2018

Dynamic Hotspot Identification for Limited Access Facilities using Temporal Traffic Data

Samer Al Amili
University of Central Florida

DYNAMIC HOTSPOT IDENTIFICATION FOR LIMITED ACCESS FACILITIES USING TEMPORAL TRAFFIC DATA

by
SAMER AL AMILI
B.Sc Al-Mustansiriya University, Iraq, 2002
M.S. Al-Mustansiriya University, Iraq, 2007
M.S. University of Central Florida, 2013

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

Spring Term
2018

Major Professor: Mohamed Abdel-Aty

ABSTRACT

ABSTRACT

Crash frequency analysis is the most critical tool to investigate traffic safety problems. Therefore, an accurate crash analysis must be conducted. Since traffic continually fluctuates over time and this effects potential of crash occurrence, shorter time periods and less aggregated traffic factors (shorter intervals than AADT) need to be used. In this dissertation, several methodologies have been conducted to elevate the accuracy of crash prediction.

The performance of using less aggregated traffic data in modeling crash frequency was explored for weekdays and weekends. Four-time periods for weekdays and two time periods for weekends, with four intervals (5, 15, 30, and 60 minutes). The comparison between AADT based models and short-term period models showed that short-term period models perform better.

As a shorter traffic interval than AADT considered, two difficulties began. Firstly, the number of zero observations increased. Secondly, the repetition of the same roadway characteristics arose. To reduce the number of zero observations, only segments with one or more crashes were used in the modeling process. To eliminate the effect of the repetition in the data, random effect was applied. The results recommend adopting segments with only one or more crashes, as they give a more valid prediction and less error.

Zero-inflated negative binomial (ZINB) and hurdle negative binomial (HNB) models were examined in addition to the negative binomial for both weekdays and weekends. Different implementations of random effects were applied. Using the random effect either on the count part, on the zero part, or a pair of uncorrelated (or correlated) random effects for both parts of the model. Additionally, the adaptive Gaussian Quadrature, with five quadrature points, was used to increase

accuracy. The results reveal that the model which considered the random effect in both parts performed better than other models, and ZINB performed better than HNB.

ACKNOWLEDGMENTS

I would like to present my most profound gratitude and thanks to my advisor, Prof. Dr. Mohamed Abdel-Aty, for his invaluable support, guidance, patience, advice and support and encouragement toward successful completion of his doctoral course. The author wishes to acknowledge and thank his committee members, Dr. Essam Radwan, Dr. Morgan Wang, Dr. Naveen Eluru and Dr. Jaeyoung Lee.

Also, I would like to thank all of my friends and colleagues for there concern and encouragement.

Last but not least, many thanks go to my families. Especially my father who always keeps me in his prayers and my brother too.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

AADT -- Average Annual Daily Traffic

AASHTO -- American Association of State Highway and Transportation Officials

AVI -- Automatic Vehicle Identification

CF -- Crash Frequency

CFX -- Central Florida Expressway Authority

CR -- Crash Rate

CRP -- Continuous Risk Profile

DMS -- Dynamic Message Signs

DUI -- Drive Under Influence

EB-- Empirical Bayes

EPDO -- Equivalent Property Damage Only

F+I-- Fatal and Injury crashes

FDOT -- Florida Department of Transportation

FHWA -- Federal Highway Administration

HCM -- Highway Capacity Manual

HOT -- High-Occupancy toll lanes.

HSM -- Highway Safety Manual

INAR -- Integer-Valued Autoregressive.

ITS -- Intelligent Transportation System

Log AADT -- Natural Logarithm of AADT

LogVol -- Natural logarithm of Volume

MAD -- Mean Absolute Deviation

MDACP -- Multivariate Autoregressive Conditional Double Poisson model.

ME -- Measurement Error

MP -- Milepost

MSPE -- Mean Square Prediction Error

MVDS -- Microwave Vehicle Detector Sensor

NB – Negative Binomial

P -- Proportion Method

PDO -- Property Damage Only

PDO --Property Damage Only

PFI -- Potential for Improvement

PS -- Peak Searching PS

PSI -- Potential for Safety Improvement

RCI -- Road Characteristics Inventory

RITIS -- Regional Integration Transportation Information System

S4A -- Signal for Analytic

SARIMA -- Seasonal autoregressive integrated moving average.

SMW -- Sliding Moving Window

SPFs -- Safety Performance Functions

UCF -- University of Central Florida

V/C -- Volume-to-Capacity Ratio

VSL -- Variable Speed Limit

CHAPTER 1: INTRODUCTION

1.1 Overview

Traffic safety including hotspot identification is among the most critical issues of the transportation system. Many transportation engineering and government officials' efforts were considered to reduce the number of fatality crashes or its severity in the United States. However, according to the (NHTSA., 2017) data, 37,461 people died in traffic fatality crashes in 2016 with an increase of 5.6 percent from the calendar year 2015. Thus, it is essential to dedicate more effort to reduce this number by accurately identifying the hotspots on the roadway and utilize the safety resources wisely. Crash frequency analyses are the most important tools to predict the number of traffic crashes and quantify the significant contributing factors that cause these crashes.

In 2010, the first edition of the Highway Safety Manual (HSM) had been released by the American Association of State Highway and Transportation Officials (AASHTO). A six-step safety management process has been introduced by the HSM to provide safety engineers with a scientific and systematic approach to managing road safety. Network screening is the first and the essential step in this process to identify the hotspot locations that need to be treated to improve the roadway safety. The output obtained from the network screening is a list of high concentration locations of crashes. By ranking these locations based on their potential for safety improvement, the road authorities will be able to allocate their limited resources to the most critical locations. The network screening methods that are suggested by the HSM and other research are mainly based on the Safety Performance Functions (SPFs) which is based on the Average Annual Daily Traffic (AADT) to estimate the predicted number of crashes. Though these methods have been proven to be effective, using high aggregated traffic data may cause five problems. First, the AADT along the segments cannot represent the traffic condition at the time of the crash. Two expressways or freeways one with high traffic volume during the peak hours would have a different effect on crash occurrence than another one with the same AADT but even hourly traffic distribution. Second, AADT does not account for the variation in traffic volume in each direction. Third, it is impossible to know the impact of the temporal factors by using AADT, e.g., morning peak, off-peak, evening peak, and nighttime (Mensah & Hauer, 1998; B. Persaud & Dzbik, 1993). Fourth, there is a different traffic pattern on weekdays and weekends that AADT does not account for (Yu & Abdel-Aty, 2013a). AADT is not based on the whole year data collection. Based on the SPFs with the AADT we cannot know when the riskiest time during the day and in which direction it is and what the practical, useful, and official ways are to treat these locations and utilize the safety resources efficiently.

In the past few decades, traffic detection technology is the main data source of any Intelligent Transportation System (ITS); on freeways and expressways, there is a wide range of vehicle detection devices in use than ever before. Start from the loop detectors to video and radar-based detectors. Traditionally, traffic engineering management has heavily used the traffic data that is generated from these detection devices. However, these data can be easily used to support and add further improvement to one of the most important and basic concerns of the transportation system which is the traffic safety. The efficiency of the whole transportation system could be improved and enhanced by improving the traffic safety. Researchers in traffic safety have been focusing on the contributing factors leading to crashes in hope to better understand the crash mechanisms that would aid professionals to come up with better traffic system design. To uncover the crash patterns, more detailed data is required especially regarding traffic patterns during a different time of the day per direction not for both directions together. These days with the broad application of ITS technologies, more detailed data are available.

Traffic safety researchers are particularly interested in freeways and expressways, as they provide a high degree of mobility and connection between different parts of the metropolitan area and are necessary with the development and expansion of many areas. Freeways and expressways carry out a significant amount of traffic traveling at high speed between different areas. These types of roads are considered as the spine of modern society. Safety on these types of roads is so important to keep the mobility and prevent farther delay and congestion.

The Central Florida Expressway Authority (CFX) equipped the expressways with Microwave Vehicle Detection Sensors (MVDS) and other detectors that provide a very detailed data of these expressways. Regional Integration Transportation Information System (RITIS) has many detectors that are mainly radar detectors that provide traffic data for other roadway types. Sufficient Utilization of these resources of data is expected to lead to more accurate hotspot identification and better utilization of the resources to improve the safety on these roads.

Therefore, the objective of this study is to utilize the available less aggregated traffic data to investigate traffic safety problems by different time intervals for weekdays and weekend with consideration of both directions and to enhance hotspot identification. Also, offering more clear insight to the traffic engineers to select better solutions by different intervals and better utilization of the resources.

1.2 Research Objectives

The current work of this study focuses on investigating the viability of using less aggregated traffic data to get more accuracy modeling estimation and to find the crash contributing factors for the expressway (SR 408) based on currently available traffic detection data, then identifying hotspot locations on this road. The detailed objectives will be achieved by the following tasks;

Does the less aggregated traffic data that is provided by ITS detectors give a better prediction and better hotspot identification? So, a preliminary analysis was conducted by using less aggregated traffic data;

Addressing the zero-inflated and repeated measurements in the data and develop several full SPFs for different intervals with different time periods for each direction for weekdays and weekends and suggested appropriate way to overcome these issues to improve the prediction and hotspot identification, and;

Explore the potential of using dual-state models with different proposed random effect implementation and recommend the best model to use. Also, identify the significant factors affecting crash occurrence for different time intervals.

Each of the above objectives has been achieved by the following tasks. Objective 1 has been achieved in Chapter 3:

Filtering the data, dividing it into weekdays and weekends then divided each of them into different intervals (5 minutes, 15 minutes, 30 minutes, and 60 minutes) for different time periods (morning peak, off-peak, evening peak, and night time) for weekdays and high volume and low volume periods for weekends.

Developing SPFs for a 21-mile expressway segment on SR 408 that has the largest traffic and crashes on the CFX expressway system for different time periods and intervals. Traffic data from MVDS (only the volume in this stage has been used) with crash data (total crashes) and geometric data (speed limit, the existence of curvature, the existence of auxiliary lane, and the number of lanes).

Adopt appropriate hotspot identification method to compare the less aggregated suggested data with the traditional method and suggest the best time interval.

Objective 2 has been achieved by the following tasks, and it is presented in Chapter 4:

Prepare different datasets:

- A dataset that has the data of the segments that at least one crash occurred on them during the study period.
- A dataset that has the data of the segments that at least two crashes occurred on them during the study period.
- A dataset that has the data of the segments that at least three crashes occurred on them during the study period in addition to the whole data.

Develop multiple SPFs for all the prepared data.

Comparing the modeling results and recommend the best dataset.

The following tasks have been implemented to achieve the last objective:

Developing several dual-state models with the consideration of the implementation of random effect in several suggested ways in the models.

Adopting appropriate goodness of fit measures to compare the models' performance and suggest the best one.

1.3 Dissertation Organization

The structure of the dissertation is as follows: Chapter 2 provides a literature review on factors contributing to crash occurrences, hourly traffic volume studies, real-time crash studies, then network screening studies. Chapter 3 gives a brief description about the expressway of interested, data collection, data preparation, methodology and finding of the PSI conducted on SR 408. Chapter 4 develops several full SPFs for the different proposed data set. Then a comparison between the developed models was presented. In chapter 4, two zero-inflated models were developed and compared, then the best model was suggested. Finally, Chapter 5, summarizes the overall dissertation work.

CHAPTER 2: LITERATURE REVIEW

2.1 Crash Frequency Studies

Road safety studies have been a continuously researched topic in the past few decades. Researchers have extensively conducted research to gain a better understanding of crash occurrence mechanisms at both macro and micro levels. Micro-level traffic safety studies examine the occurrence of vehicle crashes on specific locations (e.g., roadway segments and intersections). Crash frequency distribution has commonly been presented using Poisson (Joshua & Garber, 1990; Miaou, 1994) and Negative Binomial NA (Hauer & Hakkert, 1988; Miaou & Lord, 2003; Poch & Mannering, 1996). Lord and Mannering (2010) summarized the advantages and disadvantages of the existing frequency models. In this proposal, the main statistical technique adopted is the Negative Binomial model.

El-Basyouny and Sayed (2013) investigated the relationship between collisions and conflicts. In his study, he used the lognormal model for predicting the conflicts using traffic volume, some geometric-related variables, and area type as the covariate. He then predicted collusion using conflicts-based negative binomial (NB) safety performance function. They found from the scaled deviance and Pearson X2 goodness of fit that the proposed NB has adequately fitted the data. El-Basyouny and Sayed (2010) used measurement error (ME) model conjected with Negative Binomial SPF to overcome the bias in predicting the number of crashes; then compared the results with the traditional NB technique. They found that both approaches give comparable results when the variance in volume between years is small.

2.2 Factors Contributing to Crash Occurrence

Various factors such as driver characteristics, roadway geometric, traffic condition, and weather can affect on roadway safety (Huang & Abdel-Aty, 2010; Yu & Abdel-Aty, 2013a). Identification of these factors helps traffic engineering to select the best countermeasures. The effects of traffic variables (speed, volume, and occupancy) on crash occurrence have been incorporated in traffic safety studies. One of the most widely used traffic data sources are the loop detectors (Golob et al., 2008; Lee & Abdel-Aty, 2008; Lee et al., 2006; Pande & Abdel-Aty, 2006; Xu et al., 2013). Loop detectors can provide traffic data (speed, volume, and occupancy) based on 60 or 30 seconds. However, loop detectors perform weakly in adverse weather condition, such as heavy rain, during which the accuracy of data could be greatly reduced. Additionally, the maintenance of the loop detectors could be very complicated. Recently developed traffic detection systems enable newly developed nonintrusive detection devices such as: Video Detectors (Hourdos et al., 2006; Laureshyn et al., 2010; Versavel, 1999; G. Zhang et al., 2007), Automatic Vehicle Indemnification (AVI) sensors (Abdel-Aty et al., 2012; Ahmed et al., 2012), Microwave Vehicle Detection System (MVDS) (Ahmed & Abdel-Aty, 2013; Akin et al., 2011; Yu et al., 2013), and Regional Integrated Transportation Information System (RITIS) (Pack et al., 2008; Smith & Venkatanarayana, 2007) to be used for traffic data collection. MVDS traffic data provides similar traffic information as the loop detectors. MVDS recognizes the length of passing vehicles and classifies them into four categories: less than 10 ft, between 10 ft and 24 ft, between 24 and 54 ft, and greater than 54 ft. The advantage of MVDS data compared to loop detectors is that MVDS is installed along the side of the road. Therefore, its maintenance is easier, and it's not greatly affected by adverse weather conditions. MVDS also provides time mean speed.

2.2.1 Traffic Characteristics

Numerous studies have explored the effect of traffic characteristics on crash occurrence. Most of the studies explore the effect of speed, and the variance of speed, on the crash occurrence. Several study results confirm that higher speed will lead to a higher number of crashes or higher crash rates. Taylor et al. (2002) implemented cross-sectional analysis on rural road segment in Britain. The results show a positive relationship between crash frequency and average speed. Other researchers found that increased speed leads to more severe crashes (Hauer, 2009; Kockelman & Kweon, 2002; O'Donnell & Connor, 1996; Shankar & Mannering, 1996; Xu et al., 2013). Nilsson (2004) found in his study, a positive relationship between the number of crashes and changes in speed with different magnitude depends on crash types. Aarts and Van Schagen (2006) found that high speed led to higher crash rates when they reviewed the relationship between the driving speed and the risk of road crashes. Several researchers (Ahmed et al., 2011; Baruya, 1998; Yu et al., 2013) found an opposite effect of speed on crash occurrence. In their studies, they found that the likelihood of crashes increased when average speed decreased 5 – 10 minutes before the crash occurred. Generally speaking, speed has different effects on the likelihood of crash occurrence. Other studies focused on the speed variation instead of speed itself. In an early study by Garber and Gadirau (1988) about the effect of speed variance on the crash occurrence, they found that the crash rate does not necessarily increase with an increase in average speed; but does increase with an increase in speed variance. Golob and Recker (2003) found that on urban freeway left-lane crashes are more likely caused by volume effects, while right-lane crashes are more likely induced by adjacent lane speed variance. Lave (1985) found fatality rate was strongly associated with speed variance rather than average speed.

Annual Average Daily Traffic has been widely used as a traffic flow indicator. Most traffic safety studies have used this variable as an exposure factor. Traffic safety studies found that traffic flow is positively related to crash frequencies (Chin & Quddus, 2003; Hauer et al., 2002; Wang et al., 2009; Zhang et al., 2012). Recent works begin to explore the potential of whether a surrogate measure of disaggregate volume is worth investigation and if it could be used as an alternative to aggregated volume indicators such us AADT. Yu and Abdel-Aty (2013a) used traffic volume on weekdays and weekends to compare and reveal the features of crashes between weekdays and weekends. Hossain and Muromachi (2013) used high-resolution traffic data to identify factors influencing crashes on an urban expressway. They found that congestion level and speed difference, upstream and downstream, have the highest influence on crash and crash types.

2.2.2 Roadway Geometric Characteristics

Roadway geometric design plays a major role in traffic safety. Many studies have taken geometric characteristics into account when performing traffic safety evaluation. The geometric information could be, in most cases, gathered in two ways: fixed length segment (Shankar et al., 1995) or homogenous segments (Milton & Mannering, 1998). In this case of fixed length segment, the roadway section is divided into equal segments. For homogenous segments, a new segment will be considered if any of the geometric characteristics have changed. Both segmentation methods have disadvantages. Fixed length segment could have different geometric characteristics within it. Homogenous segments could be too short for traffic safety analysis. Crash locations are reported to the nearest milepost, and with short segments, this could lead to misplacement of the crash. Recently, researchers have used the homogeneous segmentation method but combined the too short segments with the most appropriate adjacent segment, as illustrated by Ahmed et al. (2011).

Miaou et al. (1992) investigated the relationships between trucks crashes and the highway geometric design variables by developing a count model. The results showed that there were significant correlations between the number of truck crashes and the average annual daily traffic per lane, vertical grade, and horizontal curvature.

Shankar et al. (1995) found the maximum grade and the number of horizontal curves had a positive relationship with accident frequency in Seattle.

Milton and Mannering (1998) found that sharp horizontal curves and narrow lanes (less than 3.5m) decrease crash frequency in Eastern Washington.

Abdel-Aty and Radwan (2000) modeled traffic crash occurrence and involvement on SR 50 in Central Florida. They found that narrow lane width, narrow shoulder width, reduced median width, and a larger number of lanes increased the likelihood of crash involvement. Also, it was found that crash occurrence was positively related to horizontal curvature.

Noland (2003a, 2003b) utilized county-level highway crash data from 1987 to 1990 in the state of Illinois and found that fatality crashes increased when the number of lanes and lane width increases. However, increasing the outside shoulder width decreased the crashes.

(Haynes et al., 2007; Haynes et al., 2008) suggested that more curved roads in an area resulted in fewer road crashes; furthermore, they showed that road curvature has an inverse effect on fatal crashes. Curvature was found to be a protective factor.

In Kononov et al. (2008) study about urban freeways. They found that an increase in the number of lanes increased the number of crashes.

Park and Lord (2009) studied the effect of freeway design on safety. They found that the number of lanes and curvatures were positively associated with crash frequencies; while the median width has a negative relationship with the crash occurrence.

Ahmed et al. (2011) and Yu et al. (2013) investigated the hazardous factors on a mountainous freeway segment in Colorado. They found that vertical curves have a great effect on crash occurrence. Also, they found that crash likelihood is negatively affected by the wider median width, increased number of lanes, and a higher degree of curvatures.

For several roadway geometric characteristics, consistent conclusions of their effect on traffic safety were reached; such as the number of lanes, the median width, lane width, and shoulder width. In contrast, the effect of the curvature is not consistent.

2.3 Hourly Volume Traffic Safety Studies

Precious findings on crash prediction, based on hourly volume, are presented in this subsection. Gwynn (1967) collected 5-years data from a 3.8-mile US Route, where there were no traffic signals and grade crossings. The study tried to find the relationship between hourly volume and accident rates. The result showed that highest crash rates occurred in the low and high volume level, and lowest crash rate occurred in the mediate volume level. These results would not be achievable if AADT data were used. The whole section of road had the same traffic exposure (i.e., ADT) since it did not have any access.

Persaud and Mucsi (1995) used the hourly traffic volumes for estimating crash number on two-lane rural roads. They calibrate their model based on different time periods (all day, daytime, and nighttime) and geometric characteristics (e.g., shoulder width, lane width) for total crashes and F+I. Their result proved that effect of day/night was different for single- and multi-vehicle crash.

Zhou and Sisiopiku (1997) investigate the relationship between the hourly volume-to-capacity ratio (V/C) and crash rates on an urban freeway. When considering the day of the week (weekday and weekend), different crash type (turnover, rear-end, and fixed object), different severity level (PDO, F+I); they found that the relationship between total crash rates and hourly V/C ratio followed a general U-shaped pattern. That is to say, the crash rates were high when V/C ratio was either high or low, but crash rates were low when the V/C was in between.

Chang et al. (2000) used five years of freeway data to examine the relationship between crash rate and hourly V/C ratio. In their research, three different freeway sections were studied: basic freeway section, tunnel section, and toll gate section. They found that the relationship between crash rates and V/C ratio had a U-shape relationship; however, the U shape of these three sections was not the same. For example, the toll gate section U-shape was above the other two (when the V/C was the same, crash rate of toll gate section was higher). The author's recommended to include more geometric and other traffic factors in the model since the R2 values (0.4209 to 0.5161) were low.

Martin (2002) studied the relationship between the crash rate and hourly traffic volume with the consideration of time of day (night or day), the day of the week (weekday or weekend) and the number of lanes. He found that Property Damage Only (PDO) and injury crash rates were high when the hourly traffic volume was low. During the night, there were more severe crashes when the hourly traffic volume was low.

