., (2002) Analysis of Crash Precursors on Instrumented Freeways. Presented at the 81st Annual meeting of Transportation Research Board, Washington D.C.

Metcalf, A., Aljanahi, A., and Rhodes, A., (1999) Speed, Speed Limits and Road Traffic Accidents under Free-flow conditions. Accident Analysis and Prevention, Vol.31, 161-168, 1999.

National Traffic Safety Board (NTSB), 1980. Fatal Highway Accidents on Wet Pavement – The Magnitude Location and Characteristics, HTSB-HSS-80-1. NTIS, Springfield, VA.

Oh, C., Oh, J., Ritchie, S., and Chang, M., (2001) Real-Time Estimation of Freeway Accident Likelihood. Presented at the 80th Annual meeting of Transportation Research Board, Washington D.C.

Okamoto, H., and Koshi, M., (1989) A Method to cope with the Random errors of observed Accident Rates in Regression Analysis. Accident Analysis and Prevention. Vol.21, 371-332.

Pasupathy, R., Ivan, J., and Ossenbruggen, P., (2000) Single and Multi-Vehicle Prediction Models for Two-Lane Roadways. Final Report, Project UCNR9-8, United States Department of Transportation.

Polanis, Stanley F. (1995) "Some Thoughts About Traffic Accidents, Traffic Safety and the Safety Management System." ITE Journal, Vol. 65, No. 10, pp. 32-34, October.

Rencher, Alvin C., 2002. Methods of Multivariate Analysis. Wiley Series in Probability and Mathematical Statistics. New York John Wiley & Sons, Inc. U.S.A.

Shankar, V., Mannering, F., and Barfield, W., (1995) Effect of Roadway Geometrics and Environmental factors on Rural Freeway Accident Frequencies. Accident Analysis and Prevention, Volume 27, Issue 3, pp. 371-389.

Souleyrette, R., Kamyab, A., Hans, Z., Knapp K., Khattak, A., Basavraju, R., and Storm, B., (2001) Systematic Identification of High Crash Locations. Final Report, Center for Transportation Research and Education, Iowa State University, Iowa.

The World Almanac and Book of Facts: 1996. New Jersey: World Almanac Books.

Wright, Paul H. and Radnor J. Paquette (1987) Highway Engineering 5th ed. New York:Wiley and Sons.

Washington, S., Guevara, F., and Oh, J., (2004) Forecasting Crashes at the Planning Level: A Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona. Presented at the 83rd Annual meeting of Transportation Research Board, Washington D.C.

Washington, S., Karlaftis, M., and Mannering, F. (2003) Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC. Boca Raton, Florida.