Lord et al. (2005) investigated the effect of hourly traffic parameters, e.g., volume, density, and V/C ratio on the crashes upon rural and urban freeway segments. They found that the higher density and V/C ratio caused increased crash count. The authors recommended that a separate function formula for single- and multi-vehicle crashes should be developed. The authors also suggested that crash rate should not be used in the crash prediction model since nothing proved that crash rates follow a normal distribution.

Kononov et al. (2012) investigate the relationship between hourly crash rate and hourly speed along with density for urban freeways. The result showed that crash rate increased when the density was high, and the speed did not decrease. Additionally, the crash rate stayed stable when the speed was high, and the density was between low and moderate.

2.4 Real-time Crash Prediction

There have been numerous studies on real-time crash prediction models to find the likelihood of crashes with less aggregated traffic data (Madanat & Liu, 1995). With the development of traffic detection technique, short time traffic data generated on 30 or 60 seconds enables researchers to look at a crash with microscopic traffic data more easily. Common predictors in the real-time crash predictions include average speed, the standard deviation of speed, and coefficient of variation of speed, traffic volume, and occupancy aggregated at upstream and downstream detector locations. Recent research has seen real-time information incorporated and their effects were found significant, especially in areas where inclement weather conditions are common.

Abdel-Aty and Pande (2007) identified and classified crash propensity factors using real-time traffic and crash data for the I-4 corridor in Orlando. The authors found that 70% of the crashes could be identified based on the 10 to 15 minutes speed variation before the crash occurs. A real-time crash risk assessment model for lane-change and rear-end crashes have been developed by (Abdel-Aty et al., 2007). The authors used loop detector data collected from I-4 for four years to assess the crash risk on a real-time basis. Lee et al. (2006) also used 4-years of crash data from I-4 freeway in Orlando and developed logistic regression models to identify real-time indicators for rear-end and sideswipe crashes. The authors found that the variation in traffic flow and peak and off-peak periods are correlated and have an important effect on sideswipe crashes.

Numerous studies have addressed the modeling of crash risk prediction for freeways and linked the crash risk with several real-time traffic flow characteristics. Lee et al. (2003) used real-time traffic data and log-linear models to estimate crash risk. The results showed that the crash risk is significantly correlated with variation in speed, the difference in speed between upstream and downstream, and the traffic density. (Abdel-Aty & Pande, 2005; Abdel-Aty et al., 2004) utilized the matched case-control logistic regression method to predict crashes in real-time. The matched case-control analysis was employed to explore the effects of traffic flow variables while controlling the effects of other confounding variables through the design of the study. The authors noticed that multi-vehicle crashes on freeways under high- and low-speed traffic conditions differed in severity and mechanisms. Two separate models were evaluated in the matched-case control framework. They concluded that the low-speed crashes occurred mostly in persisting congestions; however, the queues resulted from the crash dissipated quickly. In contrast, high-speed crashes often occurred under smooth traffic conditions; therefore, disruptive traffic conditions originating from downstream could cause driving errors.

Zheng et al. (2010) studied the impact of traffic oscillations, which was also known as stop- and-go driving, on freeway crashes in real-time. The matched case-control model showed that speed deviation was a significant variable, and it had a positive impact on crash occurrence.

Pande and Abdel-Aty (2006) used a classification tree and a neural network to develop a crash risk prediction model to identify real-time traffic conditions that are prone to lane-change crashes. The results showed that average speeds upstream and downstream, the difference in occupancy on adjacent lanes, and standard deviation of volume, and speed downstream of the crash location were significantly correlated with the lane-change crash risk.

Recently, Xu et al. (2012) conducted a k-means clustering analysis to test the connection between the traffic flow states and the crash risk on freeways. Crash risk prediction models were developed for these states. The results revealed that traffic flow characteristics have different impacts on the crash risk for different traffic states. Ahmed and Abdel-Aty (2012) developed a matched case-control logistic regression model for real-time crash prediction. The results showed that the average speed and the standard deviation of speed are statistically related to the crash likelihood. Yu and Abdel-Aty (2013a) conducted a multi-level traffic analysis to reveal the different characteristics of weekday and weekend crashes based on real-time traffic data. The results exhibited that the weekday crashes occurred mostly during the peak period while the weekend crashes occurred during free flow conditions. Ahmed et al. (2012) built a Bayesian logistic regression model and combined the space mean speed collected in real-time data from an Automatic Vehicle Identification (AVI) system, and real-time weather and geometric data to investigate the effect of these variables on the occurrence of crashes on a mountainous freeway. The results indicated that roadway geometrics, real-time weather, and AVI data have a considerable effect on the crash occurrence. Yu et al. (2013) employed Bayesian random effect models to incorporate real-time weather, traffic data, and geometric variables in their crash frequency study. The results showed that crash mechanisms between single and multi-vehicle crashes were different and based on different seasons, so different active traffic management strategies should be applied. Xu et al. (2013) developed a sequential logit model to link the likelihood of crash occurrence at different levels to the various traffic flow characteristics collected from loop detectors on I-880 in California. The results showed that the contribution of traffic flow characteristics on the likelihood of crashes was quite different for different severity levels. Hassan and Abdel-Aty (2013) used loop detectors and radar sensor data from freeways to investigate whether real-time traffic data can be used to predict the crash occurrence during reduced visibility conditions. The authors also wanted to compare the significant variables that contributed to crashes in reduced visibility conditions versus those in clear visibility. The results revealed that the contributing factors for crashes during reduced visibility are slightly different from those for crashes during clear visibility.

Hourdos et al. (2006) employed video data to identify crash-prone conditions at freeway high-crash locations in Minnesota. The logistic regression model was used to find indicators of crash- prone conditions. Average speed, speed variation, wet pavement, and reduced visibility were found to be significant factors increasing the crash likelihood. The model achieved a 58% accurate detection rate. One weakness of logistic regression is it has no control over other potential factors besides the ones being tested.

Bayesian matched-case logistic regressions have been employed in the study of visibility related crashes by Abdel-Aty et al. (2012). The advantages of using the Bayesian approaches, as the authors pointed out, include (1) it provides a natural and principled way of combining prior information with the data to yield a posterior belief, (2) it presents a full distributional profile of parameters rather than single coefficient estimates to fully account for the uncertainty associated with single parameter estimates in classical statistics and (3) small sample inference proceeds in the same manner of a large sample. Both loop detector data and AVI data were used in this study, it was found that the model estimated from loop detector data indicated the average speed observed at the nearest downstream station along with the coefficient of variation in speed observed at the nearest upstream station at 5-10 minutes prior to the crash time, were significant to visibility related crashes. The AVI data suggested only the coefficient of variation in speed was significant.

Yu and Abdel-Aty (2013b) employed Bayesian multi-level logistic regression to study the single- and multi-vehicle crash mechanisms on a mountainous freeway segment. The modeling approach enabled the authors to account for the seasonal variations, crash-unit-level diversity, and segment-level random effects on crash occurrence. The author stated that by Bayesian inference techniques, more unobserved heterogeneity could be captured and the better classification ability the model would have.

Random forests method is an ensemble classifier that consists of many decision trees. Compared with traditional classification trees, the random forest could obtain unbiased error estimates with no need for a separate cross-validation test data set. Ahmed and Abdel-Aty (2012) implemented the random forest technique to identify the significant traffic factors affecting crash occurrence using AYI data on OOCEA's system. The authors concluded that AYI data were promising in providing a measure of crash risk in real time. However, they suggested it is useful when AVI segments are within 1.5 miles on average.

2.5 Statistical Models

Many statistical models have been developed and used in crash frequency analyses. A summary of the main statistical models that were used in the crash frequency analysis was presented by Lod and Mannering (2010) and Mannering and Bhat (2014). Crash counts are non-negative integer events. The traditional starting modeling for crash frequency analysis is the Poisson model (Jovanis and Chang, 1986; Joshua and Garber, 1990; Sheather and Jones, 1991; Miaou and Lum, 1993). The Poisson model assumes that there is no difference between the mean and variance; for this reason, the Poisson model cannot be used with the data that is over-dispersed (i.e., the variance is more than the mean).

An alternative to the Poisson model is the negative binomial (NB) model which is an extension of the Poisson model and can deal with the over-dispersion problem. The NB model adds an error term ei to relax the equal mean variance assumption of the Poisson model. Numerous crash frequency analyses have been conducted using the NB model (Maycock and Hall, 1984; Persaud, 1994; Kumala, 1995; Karlaftis and Tarko, 1998; Abdel-Aty and Radwan, 2000; Carson and Mannering, 2001; Miaou and Lord, 2003; Alaluusua et al., 2004; Ladron de Guevara et al., 2004; Lord et al., 2005b; Kim et al., 2006a; Wang et al., 2006; Graham et al., 2010; Abdel-Aty et al., 2011a). The NB model can handle the over-dispersion resulting from temporal dependency and unobserved heterogeneity, but it may not properly account the over-dispersion caused by excess zero counts (Rose et al., 2006).

Analyzing zero-inflated count variables is one of the challenging methodologies. Poisson and NB models are insufficient to account for the zero inflated count data. The underlying assumption of the presence of zero counts may come from two states: inherently safe state (where there are no crashes) and non-zero state (where there is a crash). The non-safe state can be either a true zero (safe location) or sampling zeros (where excess zeros are results of underreporting crash data) (Shankar et al., 1997; Miaou,1994).

To account the issue of excess zero in the count data, two possible relaxations of the single-state count models were proposed. The first one – the zero inflated (ZI) model – is used to account for both true and sampling zero, and has been used by several crash frequency analysis studies (Shankar et al., 1997; Chin and Quddus, 2003). The second one – the hurdle model – which has rarely been used in transportation safety studies, is used to account for the sampling zeros.

2.6 Hotspot Identification Studies

Network screening is defined as the identification of crash hazardous locations; it is also referred to as crash hotspots, hazardous road locations, and accident-prone locations, sites with promise, black spots, high-risk locations, or priority investigation locations. Network screening is the most important first step in the highway safety management process. Network screening is a vital step to identify crash hotspots. Identifying safe spots as unsafe spots will let to waste the resources on locations that are incorrectly identified as unsafe while those that are true unsafe locations will remain untreated so the wasting will be more than twice. From a methodology perspective, different methods have been used to identify hotspots.

Monsere et al. (2007) used five ranking methods (frequency, critical rate by functional class, critical rate by functional class and climate zone, potential for crash reduction, and expected frequency adjusted by empirical Bayes) to present the empirical analysis of screening and ranking on rural one-mile sections highway in Oregon for a specific crash type (speed and ice related crashes). Safety Performance Function for Empirical Bayes using Negative Binominal model was developed. Top twenty-one-mile sections were identified for each method and compered. The results showed similar rank-order segments identified by rate-base methods, EB-based method and simple frequency method identified well-compared segments. But they could not specify which method was superior.

Montella (2010) compared between several hotspot identification methods (crash frequency (CF), equivalent property damage only (EPDO) crash frequency, crash rate (CR), proportion method (P), empirical Bayes estimate of total-crash frequency (EB), empirical Bayes estimate of severe crash frequency (EBs), and potential for improvement (PFI) using five years of crash data and four testing methods (the site consistency test, the method consistency test, the total rank differences test, and the total score test). The results showed that the EB method performed better than other hotspot identification methods.

Cheng and Washington (2005) evaluated three hotspot identification methods (simple ranking, confidence interval, and Empirical Bayes) using experimentally derived simulated data. The results showed that the Empirical Bayes significantly outperforms ranking and confidence interval techniques.

Chung and Ragland (2009) introduced a new method called (Continuous Risk Profile CRP) to detect high collision concertation locations. The new method is not affected by the spatial correlation and does not require roadway segmentation. They proved that this method has a lower false positive rate (identified safe spots as unsafe spot) than the commonly used methods (sliding window and peak searching methods).

Kwon et al. (2012) evaluated the performance of the most three network screening methods (Sliding Moving Window SMW, Peak Searching PS, and continuous risk profile CRP). Traffic crash data were used to estimate excess expected average crash frequency with Empirical Bayes using SPF. The study found that CRP performed better than other methods.

In the previous researchers used mostly the Annual Average Daily Traffic (AADT) to develop SPFs, which does not account for the variation in traffic flow by time periods and direction. In my proposal, some models have been developed for different time periods for weekdays and weekends for each direction using shorter time intervals to investigate more precisely the locations and time of the hotspots on the roadway.

2.7 Summary

Considerable crash frequency analysis studies have been conducted to find the relationship between the number of crashes and traffic and geometric characteristics of the roadways. These studies were either considering AADT in their studies which is very aggregated traffic data that does not represent the real traffic conditions when the crash happened, or a very real-time traffic data to find the most hazardous variables that may cause a crash to happen. Compared the highly aggregated traffic data studies with the safety analysis using microscopic traffic data, i.e., real-time traffic data, traffic safety studies utilizing microscopic traffic data performs better in providing a valuable detail about crash mechanisms. There are plenty of real-time crash analysis and very limited number of hourly safety studies that do not count for different time periods, intervals, different directions, and weekdays and weekends. Regarding the methods that were used in the microscopic analyses, the dominant model is the NB for the hourly crash frequency analysis, while for the real-time crash studies the dominant model is the logistic regression model.

CHAPTER 3: SYSTEM OVERVIEW AND PSI CALCULATION FOR SR 408

3.1 SR 408 Expressway Road

To explore the viability of using less aggregated traffic data from the traffic detection system in crash frequency analyses, one of the most important expressways in the East-Central Florida that is located in a very densely area has been selected. The toll roads in the East-Central Florida Area are managed by the Central Florida Expressway Authority CFX. Currently, the CFX manages partly or completely five toll roads. State Road 408, is considered the backbone expressway out of 109 miles of the expressway. Its length is 22 miles starts from Florida's Turnpike in Ocoee and ends at State Road 50. It severs an estimated 125,000 – 135,000 vehicles per day (https://www.cfxway.com/TravelersExpressways/Expressways/CurrentExpressways/408EastWestExpressway.aspx).

Figure 3.1 shows the selected Expressway (SR 408).

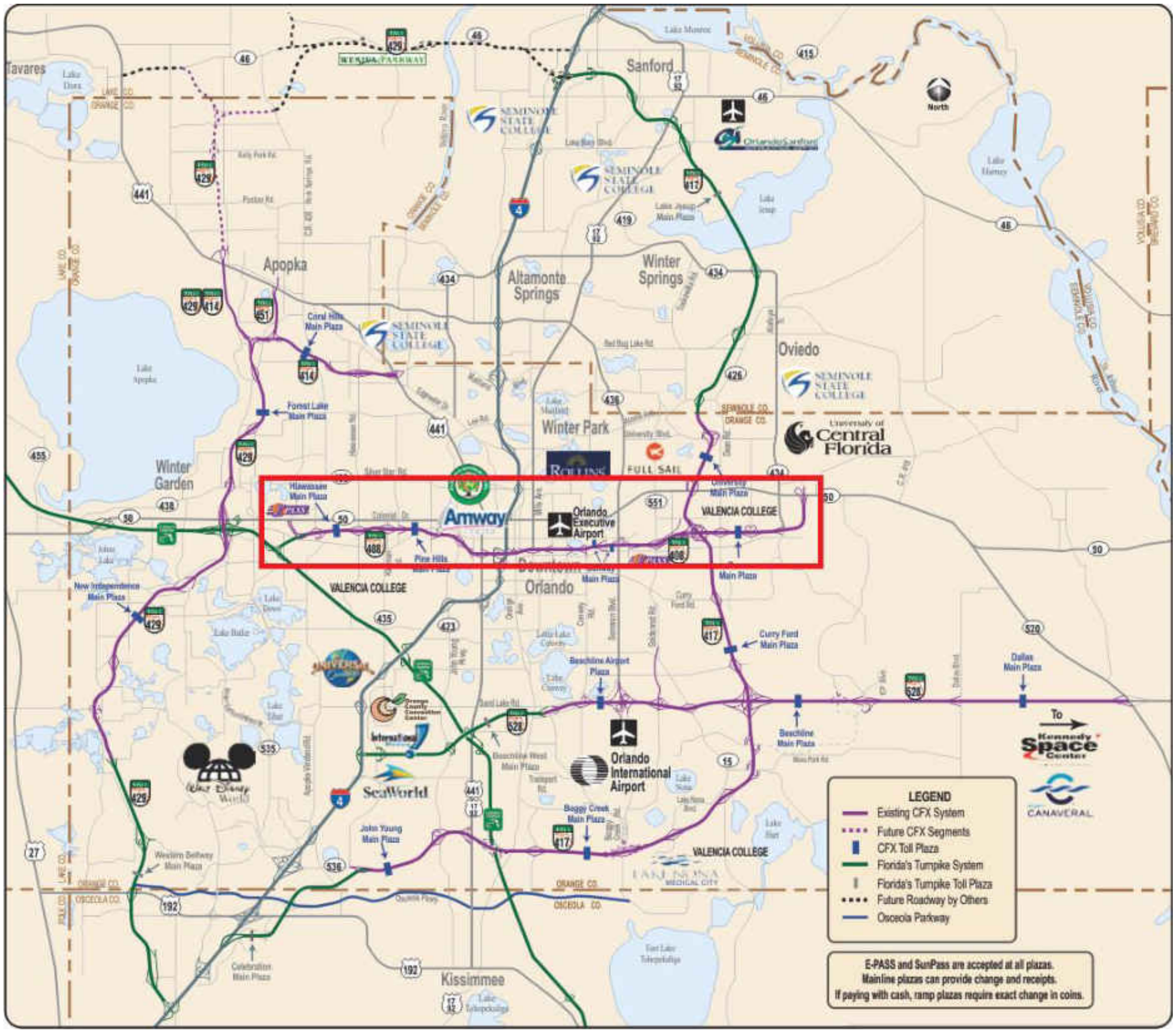


Figure 3-1: The Selected Expressways SR 408 (2014).

3.2 Traffic Data

Comprehensive time and effort have been made to collect and filter the data from the detections system. In 2012, CFX installed the Microwave Vehicle Detection System MVDS sensors on their system which is designed for traffic monitoring. The traffic data from the installed MVDS sensors have been collected for this study from July 2013 to July 2014. Figure 3.2 displays the MVDS sensors on SR 408.

Figure 3-2: MVDS Sensors on SR 408.

In total, 110 detectors are installed in both directions of the SR 408. The traffic information includes speed, volume, occupancy and vehicle type per lane at the installed locations. These data are generated based on one-minute interval. The traffic data that is provided by MVDS sensors is very rich. It contains the record of the total traffic volume and original speed not for just the mainline traffic, but also for the traffic at the ramps and toll plazas. In this study, only mainline traffic was considering. MVDS recognizes the length of passing vehicles and classifies them under four groups: the vehicles which are less than 10 ft long belong to group 1; between 10 and 24 ft to group 2; between 24 and 54 ft to group 3; and greater than 54 ft to group 4. The average spacing between MVDS is 0.385 miles. Table 3-1 shows the number of MVDS sensors on each direction with the average spacing between the sensors.

Table 3-1: MVDS sensors on each direction and the spacing between them.

Route	Length (mi)	Direction	MVDS Detectors					
			Total	Mainline (including TP Express)	The average spacing between adjacent detectors			
					Mean	Std Dev	Min	Max
SR 408	21.4	EB	57	55	0.38	0.18	0.1	1
		WB	56	55	0.39	0.18	0.1	1

3.3 Roadway Geometric Data (RCI Data)

Efficient roadway geometric design has a significant impact on the roadway operation and safety, and this has been verified in previous research and study (M. Ahmed, Huang, et al., 2011; Christoforou et al., 2011; Hossain & Muromachi, 2012; Le & Porter, 2012; Milton & Mannering, 1998; Park & Lord, 2009; Shankar et al., 1995; Venkataraman et al., 2011; Yu & Abdel-Aty, 2013b). We have collected the roadway geometry data for 2013. The Road Characteristics Inventory (RCI) database has been maintained by the Florida Department of Transportation (FDOT), and it has the complete roadway geometry data in addition to other relevant information. The RCI records 323 features and characteristics of the roadway system. For the data preparation, only some relevant variables were chosen, including AADT, number of lanes, auxiliary lane, horizontal degree of curvature, and speed limit. The expressway is divided into homogenous segments based on the selected variables. When the segment length is too short (shorter than 0.1 miles), this segment combined with the adjacent segment that shares almost the same characteristics.

3.4 Crash Data (S4A)

Florida has two types of crash reports, namely long-form crash report and short form crash report. The long form crash report involves crashes including but not limited: fatal or injury crashes, hit – and – run, criminally related crashes, DUI, and government vehicle-related crashes. When the crash does not meet the criteria under the long form crash report, short form crash reports are used to document the traffic crash. Having the two data sets together, we will have the most complete and accurate crash data which is currently available. We collect the crash data since January 1st, 2011 till December 30th, 2014 from the Signal Four Analytics online database (S4A). S4A provides many important information (i.e., time of the crash, crash coordinate (longitude and latitude information), number of vehicles involved, type and severity of the crash, the number of injuries and/or fatalities involved, weather, road surface and light condition, etc.). However, it did not provide the milepost of crashes and the direction. To locate these crashes on the selected expressway and specify their direction, ArcGIS has been used.

3.5 Data Preparation

To achieve the research objectives, data from a 19-mile section of one of the main expressways in Central Florida have been collected. The studied expressway (SR-408) sections start from the cumulative milepost 0.907 to the milepost 14.57 and from 17.376 to 21.958. The other segments have been removed because these segments were under improvement (widening) at some point during the study period. Both directions of the expressway are used in this study. Traffic data have been collected, organized, and prepared for the short-term aggregation (e.g., 5 minutes, 15 minutes, 30 minutes, and 60 minutes) and AADT analysis.

3.5.1 Short-term Aggregation Data for Developing SPFs

Each direction of the expressway has 55 detectors installed on the studied segments. The average spacing between the detectors is less than 1 mile. The real-time traffic data were obtained from the Microwave Vehicle Detection System (MVDS) which is maintained by the Central Florida Expressway Authority (CFX). Traffic data from July 2013 to the end of March 2014; and from May 2014 to the end of July 2014 were obtained for the analysis (April 2014 data were not archived).

To develop SPFs and to validate the results, the traffic data has been divided into two data sets. The first data set was used to develop SPFs and covers six months from July 2013 to December

2013. The remaining six months' traffic data were used for validation. Figures 3-3 and 3-4 show the traffic volume variation (based on 15 minutes accumulated volume for all segments) for the eastbound and westbound directions during weekdays and weekends, respectively. It also shows that it is worth investigating the weekdays and weekends, separately. In Figure 3A, the volume of all segments for the whole weekdays have been summed and plotted with time to present the variation in volume with time and the same has been done for the other (Figures 3-3B, 3-4A, and 3-4B).

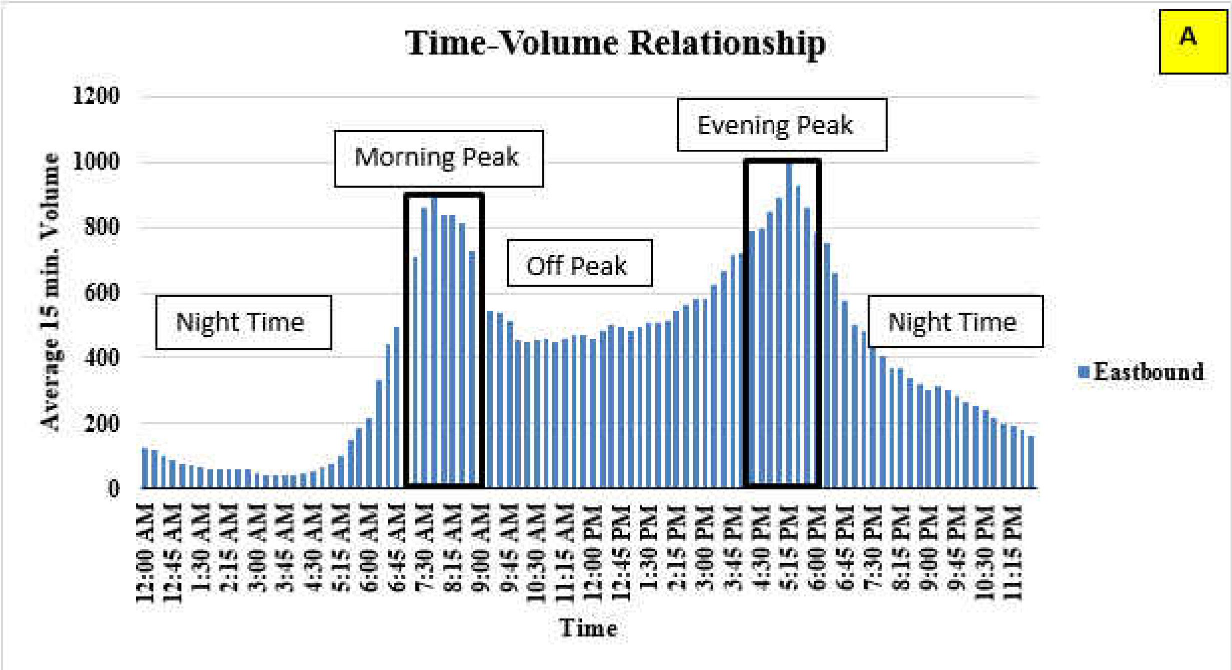
Sixteen sets of traffic data were prepared to develop different SPFs for different time intervals (5 minutes, 15 minutes, 30 minutes, and 60 minutes) for both weekdays and weekends for both directions. For weekdays, each set has been divided into four time periods (Morning Peak 7:00 am – 8:59 am, Off Peak 9:00 am – 3:59 pm, Evening Peak 4:00 pm – 5:59 pm, and Night Time 6:00 pm – 8:59 am). Figures 3-5 to 3-8 depict the variation in traffic volume based on 15-min intervals for weekdays and weekends for both directions. Figure 3-9 illustrates the traffic volume based on AADT.

While for weekends, since there is no significant variation in traffic volume as, during weekdays, only two time periods (High volume and Low volume) were considered. Since the traffic variations on each side is different, two different high volume and low volume duration were considered for each direction, eastbound-weekends (High volume 10:00 am – 10:59 pm and Low Volume 11:00 pm – 9:59 am) and westbound-weekends weekends (High volume 8:00 am – 10:59 pm and Low Volume 11:00 pm – 9:59 am) as shown in Figure 3-4.

The volume for each time interval (5 min., 15 min., 30 min., and 60 min.) has been calculated by aggregated all the volume during this time interval for the whole weekdays and weekends for each segment for each direction, separately. Then these aggregated volumes have been divided into different time periods and used to develop the models.

The crash data have been collected for four years from 2011 until the end of 2014. The crash data from 2011 to 2013 were used to develop SPFs, and the crash data in 2014 were used for validation. Crashes that occurred between Saturday 12:00 a.m. and Monday 12:00 a.m. were considered as weekend crashes while the other crashes labeled weekday crashes. From 2011 to 2014, there were 2090 crashes (740 crashes on the eastbound in weekdays, 217 crashes on the eastbound on weekends, 939 crashes on the westbound on weekdays, and 194 crashes on the westbound on weekends).

Roadway geometric characteristics' data were downloaded from the Roadway Characteristics Inventory (RCI) database which is maintained by the Florida Department of Transportation (FDOT). The collected roadway geometric data include the number of lanes, speed limit, auxiliary lanes, horizontal curvature, and AADT. The expressway is divided into 68 homogeneous segments based on these geometric characteristics. Most of these segments have one detector, some have more than one detector, and some have no detectors. In case there is more than one detector in the segment, the average of the traffic data has been calculated to represent this segment, while in case there is no detector in the segment, the traffic data of the closest detector to this segment has been used. After dividing the expressway into 68 homogenous segments based on the geometric data and combining them with the real-time traffic data and crash data, SPFs have been developed for all time periods and intervals for both weekdays and weekends and both directions.



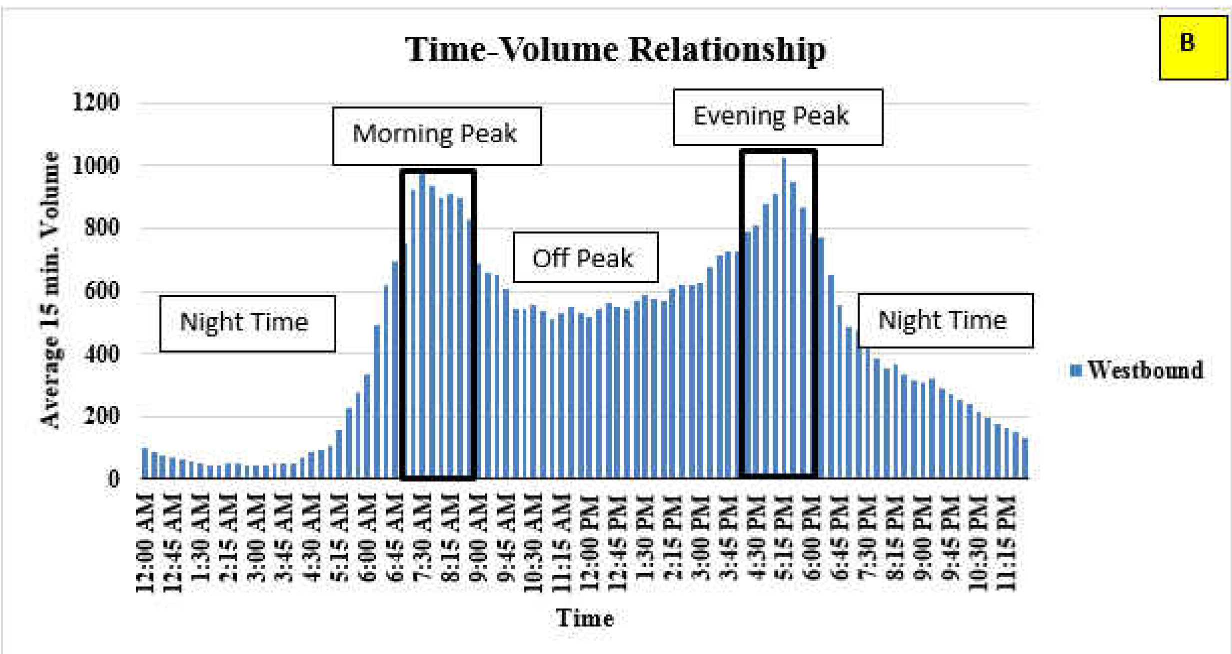


Figure 3-3: Traffic volume variation for all segment combined during (A) eastbound-weekdays and (B) westbound-weekdays.

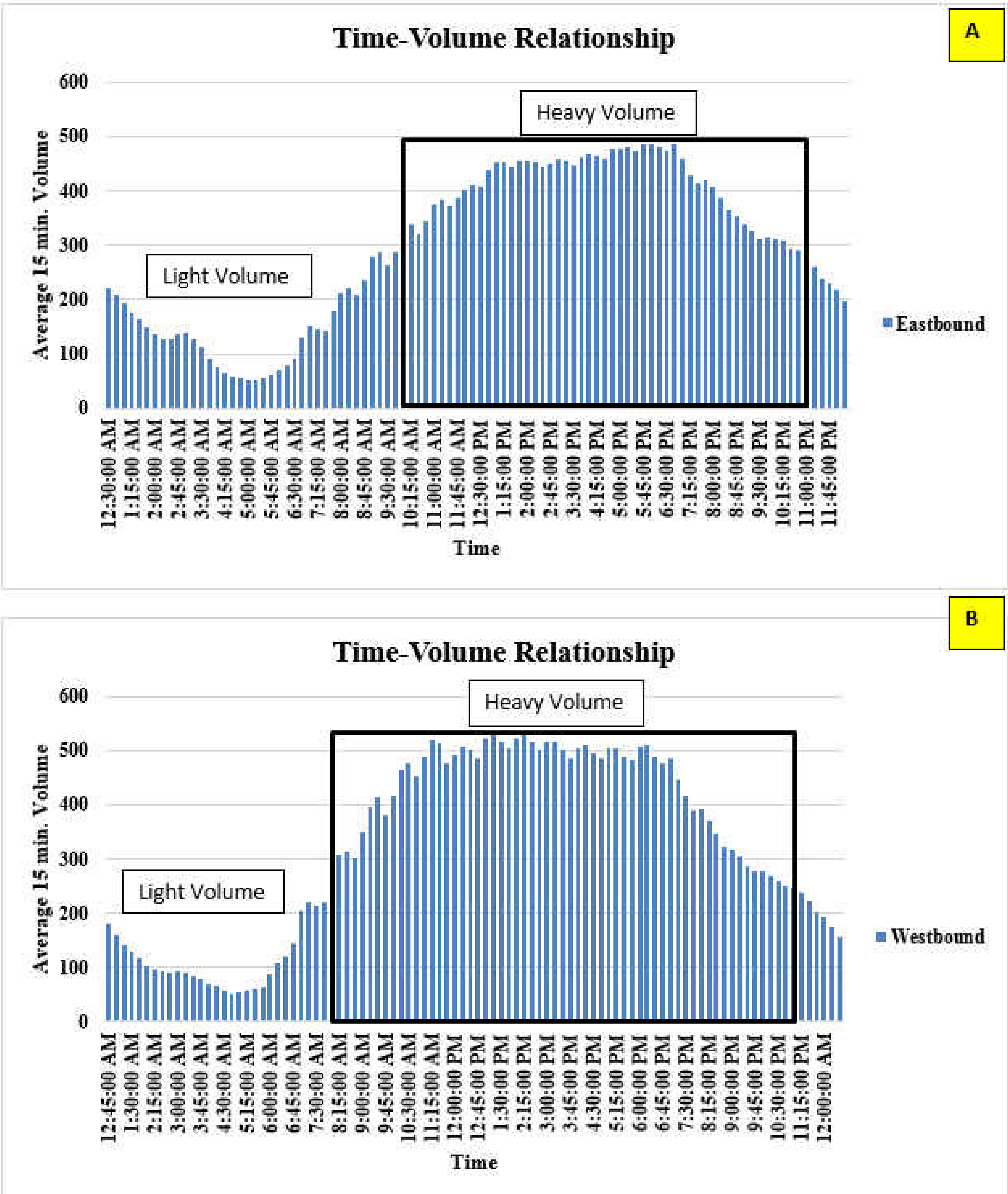


Figure 3-4: Traffic volume variation for all segment combined during (A) eastbound-weekends and (B) westbound-weekends.

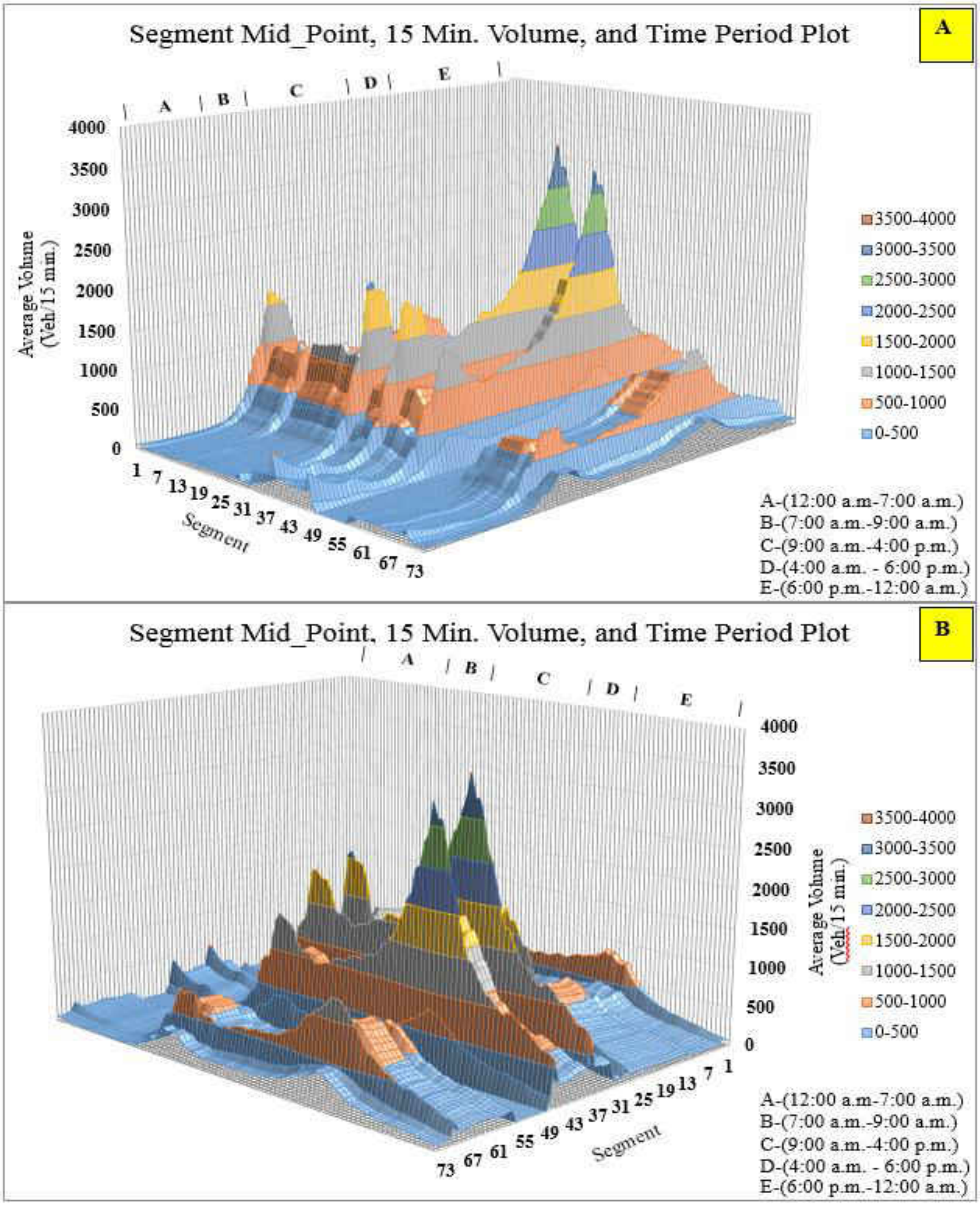


Figure 3-5: The 15 Minute Traffic Volume for each Segment during the Day Eastbound-Weekdays (A and B show different views).

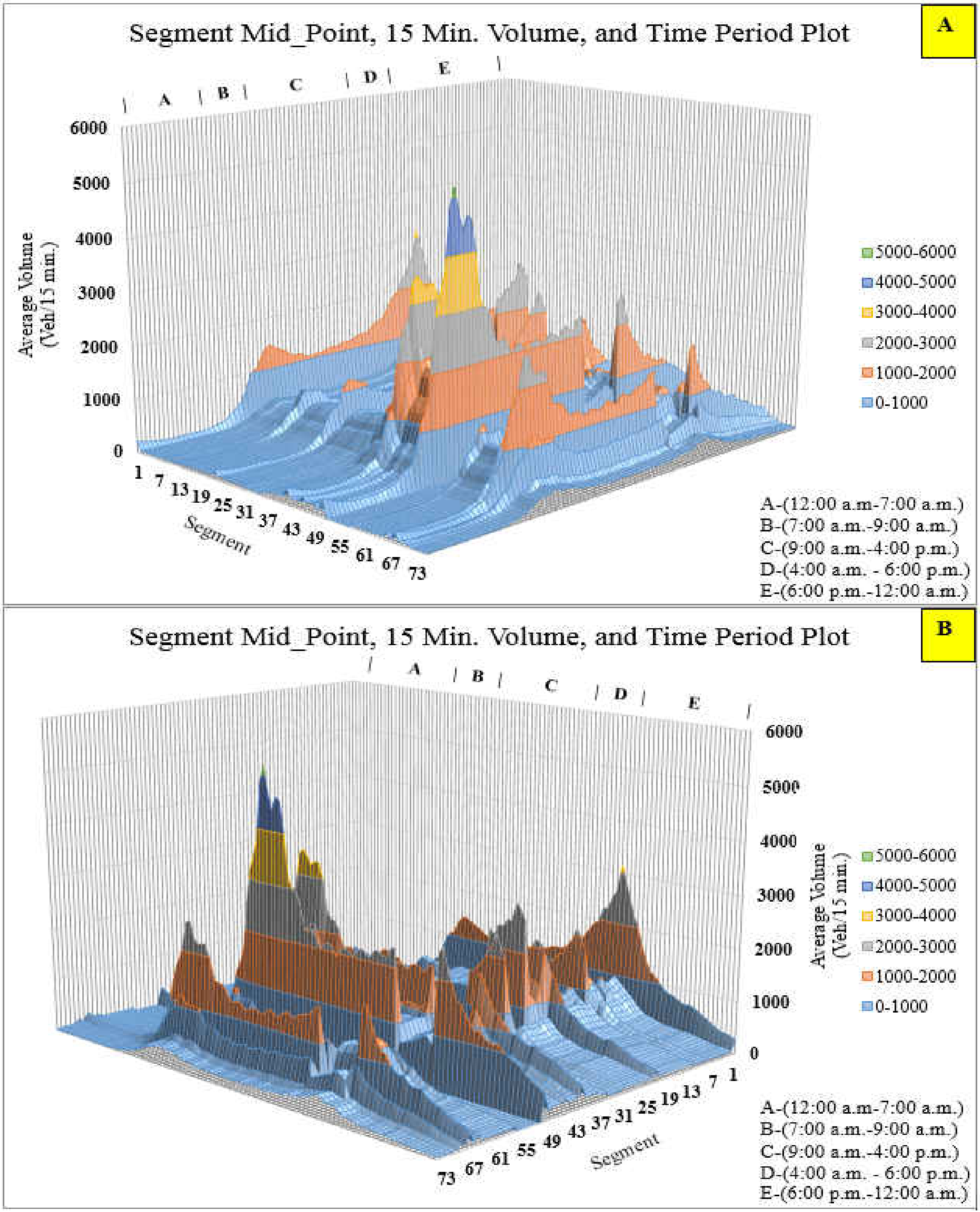


Figure 3-6: The 15 Minute Traffic Volume for each Segment during the Day Westbound-Weekdays (A and B show different views).

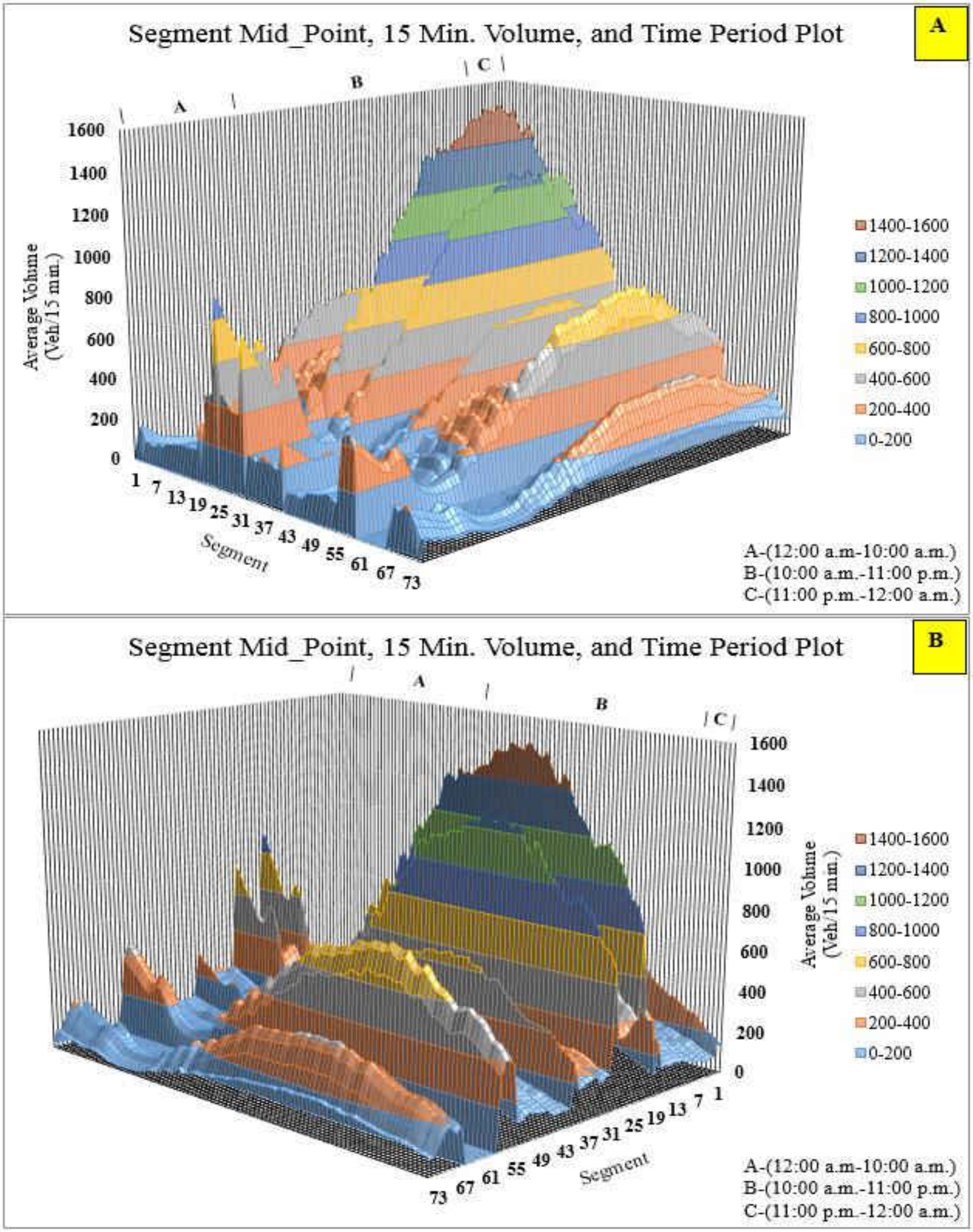


Figure 3-7: The 15 Minute Traffic Volume for each Segment during the Day Eastbound-Weekends (A and B show different views).

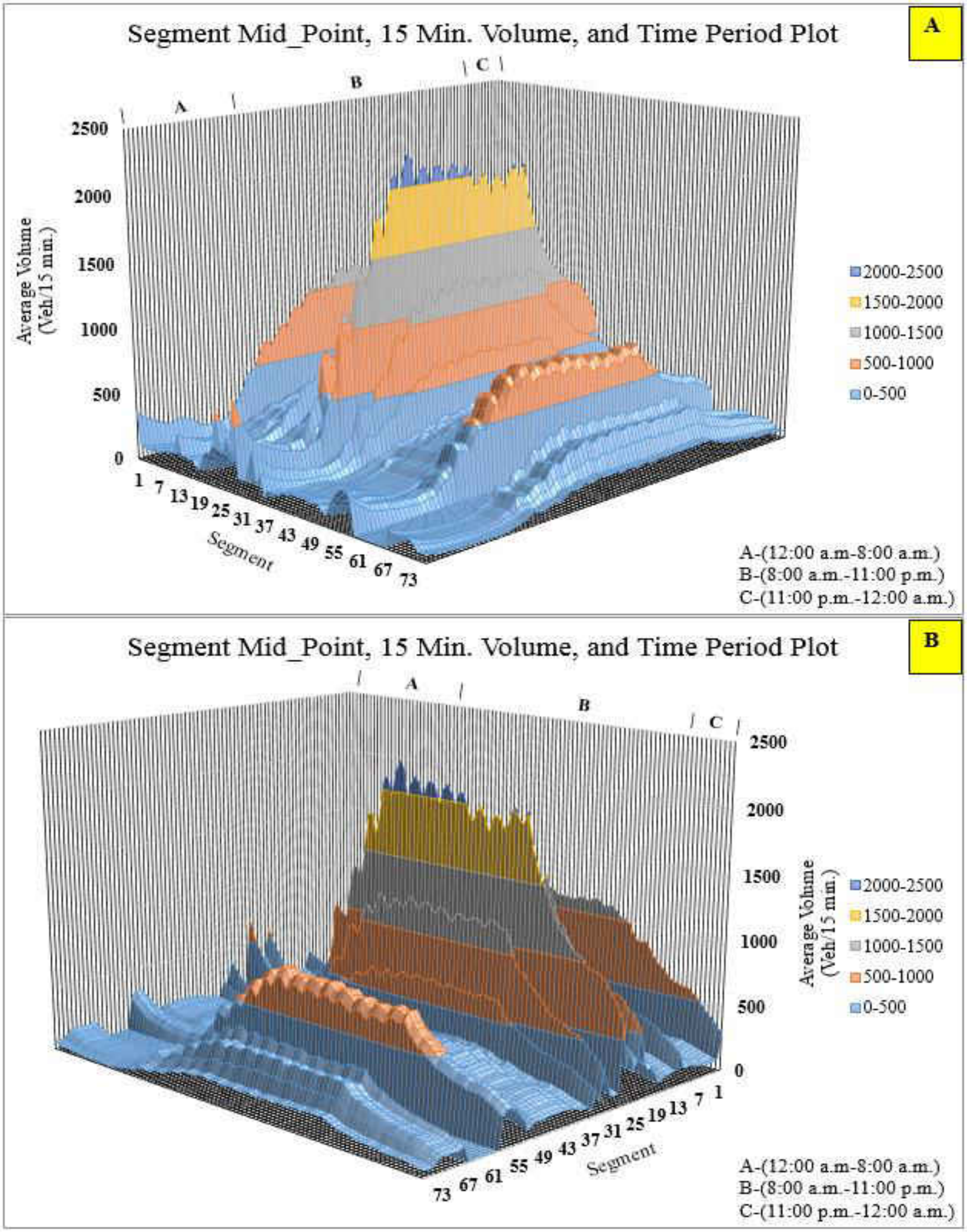


Figure 3-8: The 15 Minute Traffic Volume for each Segment during the Day Westbound-Weekends (A and B show different views).

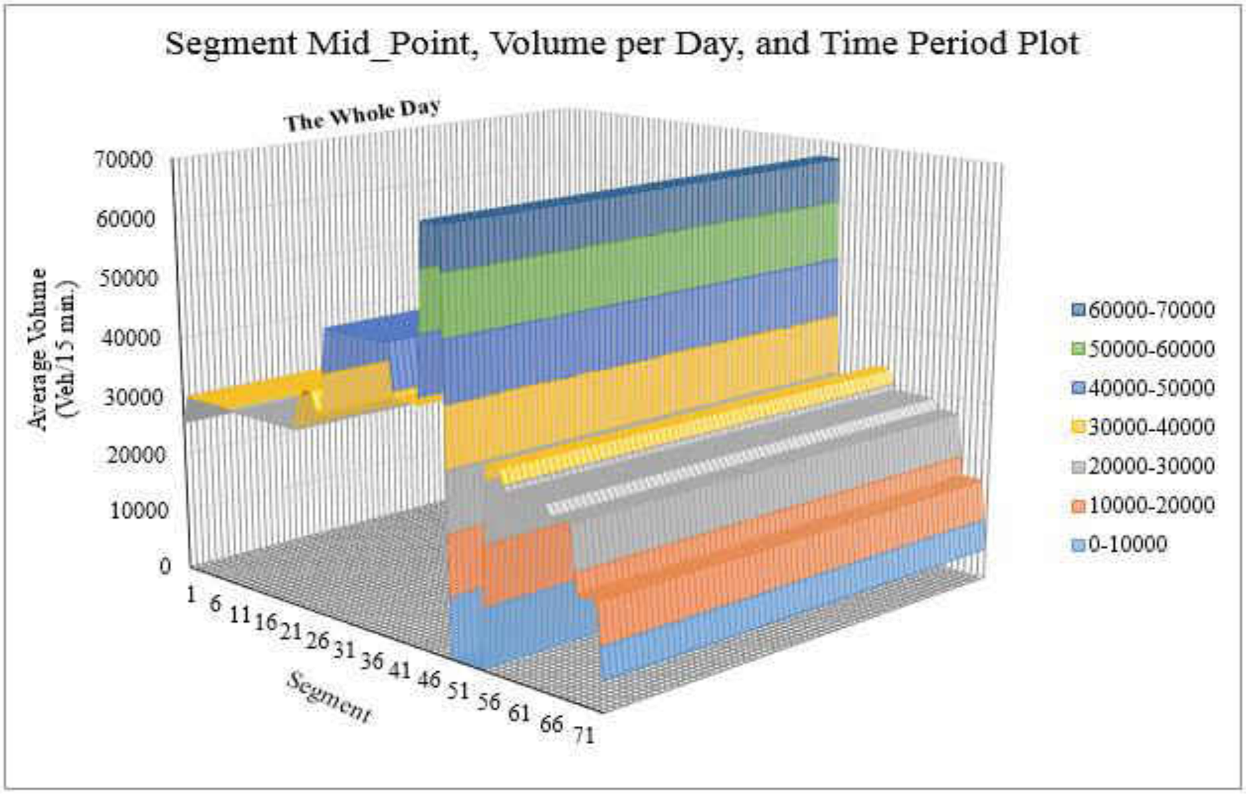


Figure 3-9: The Average Traffic Volume per Day (AADT) for each Segment during the Day.

3.5.2 Short-term Aggregate Data for Validation

The same data structure that has been used to develop SPFs for the short-term aggregated data is used, but the real-time traffic data is from January 2014 to March 2014 and from May 2014 to July 2014 (data for April 2014 was not available), while the crash data covers the entire year (2014). During 2014 there were 208 crashes for the eastbound direction (161 crashes during the weekdays and 48 crashes during the weekends) and 320 for westbound direction (270 crashes during the weekdays and 50 crashes during the weekends).

3.5.3 Aggregate Data for SPF Modeling

The aggregated traffic data (AADT) for 2013 has been collected from the RCI database. In addition to the traffic data, the directional distribution factor and other roadway geometric characteristics were collected from the same database. The crash data that has been collected for the disaggregated data has also been used.

3.5.4 Aggregate Data for SPF Validation

Traffic, roadway geometric, and crash data for 2014 have been collected from the RCI database and Signal 4 Analytics to validate the modeling results. Crashes have been assigned to each homogeneous segment and for the specific time interval. Crash data that conducted from S4A has the time of the crash, crash coordinate, number of vehicles involved, type of severity, etc. To identify the crash location for each road, based on the crash coordinate, ArcGIS software was used. After the process, all the crash data and identify the milepost of each crash, each crash assigned to the specific segment based on the mile post that we got from the ArcGIS and the time that is already included in the archived data.

3.6 Safety Performance Function Estimation

The Negative Binomial (NB) modeling approach was adopted in this study as recommended by the current Highway Safety Manual (HSM). The number of crashes is not normally distributed, due to the non-negative integers' characteristics of Poisson regression models, Poisson regression models were used to analyze crash frequency data (Jovanis & Chang, 1986). These models are easy to estimate and straightforward to explain. However, these models were criticized because of their lack of ability to handling the over-dispersion problem (M. Ahmed, Huang, et al., 2011; Guo et al., 2010; Miaou & Lord, 2003; Shankar et al., 1998; Yu et al., 2013). The previously mentioned studies used NB models and assumed a log-linear relationship between crash frequency and the explanatory variables. Kononov et al. (2011) had specified sigmoid and other non-linear relationship between crash frequency and explanatory variables to get a more accurate SPF. However, these models are sophisticated and difficult to estimate and explain (Lord & Mannering, 2010). For simplicity and to account for the over-dispersion problem and to be consistent with HSM (Manual, 2010), NB models are used in this study. Equation (1) shows the log-linear relationship between the crash frequency and the exposure variable:

$$N_{spf} = e^{[a + b * Ln(X) + Ln(L)]} \tag{3-1}$$

Where N_{spf} = the total expected number of crashes for a roadway segment, X is either the AADT or the volume on the segment, L the length of the roadway segment (miles), a and b are the regression coefficients.

The value of the over-dispersion that is associated with the calculated N_{spf} is determined as a function of the segment's length, equation (2) shows the over-dispersion formula:

$$k = \frac{1}{e^{[c + Ln(L)]}} \tag{3-2}$$

Where k = the over-dispersion parameter associated with the roadway segment length, L = the length of the roadway segment, and c = a regression coefficient used to compute the over-dispersion parameter.

To compare the performance of using different time intervals and periods for both weekdays and weekends with the performance of using only the AADT, SPFs have been developed by using the collected and prepared data.

Figure 3-10 shows the developed SPF for SR 408; the green rectangular refers to significant SPF while red rectangular indicates the SPF was either not significant or the model did not converge. Tables 3-2 to 3-5 show the details of the developed SPFs for all the suggested time periods and intervals for weekdays and weekends, respectively. For the weekdays, all the developed SPFs' independent variables are significant at the 95% confidence level. While for the weekend, only the 30 and 60 minutes interval have both High and Low Volume models significant at the 95% confidence level.

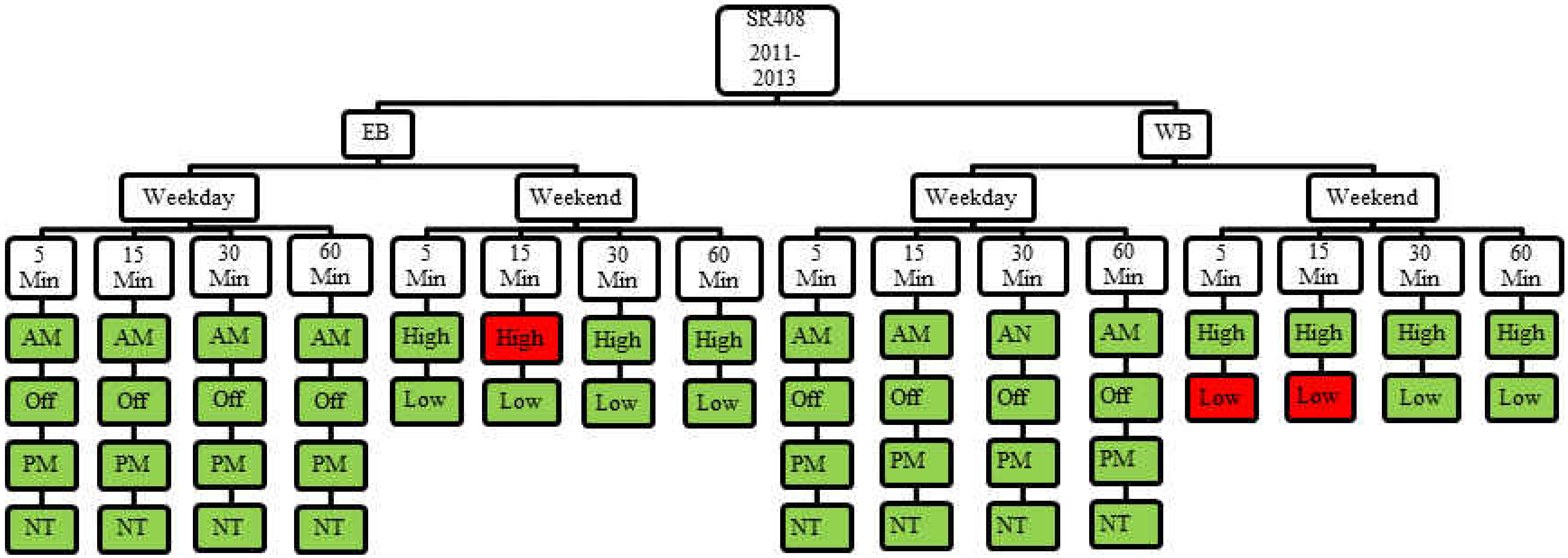


Figure 3-10: The Developed Models for SR 408.

Table 3-2. SPFs for different time periods and intervals for weekdays (Eastbound).

	5 minutes			
Parameters	Morning-Peak (7:00 a.m.-8:59a.m.)	Off-Peak (9:00 a.m.-3:59 p.m.)	Evening-Peak (4:00 p.m.-5:59 p.m.)	Night-Time (6:00 p.m.-6:59 a.m.)
Intercept	-7.1836 (<.0001)	-6.6350 (<.0001)	-8.2492 (<.0001)	-7.2171 (<.0001)
LogVol.	0.6368 (0.0267)	0.4713 (0.0087)	0.8421 (<.0001)	0.6127 (<.0001)
c	0.6903 (0.4699)	14.4669 (0.8800)	2.2229 (0.3715)	0.4029 (0.5766)
	15 minutes			
Parameters	Morning-Peak (7:00 a.m.-8:59a.m.)	Off-Peak (9:00 a.m.-3:59 p.m.)	Evening-Peak (4:00 p.m.-5:59 p.m.)	Night-Time (6:00 p.m.-6:59 a.m.)
Intercept	-7.2820 (0.0003)	-6.0617 (<.0001)	-7.9877 (<.0001)	-6.8008 (<.0001)
LogVol.	0.6979 (0.0198)	0.4762 (0.0078)	0.8304 (<.0001)	0.6148 (<.0001)
c	1.4072 (0.0070)	12.7300 (0.8903)	1.1868 (0.0600)	0.6638 (0.2105)
	30 minutes			
Parameters	Morning-Peak (7:00 a.m.-8:59a.m.)	Off-Peak (9:00 a.m.-3:59 p.m.)	Evening-Peak (4:00 p.m.-5:59 p.m.)	Night-Time (6:00 p.m.-6:59 a.m.)
Intercept	-7.2657 (0.0006)	-6.0217 (<.0001)	-7.4920 (<.0001)	-6.4759 (<.0001)
LogVol.	0.7202 (0.0114)	0.5214 (0.0081)	0.7799 (0.0001)	0.6074 (<.0001)
c	13.4574 (0.9014)	1.2059 (0.0143)	1.2744 (0.0092)	1.7926 (0.0437)
	60 minutes			
Parameters	Morning-Peak (7:00 a.m.-8:59a.m.)	Off-Peak (9:00 a.m.-3:59 p.m.)	Evening-Peak (4:00 p.m.-5:59 p.m.)	Night-Time (6:00 p.m.-6:59 a.m.)
Intercept	-7.2596 (0.0040)	-6.9286 (<.0001)	-7.3679 (<.0001)	-6.0489 (<.0001)
LogVol.	0.7062 (0.0261)	0.6916 (0.0003)	0.7757 (0.0002)	0.5893 (<.0001)
c	13.7140 (0.9200)	1.6035 (0.0007)	1.4545 (0.0005)	2.3649 (0.0204)
Parameters	**AADT**			
Intercept	-4.9636 (0.0059)			
LogAADT	0.6897 (0.0001)			
c	2.8897 (<.0001)			

c = a regression coefficient to calculate the over-dispersion parameter (see Equation 2)

Table 3-3. SPFs for different time periods and intervals for weekdays (Westbound).

	5 minutes			
Parameters	Morning-Peak (7:00 a.m.-8:59a.m.)	Off-Peak (9:00 a.m.-3:59 p.m.)	Evening-Peak (4:00 p.m.-5:59 p.m.)	Night-Time (6:00 p.m.-6:59 a.m.)
Intercept	-9.8807 (<.0001)	-9.0901 (<.0001)	-7.0806 (<.0001)	-5.7661 (<.0001)
LogVol.	1.2354 (<.0001)	0.9144 (<.0001)	0.5891 (0.0088)	0.2917 (<.0001)
c	0.6993 (0.0245)	0.5455 (0.3996)	0.6674 (0.5430)	-0.9645 (0.0046)
	15 minutes			
Parameters	Morning-Peak (7:00 a.m.-8:59a.m.)	Off-Peak (9:00 a.m.-3:59 p.m.)	Evening-Peak (4:00 p.m.-5:59 p.m.)	Night-Time (6:00 p.m.-6:59 a.m.)
Intercept	-10.3117 (<.0001)	-9.9697 (<.0001)	-6.4463 (<.0001)	-4.7622 (<.0001)
LogVol.	1.2583 (<.0001)	1.0655 (<.0001)	0.5640 (0.0143)	0.2437 (0.0012)
c	0.4420 (0.0743)	0.6142 (0.1046)	0.8544 (0.2426)	0.2107 (0.6223)
	30 minutes			
Parameters	Morning-Peak (7:00 a.m.-8:59a.m.)	Off-Peak (9:00 a.m.-3:59 p.m.)	Evening-Peak (4:00 p.m.-5:59 p.m.)	Night-Time (6:00 p.m.-6:59 a.m.)
Intercept	-10.2884 (<.0001)	-11.0883 (<.0001)	-6.5424 (0.0008)	-3.8953 (<.0001)
LogVol.	1.2254 (<.0001)	1.2195 (<.0001)	0.6209 (0.0144)	0.1805 (0.0225)
c	0.6233 (0.0207)	0.2051 (0.3762)	0.4970 (0.2419)	0.6957 (0.1134)
	60 minutes			
Parameters	Morning-Peak (7:00 a.m.-8:59a.m.)	Off-Peak (9:00 a.m.-3:59 p.m.)	Evening-Peak (4:00 p.m.-5:59 p.m.)	Night-Time (6:00 p.m.-6:59 a.m.)
Intercept	-11.3299 (<.0001)	-12.2198 (<.0001)	-5.3456 (0.0121)	-3.0205 (<.0001)
LogVol.	1.3157 (<.0001)	1.3553 (<.0001)	0.5049 (0.0484)	0.1339 (0.0981)
c	0.7023 (0.0333)	0.2283 (0.2688)	0.6572 (0.0770)	1.5188 (0.0067)
Parameters	**AADT**			
Intercept	-6.1321 (0.0074)			
LogAADT	0.8227 (0.0002)			
c	1.9243 (<.0001)			

c = a regression coefficient to calculate the over-dispersion parameter (see Equation 2)

Table 3-4. SPFs for different time periods and intervals for the weekends (Eastbound).

	5 minutes	
Parameters	**High Volume** (10:00 a.m.-10:59 p.m.)	**Low Volume-** (11:00 p.m.-9:59 a.m.)
Intercept	-7.2566 (<.0001)	-6.3976 (<.0001)
LogVol.	0.5618 (0.0243)	0.5550 (0.0009)
c	1.9489 (0.2206)	0.9615 (0.1890)
	15 minutes	
Parameters	**High Volume** (10:00 a.m.-10:59 p.m.)	**Low Volume** (11:00 p.m.-9:59 a.m.)
Intercept	-7.9930 (<.0001)	-6.6596 (<.0001)
LogVol.	0.5885 (0.0125)	0.4803 (0.0021)
c	12.5527 (-)	11.7087 (-)
	30 minutes	
Parameters	**High Volume** (10:00 a.m.-10:59 p.m.)	**Low Volume-** (11:00 p.m.-9:59 a.m.)
Intercept	-7.9684 (<.0001)	-6.6057 (<.0001)
LogVol.	0.6212 (0.0100)	0.5382 (0.0006)
c	12.6000 (0.9193)	2.2959 (0.4785)
	60 minutes	
Parameters	**High Volume** (10:00 a.m.-10:59 p.m.)	**Low Volume-** (11:00 p.m.-9:59 a.m.)
Intercept	-7.2566 (<.0001)	-6.3976 (<.0001)
LogVol.	0.5618 (0.0243)	0.5550 (0.0009)
c	1.9489 (0.2206)	0.9615 (0.1890)
Parameters	**AADT**	
Intercept	-4.9636 (0.0059)	
LogAADT	0.6897 (0.0001)	
c	2.8897 (<.0001)	

#Not significant at 10%.

Numbers in parenthesis are p-values.

Models in the shaded area were not converged.

- Values not acquired

Table 3-5. SPFs for different time periods and intervals for the weekends (Westbound).

	5 minutes	
Parameters	**High Volume** (8:00 a.m.-10:59 p.m.)	**Low Volume-** (11:00 p.m.-8:00 a.m.)
Intercept	-7.2518 (<.0001)	-6.6250 (<.0001)
LogVol.	0.3990 (0.0103)	0.2744 (0. 2351)#
c	-1.1622 (0.0342)	-2.7522 (<.0001)
	15 minutes	
Parameters	**High Volume** (8:00 a.m.-10:59 p.m.)	**Low Volume-** (11:00 p.m.-8:00 a.m.)
Intercept	-6.3982 (<.0001)	-6.0932 (<.0001)
LogVol.	0.3682 (0.0194)	0.3274 (0.0928)
c	-0.3224 (0.5344)	-1.8196 (0.0001)
	30 minutes	
Parameters	**High Volume** (8:00 a.m.-10:59 p.m.)	**Low Volume-** (11:00 p.m.-8:00 a.m.)
Intercept	-6.0115 (<.0001)	-6.1757 (<.0001)
LogVol.	0.3722 (0.0248)	0.4316 (0. 0601)
c	-0.2689 (0.5177)	-0.9282 (0. 0576)
	60 minutes	
Parameters	**High Volume** (8:00 a.m.-10:59 p.m.)	**Low Volume-** (11:00 p.m.-8:00 a.m.)
Intercept	-5.9703 (<.0001)	-5.9412 (<.0001)
LogVol.	0.4175 (0.0180)	0.4699 (0.0377)
c	0.01244 (0.9746)	-0.3525 (0. 4505)
Parameters	**AADT**	
Intercept	-6.1321 (0.0074)	
LogAADT	0.8227 (0.0002)	
c	1.9243 (<.0001)	

#Not significant at 10%.

Numbers in parenthesis are p-values.

Models in the shaded area were not converged.

- Values not acquired

3.7 Potential for Safety Improvement (PSI)

Excess Expected Average Crash Frequency with Empirical Bayes (EB) Adjustment was used as a performance measure for network screening. This performance measure is also known as the Potential for Safety Improvement (PSI). The Empirical Bayes is one of the accepted methods for obtaining a reliable expected number of crashes using weights calculated from the over-dispersion parameter. PSI represents the difference between the expected and the predicted number of crashes for the same location as in equation 3-3.

PSI = Expected number of crashes – Predicted number of crashes (3-3)

The expected number of crashes was driven from the predicted number of crashes as illustrated in equation 3-4:

$$N_{exp.} = w_i * N_{pred.} + (1 - w_i) * N_{obs.} \tag{3-4}$$

where $N_{exp.}$ = the expected number of crashes, $N_{pred.}$ = the predicted number of crashes from SPF, $N_{obs.}$ = the observed number of crashes, and w is calculated as shown in equation 3-5:

$$w_i = \frac{1}{1 + k_i * \left(\sum_{All\ study\ years} N_{pred.}\right)} \tag{3-5}$$

Where k_i = the over-dispersion parameter for the associated SPF for the specific segment (i) that was used to estimate $N_{pred.}$ and is calculated using Equation (3-2).

3.8 Comparison between the AADT Based PSI and Disaggregate Volumes Based PSI

Two different goodness-of-fit measures were used for the comparison as shown in Table 3-3. Mean Absolute Deviation (MAD) calculate mean absolute difference between the observed and the predicted values using equation 3-6 as follows:

$$MAD = \left| \left(\sum_{i=1}^{m} \left(\sum_{j=1}^{n} y_j - \sum_{j=1}^{n} \hat{y}_j \right) \right) \right| / m \tag{3-6}$$

Where m is the total number of observations for all segments. n is the observations in segment i, and y_j and $\hat{y}_j$ are the observed and predicted values for i, respectively.

Mean Square Prediction Error (MSPE) calculates mean square of the difference between the observed and the predicted values using equation 3-7:

$$MSPE = \left\{ \left[\sum_{i=1}^{m} \left(\sum_{j=1}^{n} y_j - \sum_{j=1}^{n} \hat{y}_j \right) \right]^2 \right\} / m \tag{3-7}$$

For the weekend, 30 minutes intervals have been used to calculate the PSI since it was the shortest time interval that has converted and significant models for both High and Low Volume. To be consistent with Highway Capacity Manual (HCM), to reduce the possibility of noise in the data when using less aggregated data and since there is no big difference between the 5 and 15 min time intervals in term of MAD and MSPE as shown in Table 5-6, the 15 minutes SPFs have been selected for weekdays then the hotspot locations specified by these SPFs were compared with the hotspot locations that are specified by the AADT SPF.

Table 3-6. The MAD and the MSPE for Weekday, Weekend, and AADT for different time intervals (Eastbound).

	Weekday				
	5 Minutes	**15 Minutes**	**30 Minutes**	**60 Minutes**	**AADT**
MAD	1.3714	1.3721	1.3728	1.3913	1.7495
MSPE	4.7712	4.7802	4.7902	4.9431	6.0586
	Weekend				
	5 Minutes	**15 Minutes**	**30 Minutes**	**60 Minutes**	**AADT**
MAD	N/A	N/A	0.6105	0.6101	1. 7495
MSPE	N/A	N/A	0.7268	0.7260	6. 0586

Table 3-7. The MAD and the MSPE for Weekday, Weekend, and AADT for different time intervals (Westbound).

	Weekday				
	5 Minutes	**15 Minutes**	**30 Minutes**	**60 Minutes**	**AADT**
MAD	2.7720	2.7747	2.7771	2.8128	3.2282
MSPE	39.4825	39.3675	39.4146	38.8716	47.3518
	Weekend				
	5 Minutes	**15 Minutes**	**30 Minutes**	**60 Minutes**	**AADT**
MAD	N/A	N/A	0.6623	0.6601	3.2282
MSPE	N/A	N/A	1.2724	1.2708	47.3518

3.9 Hotspot Identification

AADT based PSI and different time periods based PSI for weekdays and weekends were compared. Since we selected 15 minutes time intervals for weekdays, then for each segment there are fore expected numbers for crashes and four observed crash in one hour (i.e., one expected number of crashes in each 15-minute interval). While for weekends, there is two expected number of crashes and two observed crash in one hour since 30-minute interval has been selected. As it was mentioned before, there are four time periods for the weekdays and two time periods for the weekends, thus to obtain the time periods based PSI for each segment for each time period, the summation of all the expected number of crashes within that time period for that segment has been subtracted from the summation of all the predicted number of crashes within the same time period for the same segment. The AADT based PSI has been calculated by dividing the expected number of crashes based on AADT by 24 then multiplying it by the number of hours in each time period then the result subtracted from the total number of crashes during that time period. Figures 8 through 13 show the comparison between AADT based PSI and the different time periods based PSI. The blue line represents different time periods based PSI, and the orange line represents AADT based PSI. During the study time period, the section from the milepost 15.066 to the milepost 17.265 experienced construction and thus not considered. The comparison between the AADT based PSI and different time periods based PSI shows considerable differences in some periods. For example, during the morning-peak, the PSIs are significantly higher than the AADT based PSI. Some of the segments have negative AADT based PSI while they have positive morning-peak based PSI. The time period approach with 15 minutes interval for weekdays and 30 minutes interval for weekends is more specific and is able to capture the variation in traffic volume and the potential for safety improvement more accurately than the common approach which is based on the average volume for the whole day that does not account for the variation in traffic volume and the different pattern of crashes between weekdays and weekends.

Generally, AADT based PSI is insensitive and has smaller absolute values whereas specific time periods based PSI is more sensitive and relatively larger absolute values. Figures 3-11 to 3-16 show the comparison between the AADT based PSI and the summation of the different time periods based PSIs. In these Figures, the blue line represents the summation of the PSI for different time periods while the orange line is the AADT based PSI. Comparing the PSI value between weekday and weekend of the same segment shows that there are differences between them. For example, in Figure 3-15 at milepost 20, during the weekday, the weekday PSI value is lower than the AADT PSI value, while during the weekend, the AADT PSI is lower than the weekday PSI. Also, there are some considerable differences between directions' PSIs. For example, at milepost 11.5, AADT PSIs for eastbound for both weekdays and weekends are higher than the weekday or weekend PSI, while the case is opposite for the westbound direction.

The morning peak PSIs, which has the highest volume, are always more sensitive compared to its counterpart regardless of direction. Except for the weekday morning peak hours, the sensitivity heavily depends on its direction. The PSIs from specific temporal SPFs are more sensitive for the eastbound direction; however, they are less sensitive for the westbound, compared to AADT PSIs. This might be because that the volume distribution within the same interval for different segments is not consistence.

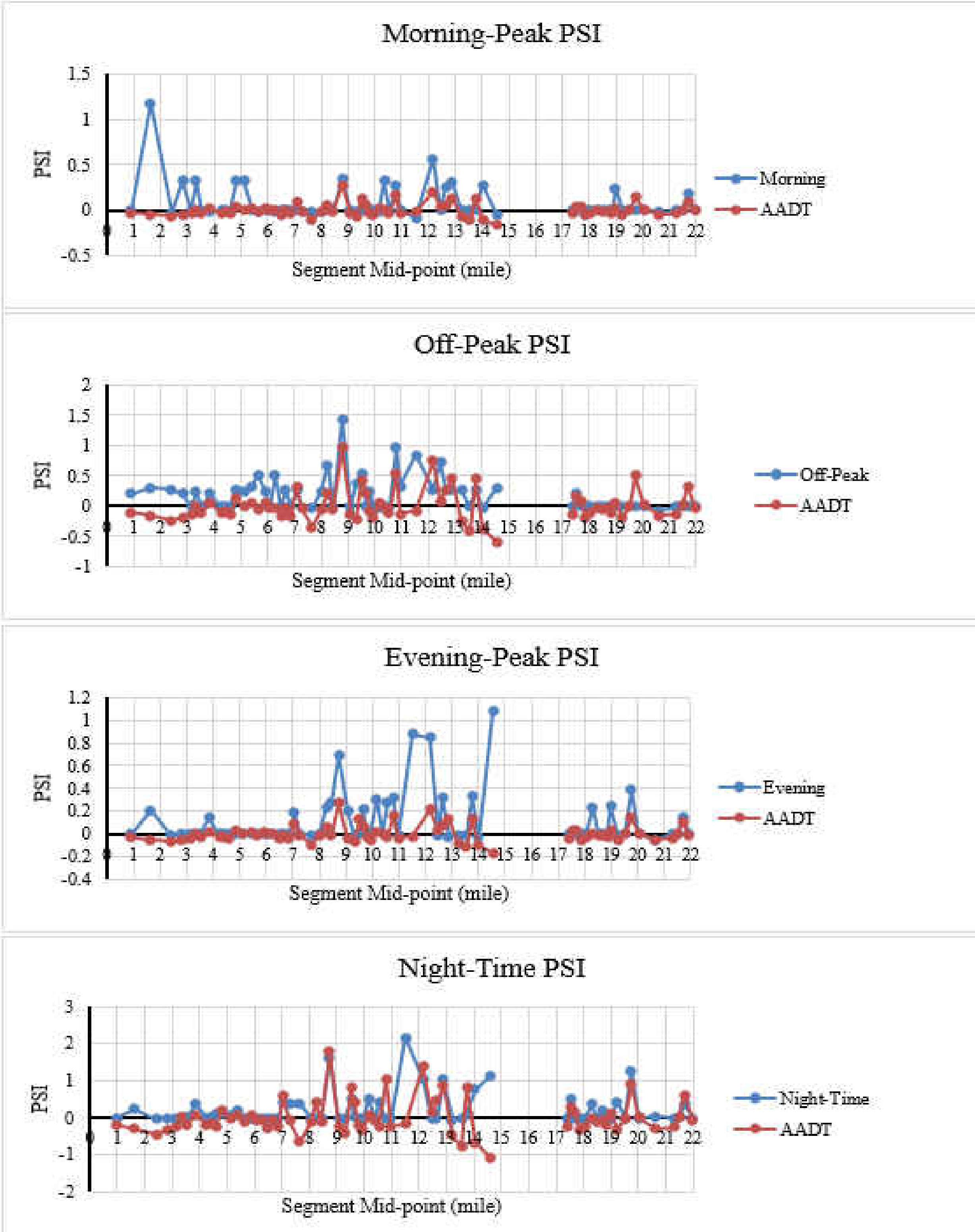


Figure 3-11. The Comparison between AADT based PSI and different time periods based PSI for eastbound weekdays.

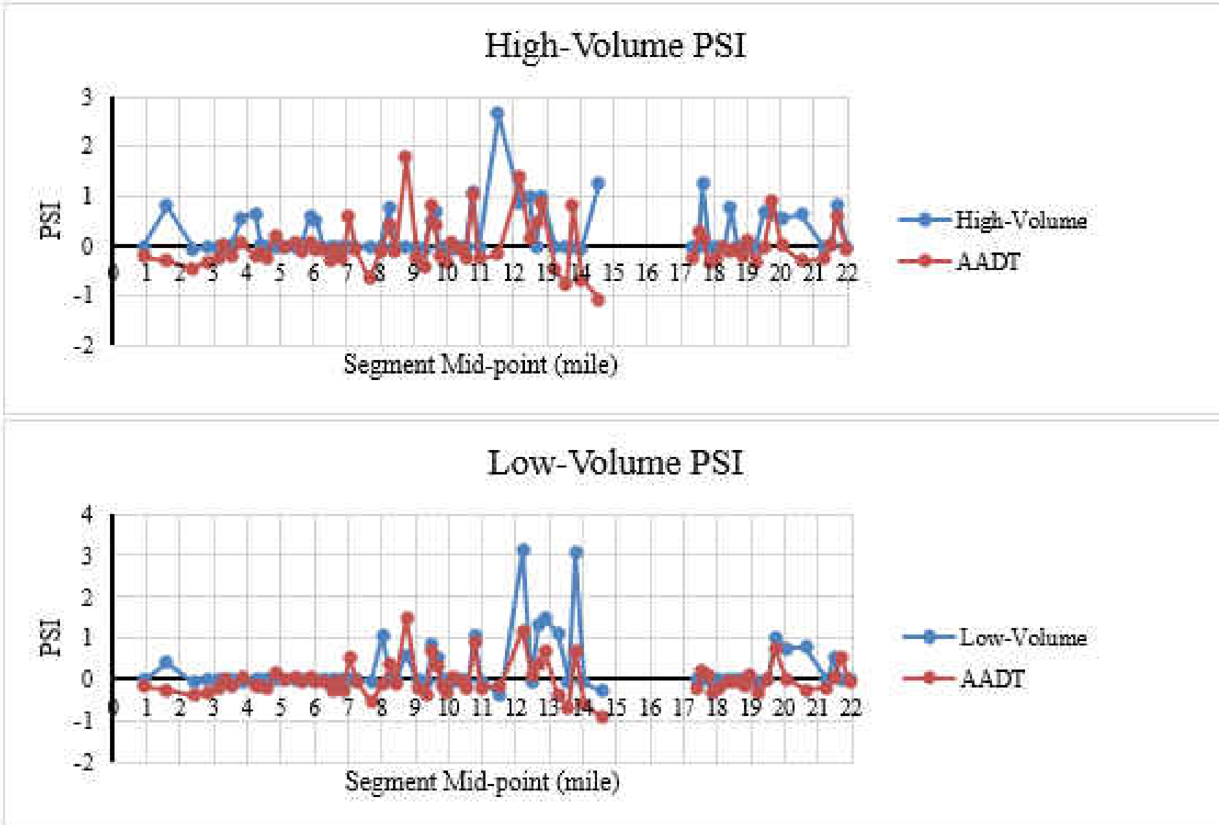


Figure 3-12. The Comparison between AADT based PSI and different time periods based PSI for Eastbound Weekends.

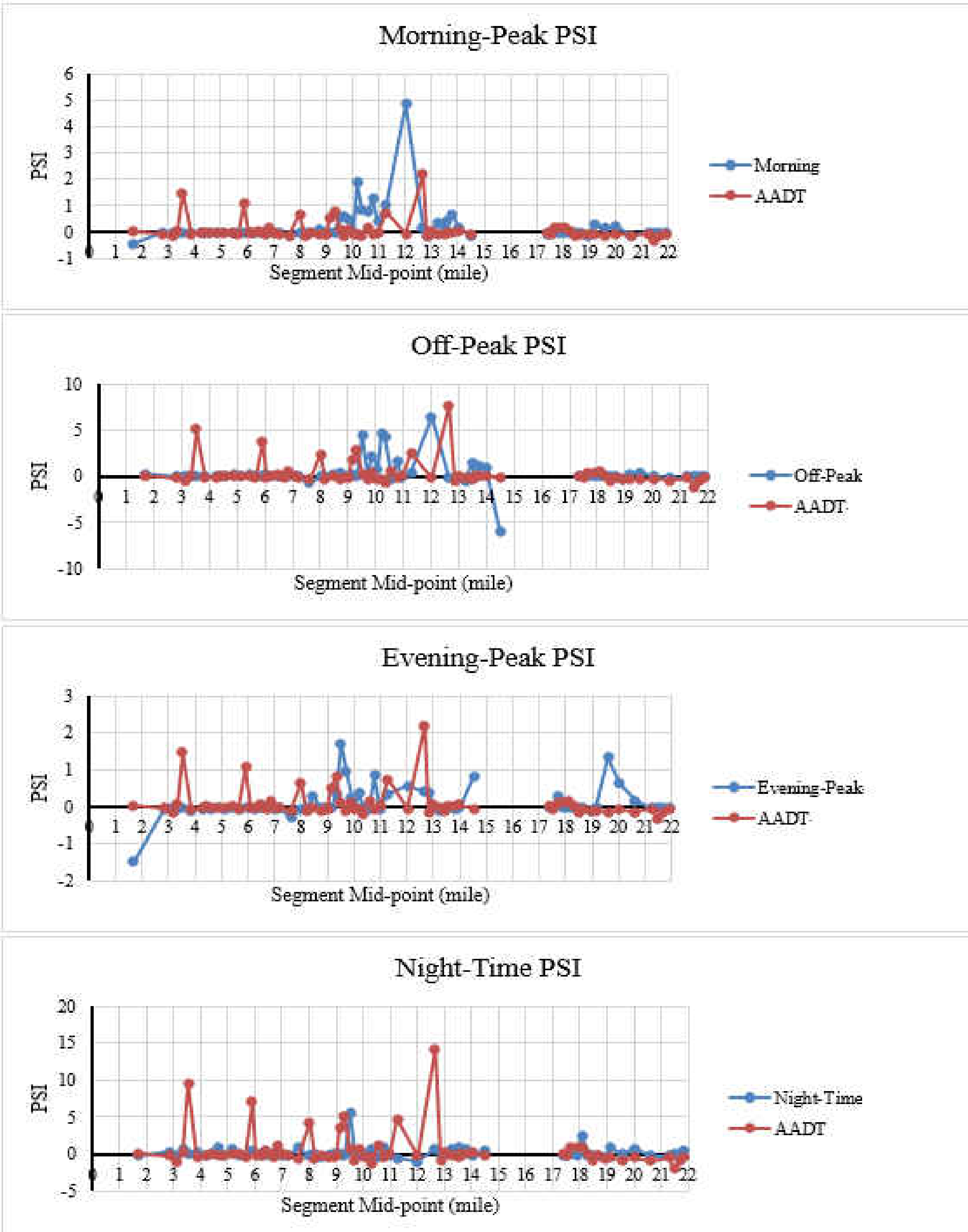
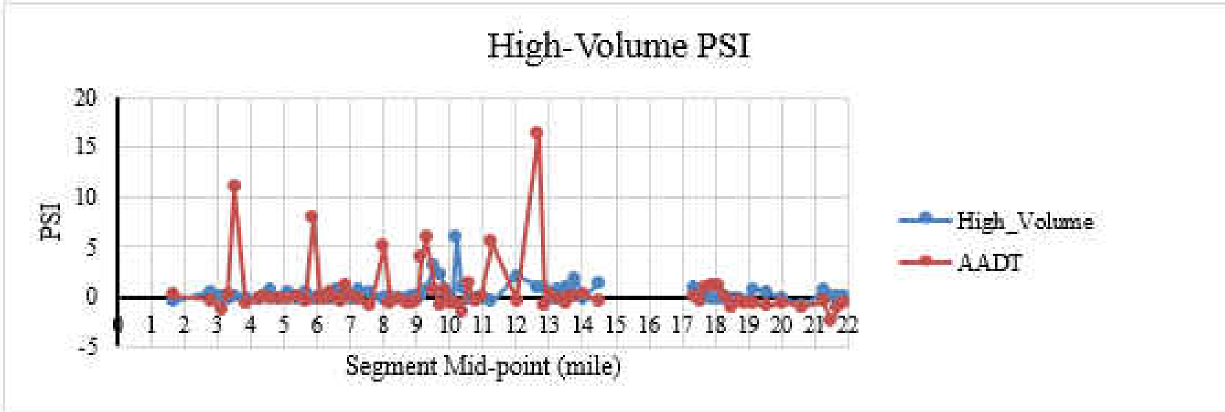


Figure 3-13. The Comparison between AADT based PSI and different time periods based PSI for Westbound Weekdays.



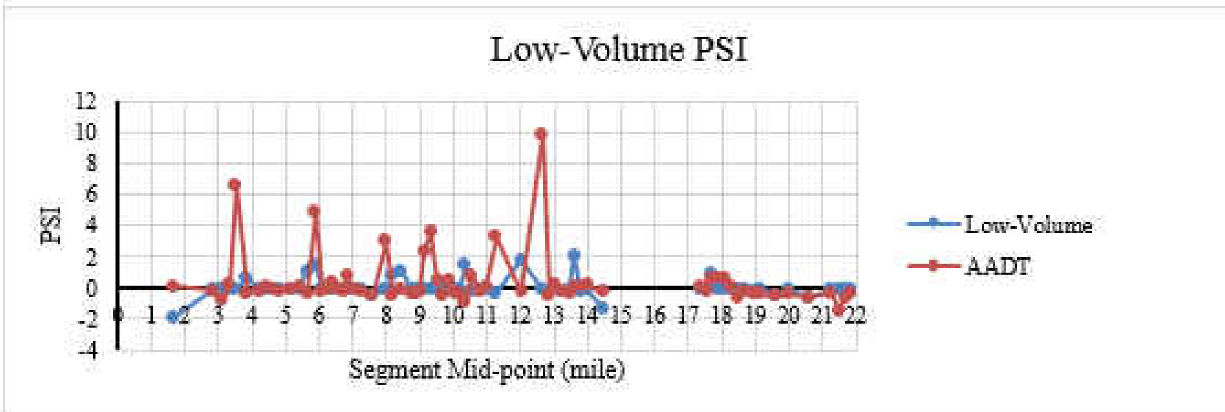


Figure 3-14. The Comparison between AADT based PSI and different time periods based PSI for Westbound Weekends.

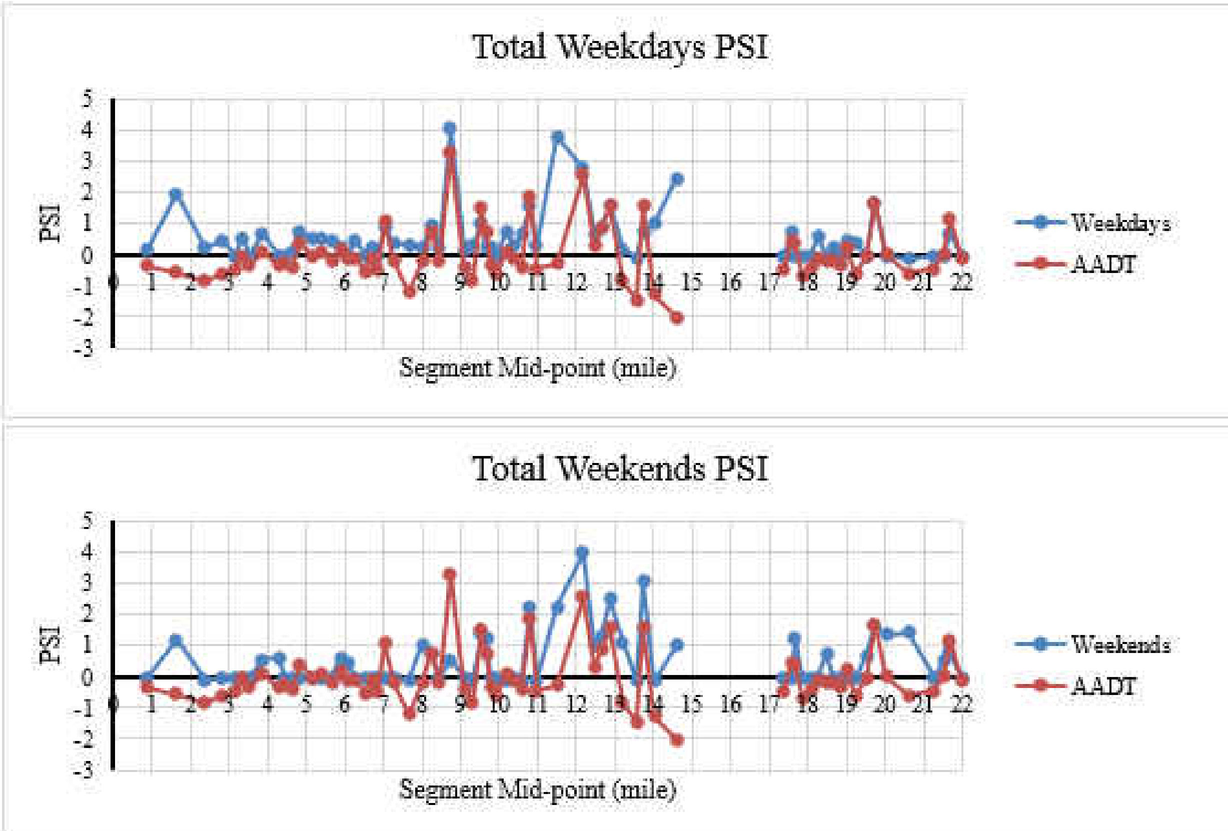


Figure 3-15. The Comparison between AADT based PSI and the summation of different time periods based PSI for weekdays and weekends (Eastbound).

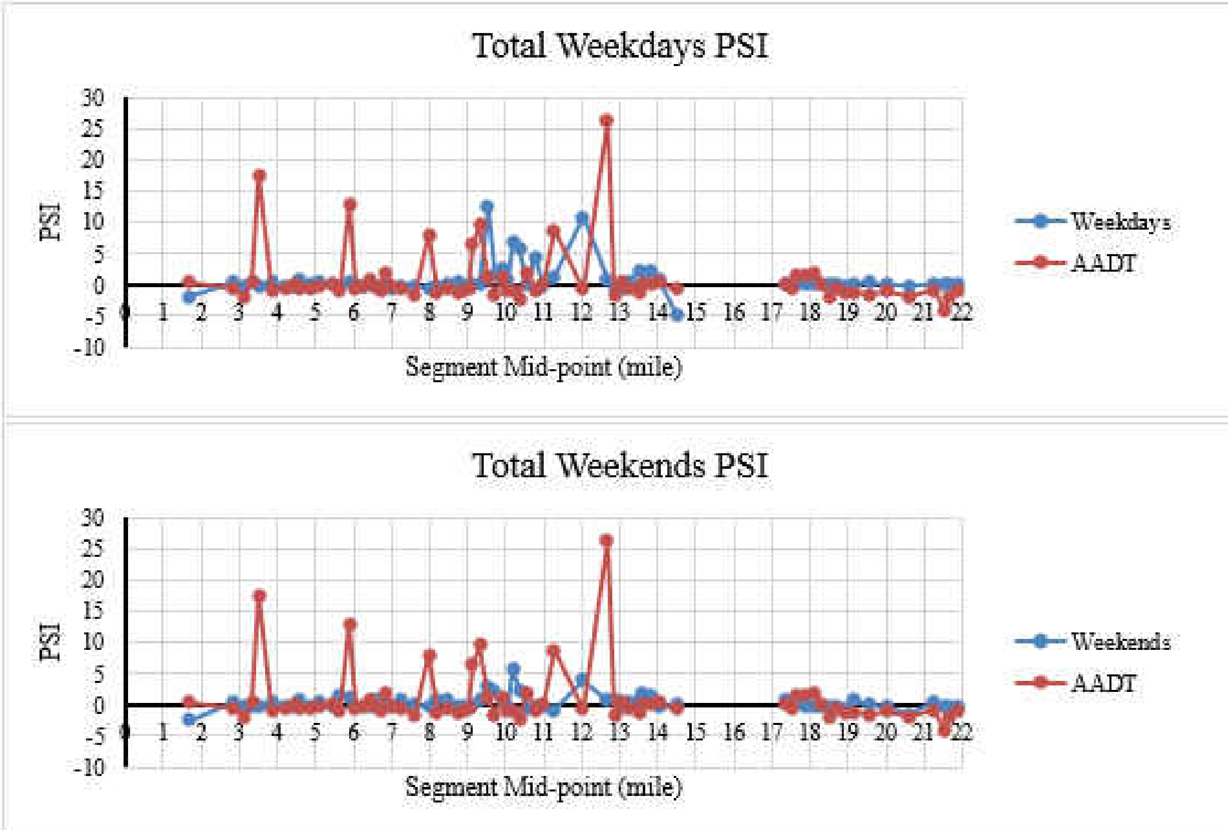


Figure 3-16. The Comparison between AADT based PSI and the summation of different time periods based PSI for weekdays and weekends (Westbound).

3.10 Conclusions

Although developing Safety Performance Functions (SPFs) are among the most important steps in traffic safety analysis, they usually have been developed using highly aggregated traffic and crash data, which may result in failure of understanding of the effect of variation in traffic volume on the crash occurrence, and their temporal relationship. Thus, a dynamic hotspot identification method is proposed in this paper using time period specific SPFs developed from less aggregated data.

The current approach for identifying hotspot locations on a roadway segment requires developing a SPF which is based on the aggregated traffic volume AADT. The disadvantage of using AADT can be listed in three points: a) AADT fails to account for the variation in traffic volume for each direction; b) It does not distinguish between the pattern of crashes and traffic during weekdays versus weekends; and c) also fails to capture the variation in the traffic volume during the different time periods in weekdays or weekends. The results from both AADT and short-term period models suggest that short-term period models perform better. Using less aggregated traffic data improved the accuracy of hotspot identification especially during the weekend in term of MAD and MSPE. Both MAD and MSPE for PSI based short-term periods have lower values than PSI based AADT. Also, in many cases, the PSI based on AADT were less sensitive to identify the hotspot locations during different time periods. Failing to identify true hotspots is much worse than identifying a safe spot as a hotspot. Also, by considering only the AADT, two different roadways with the same AADT but different traffic variance will be treated similarly. However, such different traffic pattern can be easily captured by using disaggregate traffic data.

CHAPTER 4: IMPACT OF USING SEGMENTS WITH ONE OR MORE CRASHES IN THE MODELING PROCESS ON THE ACCURACY OF THE PREDICTION

4.1 Introduction

In the previous chapter, it was found that using less aggregated traffic data improve the accuracy of the hotspot identification and reduce the error term. However, two difficulties were noticed when we considered shorter time intervals. First, as the considered time intervals get smaller, the more zero observations were observed. Second, the repeated measurement data. In this chapter, several Safety Performance Functions (SPFs) were developed with different scenarios: first, several full SPFs were developed for the whole data, in these models, a random effect was considered in the modeling process to account for the repetition in the data. Then, to reduce the number of zero observations, several Full SPFs were developed using either (segments with one or more crashes, segments with two or more crashes, or segments with three or more crashes) with the consideration of random effect to overcome the repetition in the data. Then, in term of error, a comparison between the hotspot identification based on the developed SPFs for the entire segments was adopted to identify the superior models.

4.2 Data Preparation

For the purpose of this study, traffic data (MVDS) have been collected from (July 2013 to July 2014). MVDS does not return traffic information for an individual vehicle. They aggregated and recorded traffic flow data for each lane where they are installed at one-minute interval bases. Different types of vehicles were defined by the MVDS based on their lengths:

1- Type I: vehicles 0 to 10 feet in length.

2- Type II: vehicles 10 to 24 feet in length.

3- Type III: vehicles 24 to 54 feet in length.

4- Type IV: vehicles longer than 54 feet in length.

The recorded traffic data include traffic volume, traffic volume by length, time mean speed, and lane occupancy. Also, the MVDS data include the timestamp when the sensor has recoded the data. The sensor is recorded every one minute. Additional information from the MVDS data contained sensor identifier, lane identifier, milepost, and direction. The number of lanes is counted from the inner of the roadway (the median) to the outside of the roadway (the shoulder). The roadway lanes fall into four different categories:

Mainline.

- Mainline.
- Mainline TP Express.
- Mainline TP Cash, and
- Ramp.

The types of lanes and number of lanes at each MVDS detection location can be seen in Appendix B.

There are 110 detectors on both sides of the SR 408 (55 detectors on each side). These detectors supposed to poll the traffic data every one minute for each lane in the section where they have been installed. Before starting using these data, several procedures have been made to check if the detectors were working properly for the whole day. Here are the procedures that have been done to erase all the abnormal data and incomplete daily archived data:

1- Average speed per minute should not exceed 120 mph, so any recorded average speed data that is exceeded this number considered abnormal reading and been deleted.

2- The relationship between the speed and flow of the collected data has been plotted and examined. The relationship between the speed and the volume (veh/min/lane) is presented in Figure 4-1. This Figure shows that there are several abnormal data after the volume per minute per lane exceed 60 (veh/min/ln). So, to eliminate these abnormal data, any recorded volume that exceeds 60 veh/min/lane have been deleted.

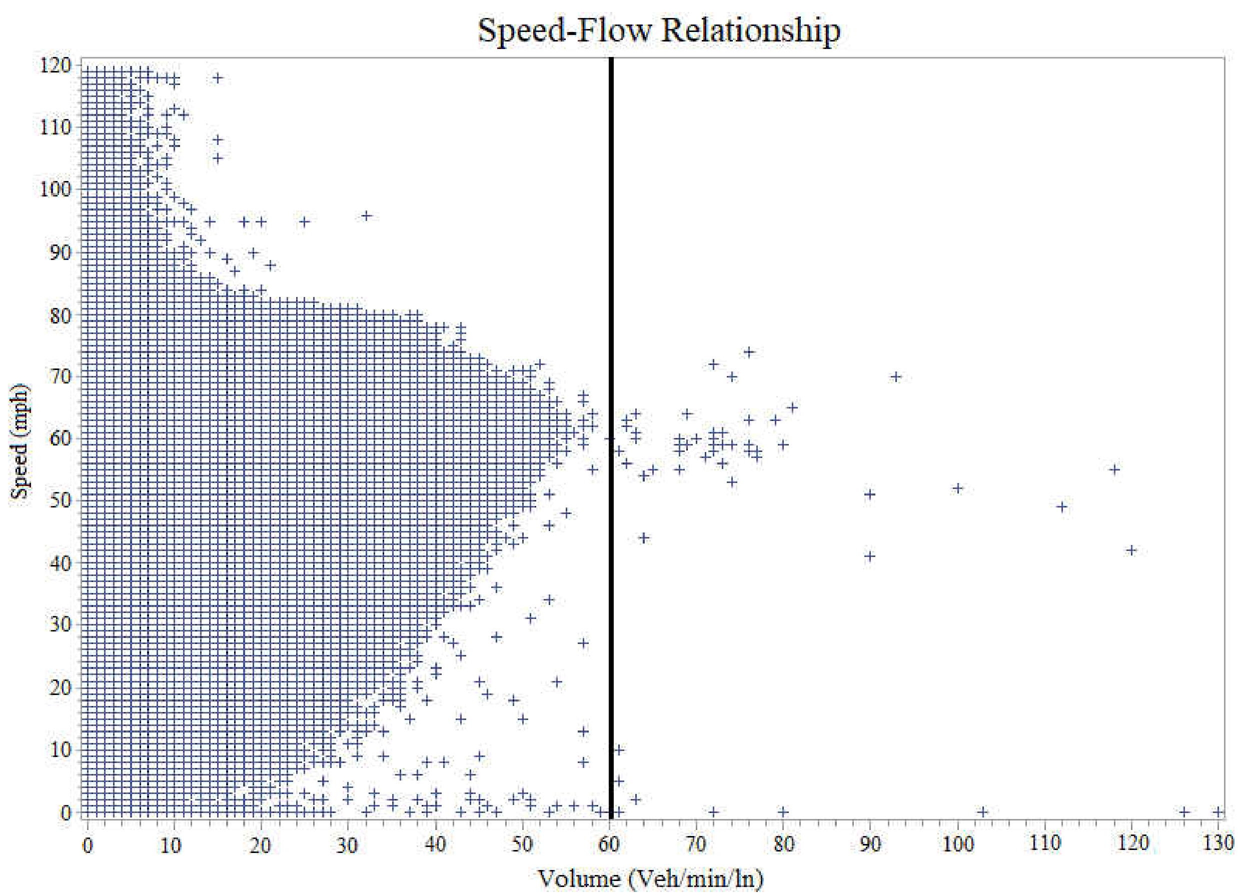


Figure 4-1 Speed-Flow Relationship.

3- Another check for the volume has been done. In the archived data, there are 4 volume categories based on the length of the detected vehicle. The summation of all four categories should be equal to the total number of the vehicle detected. The summation of the four categories has been checked with the total number of the vehicle, and in case the summation is not matched, these data have been removed.

4- In some cases, there is a recorded volume, but there is no recorded speed, this data has been removed.

5- As it has been mentioned before, the data has been archived on a one-minute base, so this issue also have been checked and several repeated recorded data for the same milepost for the same day and time has been found, in this case, only one reading data has been kept.

6- The data also consist the ramp data, so all the ramp data has been excluded since this study is focusing on segments.

7- In some cases, for one or more lanes, there is no volume recorded, but there is a speed record for that lane, the speed recorded for that lane has been adjusted to be equal to zero, and the average speed for that section is calculated with this consideration.

8- After removing all the uncorrected and unreliable recorded data, only full correct recorded days were kept and used in the data analysis (i.e., since the data is polled on a one minute bases, there should be "60 min.*24 hr. = 1440" recorded data per day to consider it a full recorded day).

After filtering the data form the abnormal observation, a comparison between actual daily volume (based on the actual collected data from the detectors) and the AADT (from RCI for 2014) for each direction has been done. The comparison shows that there is a big difference between the daily volume based on the collected data and the AADT for the same year. Figures 4-2 and 4-3 show the daily volume per direction for both the collected data from the detectors and the AADT from the RCI. As the Figures show, there is a significant difference in the daily volume from the milepost ten till fourteen (the Orlando downtown area). With this big different in the volume and with the availability of the real data, it is not recommended to use the AADT to develop SPFs and identify hotspot locations. In addition to that, AADT does not give information about the variation in the traffic during the whole day.

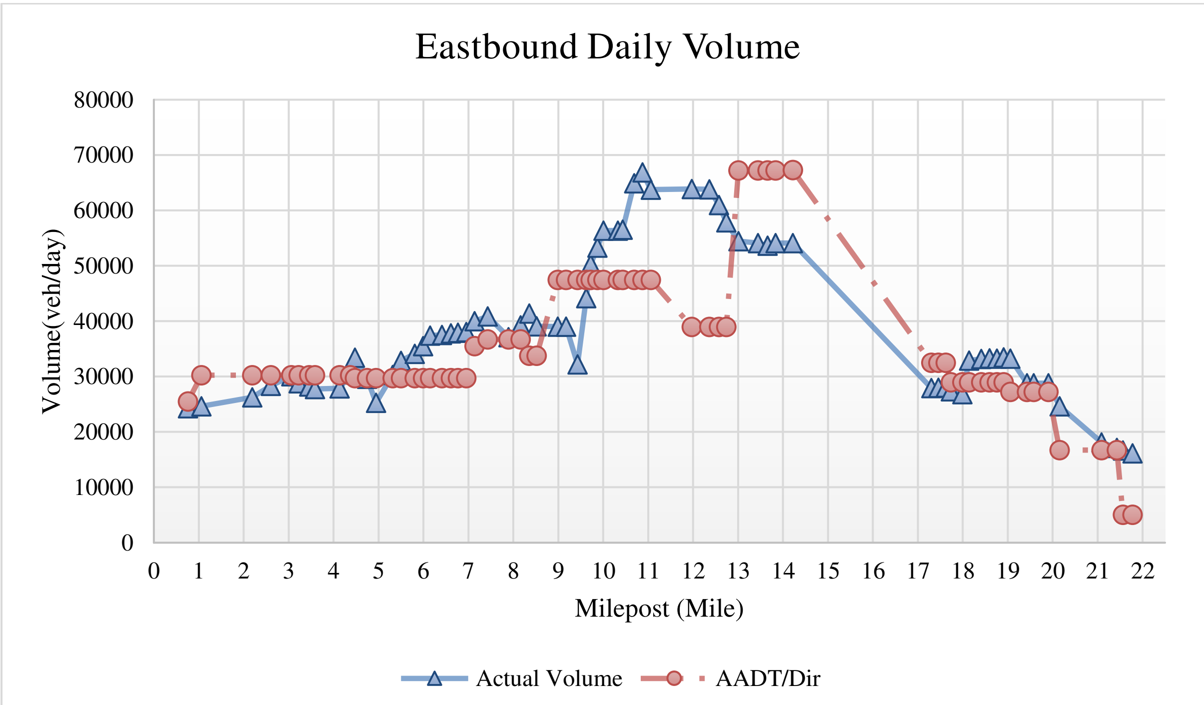


Figure 4-2 Daily volume for Eastbound (2014).

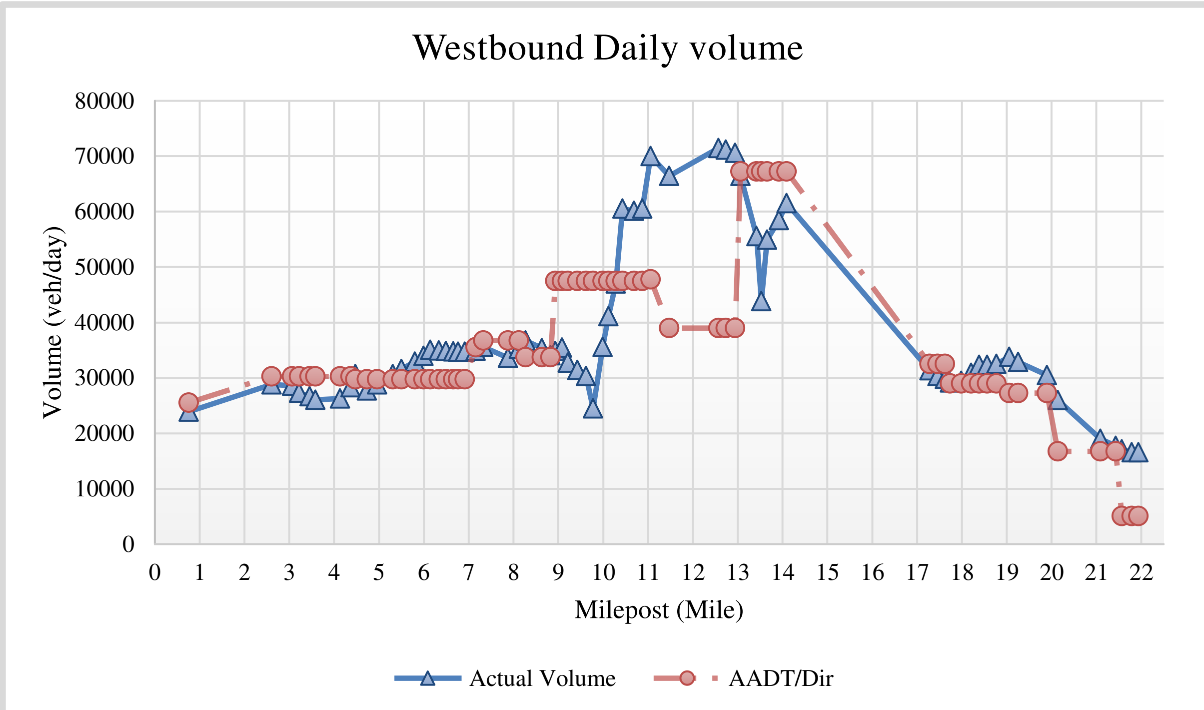


Figure 4-3 Daily volume westbound (2014).

Crashes that have occurred on SR 408 from 2011 till the end of 2013 were compiled for the analysis while the crashes that had occurred in 2014 were used for validation.

4.2.1 Short-term Aggregation Data preparation for Developing SPFs

The filtered traffic and geometric data that were used in the previous chapter with same time periods and intervals are used here to develop several full SPFs with the consideration of using the random effect. But for the SPFs development procedure and to erase the effect of excessive zero observations, some additional procedures have been done to the data before using it in developing the SPFs, so three data sets were prepared. First data set has only segments that have one or more crashes within the analysis period (i.e., excluding safe segments from the data), the second data set includes only segments with two or more crashes during the analysis period, and the last data set has only segments with three or more crashes from 2011 to 2013. Table 4-1 shows the percentage of non-zero observations for different time periods for different datasets. Figure 4-3 shows the percentage of non-zero observations.

The above-mentioned data sets were only used for developing SPFs.

4.2.2 Short-term Aggregation Data preparation for Validation SPFs

In total, 16 data sets were prepared to be used in the validation processes for different time intervals (5, 15, 30, and 60) minutes for both weekdays and weekends for both directions. These data sets were combined with the crash data for 2014 then the whole data that consist all segments (i.e. segments that have crashes and segments that have no crashes) were considered and used in the validation process for 24 models (i.e. 4 time intervals by 4 time periods "morning peak, off-peak, evening peak, and night time" for the weekdays and 4 time intervals by 2 time periods "heavy traffic and low traffic" for the weekends).

Table 4-1: Percentage of Non-Zero observations for different datasets.

Whole Data																								
	AMP				**Off Peak**				**PMP**				**NT**				**HV**				**LV**			
	5	15	30	60	5	15	30	60	5	15	30	60	5	15	30	60	5	15	30	60	5	15	30	60
0	3025	833	285	11	11246	3574	1656	697	3146	954	406	132	21060	6812	3250	1469	22880	7528	3690	1771	16334	5382	2644	1275
Non Zero	263	263	263	263	262	262	262	262	142	142	142	142	312	312	312	312	148	148	148	148	94	94	94	94
Total	3288	1096	548	274	11508	3836	1918	959	3288	1096	548	274	21372	7124	3562	1781	23028	7676	3838	1919	16428	5476	2738	1369
Non Zero%	8%	24%	48%	96%	2%	7%	14%	27%	4%	13%	26%	52%	1%	4%	9%	18%	1%	2%	4%	8%	1%	2%	3%	7%
Segments with One or More Crashes																								
0	1717	464	173	43	7639	2386	1089	457	1567	443	174	49	16869	5439	2590	1174	11402	3707	1786	827	6400	2074	994	458
Non Zero	227	184	151	119	257	246	227	201	137	125	110	93	291	281	270	256	142	141	138	135	92	90	88	83
Total	1944	648	324	162	7896	2632	1316	658	1704	568	284	142	17160	5720	2860	1430	11544	3848	1924	962	6492	2164	1082	541
Non Zero%	12%	28%	47%	73%	3%	9%	17%	31%	8%	22%	39%	65%	2%	5%	9%	18%	1%	4%	7%	14%	1%	4%	8%	15%
Segments with Two or More Crashes																								
0	935	226	71	9	5232	1603	712	283	647	163	54	9	10979	3501	1640	718	6809	2194	1043	469	2651	845	395	174
Non Zero	193	150	117	85	228	217	198	172	97	85	70	53	253	243	232	218	115	114	111	108	61	59	57	52
Total	1128	376	188	94	5460	1820	910	455	744	248	124	62	11232	3744	1872	936	6924	2308	1154	577	2712	904	452	226
Non Zero%	17%	40%	62%	90%	4%	12%	22%	38%	13%	34%	56%	85%	2%	6%	12%	23%	2%	5%	10%	19%	2%	7%	13%	23%
Segments with Three or More Crashes																								
0	583	130	35	2	3018	900	386	142	339	78	24	4	6818	2148	989	418	3132	997	466	201	1316	413	189	80
Non Zero	161	118	89	60	174	164	146	124	69	58	44	30	202	192	181	167	72	71	68	66	40	39	37	33
Total	744	248	124	62	3192	1064	532	266	408	136	68	34	7020	2340	1170	585	3204	1068	534	267	1356	452	226	113
Non Zero%	22%	48%	72%	97%	5%	15%	27%	47%	17%	43%	65%	88%	3%	8%	15%	29%	2%	7%	13%	25%	3%	9%	16%	29%

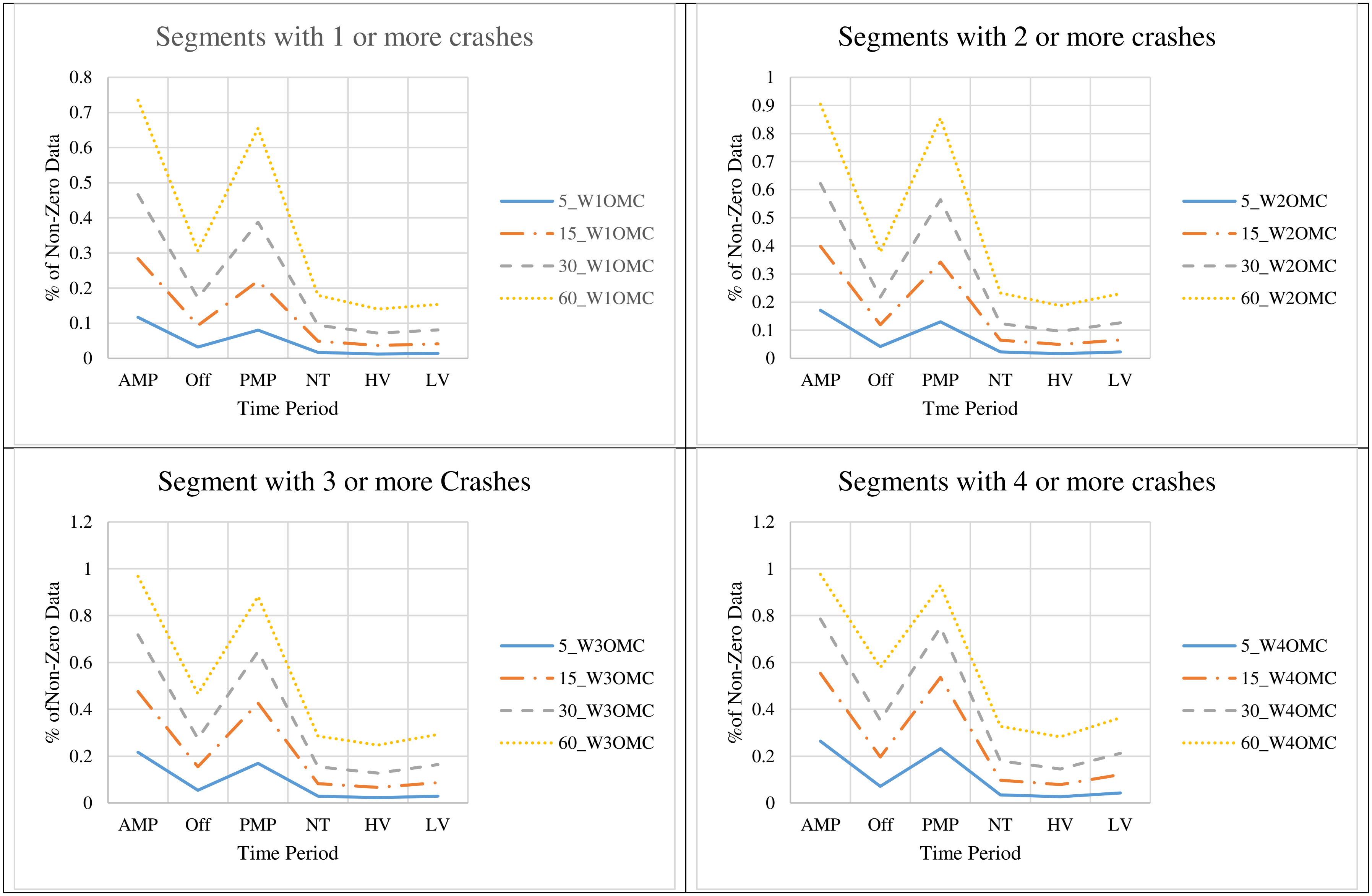


Figure 4-4 Segments with Different Number of Crashes within 3 Years.

4.3 Methodology

Observations are typically assumed to be independent in crash frequency models with a yearly data structure. But, when using short time periods with repeated measures, this assumption is violated. Correlation between those repeated measures may occur and to deal with this issue, random effect models are usually adopted. To reduce the number of zero observations and to account for the correlation and the repetition in the data, several SPFs were developed for different data sets that include only segments with one or more crashes and compare with SPFs for the whole data. The random effect negative binomial model adds an error term ε_i to the Poisson model mean to overcome the assumption of Poisson model (the equal mean and variance assumption) and to account for the over-dispersion as:

$$\lambda_i = Exp(\beta_i x_i + \varepsilon_i + R_i) \tag{4-1}$$

Where λ_i is the Poisson distribution expected number for subject i, β_i is the vector of regression coefficients, x_i is the vector of the explanatory variables, ε_i is the error term which is assumed to be gamma distributed with mean 1 and variance α, and R_i is the random effect which is assumed to be normally distributed with mean 0 and variance σ.

4.4 Modeling Results and Discussion

4.4.1 Modeling results

The modeling results of 96 models for different time intervals and different time periods for both weekdays and weekends for all the prepared data are presented in Tables 4-2 to 4-9. As shown in these tables, 10 independent variables have been used, Log (volume), standard deviation of the volume 'SD(Volume)', average speed '(Speed)', standard deviation of speed 'SD(Speed)', the speed limit 'Speed limit', a dummy variable for auxiliary lane (1 when there is an auxiliary lane), the degree of the curvature when there is a curve ‘Horizontal Curve', percentage of heavy vehicles ‘% of Heavy Vehicle', number of lanes in the section ‘Number of lanes', and a dummy variable for the direction (1 when the direction is eastbound) ‘Direction'. The modeling results show that for different time periods there are different significant variables.

Table 4-2: Modeling Results for different time intervals and periods using the whole data (Weekdays).

	Whole Data															
	Weekday Models															
	Morning Peak				Off Peak				Evening Peak				Night Time			
Parameters	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.
Intercept	1.7029 (0.5139)	0.9452 (0.7411)	-0.5022 (0.8741)	-1.8322 (0.6179)	-7.8737 (0.0003)	-7.9244 (<0.0001)	-7.7266 (0.0024)	-7.3421 (0.0067)	-0.5632 (0.8573)	-0.3504 (0.9182)	0.03419 (0.9925)	0.2581 (0.9475)	-2.2835 (0.1521)	-2.0068 (0.2195)	-1.4371 (0.3836)	-1.0051 (0.5582)
Log(Volume)	0.9257 (0.0029)	1.0788 (0.0006)	1.2161 (0.0003)	1.4421 (<0.0001)	1.2158 (<0.0001)	1.1928 (<0.0001)	1.1538 (<0.0001)	1.1028 (<0.0001)	0.7952 (0.0089)	0.7722 (0.0116)	0.7200 (0.0216)	0.7528 (0.0167)	0.5054 (<0.0001)	0.5072 (<0.0001)	0.5029 (<0.0001)	0.4720 (<0.0001)
SD(Volume)	-	-	-	-	-	-	-	-	-	-	-	-	-			
Speed	-0.05489 (0.0017)	-0.05175 (0.0031)	-0.04355 (0.0190)	-0.04900 (0.0457)	-	-	-	-	-0.1168 (0.0014)	-0.1142 (0.0020)	-0.1120 (0.0032)	-0.1168 (0.0033)	-0.06145 (0.0143)	-0.05736 (0.0241)	-0.06061 (0.0178)	-0.05879 (0.0269)
SD(Speed)	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-
Speed limit	-0.07232 (0.0029)	-0.07381 (0.0018)	-0.07525 (0.0015)	-0.07368 (0.0024)	-0.03406 (0.0998)	-0.03477 (0.0926)	-0.03568 (0.0842)	-0.03728 (0.0711)	-	-	-	-	-	-	-	-
Auxiliary Lane	-	-	-	-	-	-	-	-	-	-	-	-	0.2792 (0.0681)	0.2783 (0.0676)	0.2845 (0.0619)	0.2786 (0.0684)
Horizontal Curve	-0.3159 (0.0231)	-0.3030 (0.0260)	-0.2905 (0.0326)	-0.2898 (0.0328)	-	-	-	-	-	-	-	-	-	-	-	-
% of Heavy Vehicle	-	-	-	-	-	-	-	-	0.03242 (0.0788)	0.03239 (0.0793)	0.03316 (0.0779)	0.03503 (0.0615)	-	-	-	-
Number of lanes	-0.3712 (0.0053)	-0.4542 (0.0006)	-0.4567 (0.0006)	-0.5814 (<0.0001)	-	-	-	-	-	-	-	-	-	-	-	-
Direction	-0.9533 (<0.0001)	-0.9682 (<0.0001)	-0.9853 (<0.0001)	-0.9926 (<0.0001)	-0.03922 (0.8346)	-0.03944 (0.8332)	-0.03866 (0.8359)	-0.03973 (0.8309)	-0.3622 (0.1436)	-0.3532 (0.1529)	-0.3466 (0.1669)	-0.3439 (0.1665)	-0.1400 (0.3251)	-0.1382 (0.3281)	-0.1403 (0.3209)	-0.1371 (0.3343)
SD	-0.8056 (<0.0001)	-0.7628 (<0.0001)	0.7384 (<0.0001)	-0.7405 (<0.0001)	-0.6835 (<0.0001)	-0.6779 (<0.0001)	-0.6637 (<0.0001)	0.6562 (<0.0001)	-0.7373 (<0.0001)	-0.7361 (<0.0001)	0.7665 (<0.0001)	-0.7276 (<0.0001)	-0.3702 (0.0006)	-0.3532 (0.0013)	0.3513 (0.0016)	0.3726 (0.0005)
Gamma	7.2991 (0.8786)	12.5625 (0.9048)	3.6709 (0.0003)	14.1649 (0.8826)	11.7657 (0.9445)	3.7549 4 (0.1089)	2.8919 (0.0022)	3.2826 (0.0039)	13.1778 (0.9211)	14.9669 (0.8027)	13.4135 (0.8389)	4.4573 (0.2142)	-0.2494 (0.4390)	0.7020 (0.0484)	1.3637 (0.0012)	2.3638 (0.0005)
-2 Log Likelihood	2560.16	1761.6	1341.12	955.68	3804.32	2912.8	2379.2	1879.52	1760.8	1296.64	1031.36	788	4909.6	3862.08	3214.88	2601.76
AIC	2588.96	1790.4	1369.92	984.48	3823.52	2932	2398.4	1898.24	1783.2	1319.04	1053.76	810.4	4932	3884.48	3237.28	2624.16
BIC	2631.04	1832.48	1412	1026.56	3851.52	2960	2426.4	1926.24	1815.84	1351.84	1086.56	843.04	4964.64	3917.12	3269.92	2656.8

Table 4-3: Modeling Results for different time intervals and periods using the whole data (Weekends).

	Whole Data							
	Weekend Models							
	High Volume				Low Volume			
Parameters	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**
Intercept	0.8689 (0.7223)	2.1713 (0.3683)	2.3898 (0.3250)	3.2316 (0.1842)	-2.4257 (0.4012)	-2.8452 (0.3314)	-3.3862 (0.2639)	-3.1189 (0.3277)
Log(Volume)	-	-	-	-	1.0607 (0.0003)	1.0791 (0.0001)	1.1234 (<0.0001)	1.0255 (0.0004)
SD(Volume)	-	-	-	-	-0.06965 (0.0030)	-0.02329 (0.0020)	-0.01195 (0.0015)	-0.00563 (0.0035)
Speed	-0.09302 (0.0153)	-0.09579 (0.0114)	-0.08835 (0.0199)	-0.09064 (0.0171)	-0.09256 (0.0390)	-0.08914 (0.0406)	-0.08544 (0.0499)	-0.08268 (0.0600)
Auxiliary Lane	0.6822 (0.0066)	0.6882 (0.0049)	0.6816 (0.0054)	0.6844 (0.0052)	-	-	-	-
Horizontal Curve	-0.2297 (0.0854)	-0.2278 (0.0824)	-0.2255 (0.0858)	-0.2254 (0.0856)	-	-	-	-
Direction	-0.2992 (0.2133)	-0.2997 (0.2012)	-0.2957 (0.2074)	-0.2960 (0.2067)	0.2296 (0.4103)	0.2208 (0.4095)	0.2143 (0.4230)	0.2440 (0.3634)
SD	-0.7996 (<0.0001)	-0.7512 (<0.0001)	-0.7468 (<0.0001)	-0.7433 (<0.0001)	-0.7483 (0.0001)	-0.6431 (0.0004)	-0.6368 (0.0006)	-0.6027 (0.0022)
Gamma	-0.4001 (0.3993)	0.5786 (0.2420)	1.0762 (0.0354)	1.7045 (0.0026)	0.2904 (0.7502)	1.1377 (0.2903)	1.7802 (0.1712)	1.4200 (0.0791)
-2 Log Likelihood	2739.68	2238.4	1924.48	1609.6	1747.52	1425.6	1222.72	1029.28
AIC	2762.08	2260.8	1946.88	1632	1769.92	1448	1245.12	1051.68
BIC	2794.72	2293.44	1979.52	1664.64	1802.72	1480.8	1277.92	1084.48

Table 4-4: Modeling Results for different time intervals and periods using the data that has one or more crashes (Weekdays).

	Segments with One or More Crashes Data Set															
	Weekday Models															
	Morning Peak				Off Peak				Evening Peak				Night Time			
Parameters	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.
Intercept	2.1187 (0.3852)	1.9125 (0.4850)	1.0931 (0.7193)	0.1052 (0.1941)	-8.2640 (<0.0001)	-8.8737 (<0.0001)	-9.0809 (<0.0001)	-8.4588 (0.0017)	1.1252 (0.6798)	1.6281 (0.5784)	2.0522 (0.5082)	2.3685 (0.4690)	0.2964 (0.8516)	0.7552 (0.6428)	1.2498 (0.4480)	1.6603 (0.3335)
Log(Volume)	0.5486 (0.0709)	0.6807 (0.0289)	0.7831 (0.0168)	0.9466 (0.0088)	1.0514 (0.0019)	1.1649 (0.0007)	1.1721 (0.0009)	1.0552 (0.0059)	0.4829 (0.0617)	0.4618 (0.0740)	0.4181 (0.0986)	0.4438 (0.0912)	0.5304 (<0.0001)	0.5362 (<0.0001)	0.5287 (<0.0001)	0.4956 (<0.0001)
Speed	-0.04580 (0.0056)	-0.04411 (0.0084)	-0.03673 (0.0367)	-0.03443 (0.0948)	-	-	-	-	-0.09907 (0.0042)	-0.09591 (0.0057)	-0.09163 (0.0096)	-0.09353 (0.0093)	-0.09471 (0.0002)	-0.09361 (0.0002)	-0.09631 (0.0002)	-0.09494 (0.0004)
SD(Speed)	-		-	-	-0.2779 (0.0134)	-0.2887 (0.0131)	-0.2672 (0.0222)	-0.2649 (0.0254)	-	-	-	-	-	-	-	-
SD(Occupancy)	-	-	-	-	0.03125 (0.0016)	0.01082 (0.0019)	0.005236 (0.0053)	0.002955 (0.0095)	-	-	-	-	-	-	-	-
Speed limit	-0.05477 (0.0176)	-0.05564 (0.0154)	-0.05820 (0.0115)	-0.05956 (0.0123)	-	-	-	-	-	-	-	-	-	-	-	-
Auxiliary Lane	-	-	-	-	-	-	-	-	-	-	-	-	0.3070 (0.0293)	0.3066 (0.0297)	0.3084 (0.0289)	0.2969 (0.0357)
Horizontal Curve	-	-	-	-	0.1755 (0.0777)	0.1772 (0.0716)	0.1774 (0.0711)	0.1735 (0.0777)	-	-	-	-	-	-	-	-
% of Heavy Vehicle	-	-	-	-	0.02841 (0.0249)	0.02832 (0.0239)	0.02923 (0.0199)	0.02976 (0.0181)	-	-	-	-	-	-	-	-
Number of lanes	-0.2708 (0.0324)	-0.3427 (0.0068)	-0.3406 (0.0080)	-0.4124 (0.0031)	-0.2642 (0.0161)	-0.3105 (0.0050)	-0.3090 (0.0053)	-0.2826 (0.0127)	-	-	-	-	-0.1306 (0.0403)	-0.1444 (0.0241)	-0.1325 (0.0367)	-0.1155 (0.0710)
Direction	-0.7490 (0.0012)	-0.7550 (0.0010)	-0.7661 (0.0008)	-0.7862 (0.0007)	-0.09926 (0.5502)	-0.09783 (0.5506)	-0.09822 (0.5489)	-0.09288 (0.5733)	-0.1592 (0.4775)	-0.1468 (0.5116)	-0.1268 (0.5710)	-0.1343 (0.5484)	-0.1043 (0.4141)	-0.1061 (0.4066)	-0.1055 (0.4100)	-0.1042 (0.4183)
SD	0.6480 (<0.0001)	0.6365 (<0.0001)	0.6148 (<0.0001)	0.6282 (<.0001)	-0.4348 (<0.0001)	0.4171 (<0.0001)	0.4060 (<0.0001)	-0.4110 (<0.0001)	-0.4093 (0.0002)	-0.4085 (0.0002)	0.4126 (0.0002)	0.4079 (<0.0002)	-0.1876 (0.1511)	-0.1832 (0.1642)	0.1894 (0.1406)	0.2203 (0.0556)
Gamma	8.2933 (0.9458)	13.4331 (0.8810)	3.8951 (0.0006)	15.6325 (0.7626)	12.4570 (0.9420)	4.5592 (0.3322)	3.3219 (0.0086)	4.2806 (0.0880)	14.3850 (0.7639)	14.2828 (0.8931)	14.3924 (0.8794)	15.2658 (0.9444)	-0.1178 (0.7200)	0.8651 (0.0208)	1.6105 (0.0007)	2.7772 (0.0016)
-2 Log Likelihood	2428.16	1627.84	1208.16	824.64	3620.8	2727.36	2196.96	1699.04	1568.64	1269.78	839.2	596.16	4775.84	3727.04	3080.64	2468.32
AIC	2453.92	1653.44	1233.76	850.24	3652.8	2759.36	2228.96	1731.04	1587.84	1291.86	858.4	615.36	4801.44	3752.64	3106.24	2493.92
BIC	2484.32	1684.16	1264.32	880.8	3693.6	2800	2269.6	1771.84	1609.6	1316.75	880.16	637.12	4836	3787.2	3140.8	2528.48

Table 4-5: Modeling Results for different time intervals and periods using the data that has one or more crashes (Weekends).

	Segments with One or More Crashes Data Set							
	Weekend Models							
	High Volume				Low Volume			
Parameters	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.
Intercept	2.6075 (0.2519)	3.8681 (0.0907)	4.1455 (0.0720)	4.9873 (0.0311)	-3.9874 (0.1734)	-4.2421 (0.1667)	-4.7146 (0.1365)	-4.2512 (0.2022)
Log(Volume)	-	-	-	-	1.0361 (0.0003)	1.0398 (0.0002)	1.0760 (0.0001)	0.9865 (0.0005)
SD(Volume)	-	-	-	-	-0.06459 (0.0046)	-0.02145 (0.0033)	-0.01095 (0.0026)	-0.00516 (0.0051)
Speed	-0.1176 (0.0014)	-0.1198 (0.0012)	-0.1131 (0.0022)	-0.1157 (0.0017)	-0.1074 (0.0072)	-0.1070 (0.0077)	-0.1041 (0.0095)	-0.1037 (0.0114)
Speed limit	-	-	-	-	0.05336 (0.0349)	0.05452 (0.0310)	0.05529 (0.0280)	0.05400 (0.0347)
Auxiliary Lane	0.5885 (0.0072)	0.5944 (0.0063)	0.5906 (0.0068)	0.5925 (0.0066)	0.7308 (0.0038)	0.7299 (0.0038)	0.7198 (0.0042)	0.7359 (0.0042)
Horizontal Curve	-	-	-	-	0.2214 (0.0560)	0.2248 (0.0523)	0.2268 (0.0488)	0.2272 (0.0549)
% of Heavy Vehicle	0.03776 (0.0600)	0.03595 (0.0733)	0.03471 (0.0866)	0.03607 (0.0748)	-	-	-	-
Direction	0.09382 (0.6607)	0.08923 (0.6742)	0.08525 (0.6881)	0.08883 (0.6756)	-0.09788 (0.6741)	-0.1120 (0.6308)	-0.1189 (0.6086)	-0.09686 (0.6809)
SD	-0.4394 (0.0008)	-0.4275 (0.0012)	0.4242 (0.0014)	0.4226 (0.0016)	0.4325 (0.0015)	0.4521 (0.0125)	0.4784 (0.0158)	0.4158 (0.0114)
Gamma	-0.3222 (0.4953)	0.7242 (0.1473)	1.2576 (0.0175)	1.9426 (0.0014)	0.4530 (0.6227)	1.6777 (0.2082)	2.7916 (0.2514)	2.4352 (0.0709)
-2 Log Likelihood	2553.6	2050.24	1737.44	1422.24	1563.84	1239.52	1036.64	845.12
AIC	2576	2072.64	1759.84	1444.64	1595.84	1271.52	1068.64	877.12
BIC	2600.8	2097.6	1784.64	1469.44	1627.36	1303.04	1100.16	908.64

Table 4-6: Modeling Results for different time intervals and periods using the data that has two or more crashes (Weekdays).

	Segments with Two or More Crashes Data Set															
	Weekday Models															
	Morning Peak				Off Peak				Evening Peak				Night Time			
Parameters	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**
Intercept	-0.9599 (0.6856)	-1.6840 (0.5305)	-2.5300 (0. 4032)	-4.3502 (0. 2180)	-6.6701 (0.0002)	-7.1038 (0.0012)	-7.4843 (0.0031)	-6.5180 (0.0334)	0.7437 (0.8518)	1.1564 (0.7849)	1.3083 (0.7659)	0.8303 (0.8580)	0.1770 (0.9152)	0.6848 (0.6894)	1.2022 (0.4890)	1.6195 (0.3686)
Log(Volume)	0.8439 (0.0233)	1.0584 (0.0055)	1. 1813 (0. 0037)	1.4660 (0. 0015)	0.8520 (0.0219)	0.9792 (0.0099)	1.0317 (0.0092)	0.8701 (0.0463)	0.8026 (0.0309)	0.7923 (0.0340)	0.6879 (0.0656)	0.7521 (0.0490)	0.5544 (<0.0001)	0.5608 (<0.0001)	0.5502 (<0.0001)	0.5237 (<0.0001)
Speed									-0.1834 (0.0005)	-0.1857 (0.0005)	-0.1711 (0.0014)	-0.1690 (0.0019)	-0.09495 (0.0003)	-0.09488 (0.0005)	-0.09793 (0.0003)	-0.09751 (0.0006)
SD(Speed)					-0.4594 (0.0039)	-0.5234 (0.0016)	-0.4864 (0.0052)	-0.4698 (0.0102)								
SD(Occupancy)					0.04641 (0.0004)	0.01735 (0.0002)	0.008353 (0.0010)	0.004477 (0.0025)								
Speed limit	-0.06865 (0.0083)	-0.07119 (0.0055)	-0.07179 (0.0054)	-0.07373 (0.0033)					0.05381 (0.0699)	0.05404 (0.0693)	0.05144 (0.0815)	0.05214 (0.0786)				
Auxiliary Lane									0.8663 (0.0051)	0.8715 (0.0049)	0.8440 (0.0062)	0.8376 (0.0068)	0.4585 (0.0030)	0.4601 (0.0029)	0.4568 (0.0033)	0.4430 (0.0042)
Horizontal Curve					0.2109 (0.0370)	0.2064 (0.0383)	0.2092 (0.0361)	0.2016 (0.0453)	0.4990 (0.0263)	0.4944 (0.0277)	0.4799 (0.0316)	0.4978 (0.0267)	0.1622 (0.0904)	0.1634 (0.0888)	0.1613 (0.0940)	0.1578 (0.0978)
% of Heavy Vehicle					0.02946 (0.0181)	0.02930 (0.0176)	0.02957 (0.0169)	0.03018 (0.0159)	0.03454 (0.0977)	0.03455 (0.1083)	0.03591 (0.0925)	0.03620 (0.0922)	0.01805 (0.0525)	0.01744 (0.0636)	0.01651 (0.0801)	0.01773 (0.0573)
Number of lanes	-0.3413 (0.0369)	-0. 4679 (0. 0042)	-0. 4935 (0. 0036)	-0. 6451 (0. 0005)	-0.3507 (0.0025)	-0.4054 (0.0006)	-0.4116 (0.0006)	-0.3638 (0.0029)					-0.1601 (0.0181)	-0.1759 (0.0100)	-0.1602 (0.0176)	-0.1467 (0.0300)
Direction	-0.7729 (0.0082)	-0. 7655 (0. 0079)	-0. 7466 (0. 0095)	-0. 7696 (0. 0067)	-0.1600 (0.3395)	-0.1531 (0.3530)	-0.1508 (0.3588)	-0.1453 (0.3830)	-0.4652 (0.1339)	-0.4655 (0.1350)	-0.3806 (0.2192)	-0.3977 (0.2040)	-0.1490 (0.2712)	-0.1514 (0.2649)	-0.1511 (0.2687)	-0.1530 (0.2620)
SD	0.6548 (<0.0001)	0. 6360 (<0.0001)	0. 6182 (<0.0001)	-0. 6060 (<0.0001)	0.3196 (0.0007)	0.2983 (0.0019)	-0.2796 (0.0071)	-0.2913 (0.0060)	0.2746 (0.1085)	-0.2746 (0.1096)	0.2677 (0.1209)	0.2703 (0.1178)	-0.1248 (0.4840)	0.1234 (0.4902)	-0.1401 (0.3820)	0.1717 (0.1997)
Gamma	5.4862 (0.5011)	11.9523 (0.9890)	3. 7517 (0.0003)	16.2117 (0. 7528)	13.0535 (0.8474)	4.2428 (0.2285)	3.2174 (0.0063)	3.9799 (0.0413)	14.6192 (0.8226)	13.5650 (0.9018)	13.9891 (0.8773)	13.6532 (0.9043)	-0.1375 (0.6756)	0.8772 (0.0215)	1.6485 (0.0009)	2.9046 (0.0037)
-2 Log Likelihood	1930.08	1247.04	897.92	581.44	3054.08	2258.56	1793.76	1363.04	971.84	646.4	474.4	320.8	3998.56	3082.72	2520.64	1988.64
AIC	1952.48	1269.44	920.32	603.84	3086.08	2290.56	1825.76	1395.04	1003.84	678.4	506.4	352.8	4030.56	3114.72	2552.64	2020.64
BIC	1973.12	1290.08	940.96	624.48	3120.96	2325.28	1860.64	1429.92	1026.72	701.28	529.44	375.68	4067.04	3151.04	2589.12	2057.12

Table 4-7: Modeling Results for different time intervals and periods using the data that has two or more crashes (Weekends).

	Segments with Two or More Crashes Data Set							
	Weekend Models							
	High Volume				Low Volume			
Parameters	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.
Intercept	2.3023 (0.3922)	3.7331 (0.1667)	3.9071 (0.1525)	4.7872 (0.0807)	-8.0663 (<0.0001)	-8.2493 (<0.0001)	-8.3794 (<0.0001)	-7.8316 (0.0005)
Log(Volume)					1.2972 (0.0012)	1.2277 (0.0013)	1.2084 (0.0014)	1.0817 (0.0045)
SD(Volume)					-0.07335 (0.0088)	-0.02256 (0.0101)	-0.01071 (0.0123)	- 0.00498 (0.0222)
Speed	-0.1106 (0.0105)	-0.1153 (0.0077)	-0.1072 (0.0135)	-0.1105 (0.0109)				
SD(Speed)								
SD(Occupancy)								
Speed limit								
Auxiliary Lane	0.8124 (0.0020)	0.8144 (0.0019)	0.8153 (0.0020)	0.8176 (0.0019)				
Horizontal Curve					0.5771 (0.0180)	0.5823 (0.0170)	0.5838 (0.0160)	0.5807 (0.0217)
% of Heavy Vehicle	0.04177 (0.0696)	0.03875 (0.0916)	0.03900 (0.0931)	0.04098 (0.0778)				
Number of lanes								
Direction	-0.04530 (0.8550)	-0.05246 (0.8315)	-0.05390 (0.8274)	-0.04772 (0.8469)	0.1950 (0.5196)	0.1880 (0.5344)	0.1836 (0.5418)	0.2273 (0.4637)
SD	0.4268 (0.0044)	0.4185 (0.0056)	0.4166 (0.0063)	0.4148 (0.0069)	0.5214 (0.0214)	0.4245 (0.0254)	0.4215 (0.0031)	0.3215 (0.0025)
Gamma	-0.3374 (0.4789)	0.7128 (0.1590)	1.2412 (0.0215)	1.9231 (0.0022)	0.4645 (0.6213)	1.6291 (0.2244)	2.6115 (0.2362)	2.1450 (0.0703)
-2 Log Likelihood	1985.6	1576.96	1323.36	1067.84	952.32	739.68	607.36	484.64
AIC	2008	1599.36	1345.76	1090.24	974.72	762.08	629.76	507.04
BIC	2027.2	1618.56	1364.96	1109.44	987.04	774.4	642.08	519.36

Table 4-8: Modeling Results for different time intervals and periods using the data that has three or more crashes (Weekdays).

	Segments with Three or More Crashes Data Set															
	Weekday Models															
	Morning Peak				Off Peak				Evening Peak				Night Time			
Parameters	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**	**5 Min.**	**15 Min.**	**30 Min.**	**60 Min.**
Intercept	-1.0634 (0.7138)	-2.9532 (0.3675)	-3.8757 (0.2964)	-6.9811 (0.1211)	-4.9268 (0.0608)	-5.6051 (0.0614)	-5.9499 (0.0732)	-5.4804 (0.1468)	4.5002 (0.3250)	4.6519 (0.3421)	4.0552 (0.4109)	3.1778 (0.5378)	0.9761 (0.6223)	1.7519 (0.3856)	2.4041 (0.2294)	3.9813 (0.0576)
Log(Volume)	0.9055 (0.0561)	1.3000 (0.0085)	1.4631 (0.0063)	1.8680 (0.0026)	1.0702 (0.0105)	1.2224 (0.0044)	1.2612 (0.0044)	1.1648 (0.0122)	1.0200 (0.0304)	0.9849 (0.0371)	0.8816 (0.0583)	0.9576 (0.0457)	0.5547 (<0.0001)	0.5065 (<0.0001)	0.4646 (<0.0001)	0.4124 (<0.0001)
Speed									-0.1957 (0.0056)	-0.1948 (0.0066)	-0.1729 (0.0124)	-0.1672 (0.0164)	-0.1049 (0.0010)	-0.1057 (0.0012)	-0.1086 (0.0008)	-0.1238 (0.0003)
SD(Speed)													0.1119 (0.0523)	0.1668 (0.0229)	0.2046 (0.0073)	0.2794 (0.0020)
Speed limit	-0.07157 (0.0185)	-0.07205 (0.0144)	-0.07675 (0.0096)	-0.07433 (0.0089)	-0.3430 (0.0986)	-0.03651 (0.0699)	-0.03657 (0.0688)	-0.03718 (0.0692)								
Auxiliary Lane													0.6119 (0.0007)	0.6330 (0.0005)	0.6415 (0.0004)	0.6488 (0.0005)
Horizontal Curve					0.3357 (0.0147)	0.3439 (0.0098)	0.3452 (0.0096)	0.3375 (0.0119)					0.2157 (0.0780)	0.2237 (0.0719)	0.2298 (0.0635)	0.2368 (0.0503)
% of Heavy Vehicle	0.03624 (0.0582)	0.03660 (0.0497)	0.03328 (0.0807)	0.03205 (0.0814)												
Number of lanes	-0.4832 (0.0273)	-0.6916 (0.0018)	-0.7497 (0.0014)	-0.9390 (0.0002)	-0.4240 (0.0017)	-0.4909 (0.0004)	-0.5148 (0.0003)	-0.4734 (0.0011)					-0.1967 (0.0103)	-0.2076 (0.0074)	-0.1827 (0.0151)	-0.1714 (0.0206)
Direction	-0.8212 (0.0277)	-0.8787 (0.0158)	-0.8938 (0.0141)	-0.9952 (0.0052)	-0.2738 (0.1776)	-0.2683 (0.1700)	-0.2624 (0.1790)	-0.2778 (0.1599)	-0.4635 (0.2699)	-0.4468 (0.2889)	-0.3568 (0.3829)	-0.3686 (0.3718)	-0.2153 (0.1578)	-0.2172 (0.1581)	-0.2133 (0.1624)	-0.2336 (0.1218)
SD	0.5658 (<0.0001)	0.5316 (<0.0001)	0.5101 (0.0002)	0.4914 (<0.0001)	0.3110 (0.0052)	-0.2714 (0.0226)	-0.2526 (0.0468)	-0.2668 (0.0310)	-0.4940 (0.0129)	0.4977 (0.0131)	0.4687 (0.0134)	0.4560 (0.0222)	0.03251 (0.7523)	0.02155 (0.5641)	0.05368 (0.6874)	0.08278 (0.7896)
Gamma	4.6261 (0.2107)	4.8851 (0.1482)	3.5572 (0.0003)	14.5583 (0.8794)	11.6807 (0.9432)	4.2735 (0.2602)	3.2303 (0.0115)	4.0390 (0.0656)	13.7659 (0.8693)	14.2369 (0.8864)	13.9651 (0.9409)	5.0376 (0.5202)	0.06167 (0.8692)	1.0016 (0.0167)	1.8599 (0.0019)	3.7213 (0.0904)
-2 Log Likelihood	1529.12	955.84	668.8	417.92	2219.2	1610.24	1254.56	931.84	646.4	418.24	305.44	207.68	3045.12	2307.68	1857.44	1438.56
AIC	1554.72	981.44	694.4	443.52	2244.8	1635.84	1280.16	957.44	665.6	437.44	324.64	226.88	3077.12	2339.68	1889.44	1470.56
BIC	1573.12	999.84	712.64	461.76	2265.76	1636	1301.12	978.4	673.6	445.44	332.64	234.72	3106.08	2368.48	1918.4	1499.52

Table 4-9: Modeling Results for different time intervals and periods using the data that has three or more crashes (Weekends).

	Segments with Three or More Crashes Data Set							
	Weekend Models							
	High Volume				Low Volume			
Parameters	5 Min.	15 Min.	30 Min.	60 Min.	5 Min.	15 Min.	30 Min.	60 Min.
Intercept	3.9076 (0.3453)	5.1731 (0.2237)	5.0574 (0.2473)	5.4489 (0.2246)	2.2030 (0.5623)	2.4969 (0.5195)	2.3900 (0.5452)	2.9094 (0.4876)
Log(Volume)					0.5965 (0.0540)	0.6118 (0.0503)	0.6503 (0.0423)	0.6024 (0.0643)
Speed	-0.1327 (0.0473)	-0.1368 (0.0449)	-0.1238 (0.0739)	-0.1237 (0.0798)	-0.1311 (0.0379)	-0.1299 (0.0376)	-0.1272 (0.0404)	-0.1272 (0.0463)
SD(Speed)	0.3031 (0.0399)	0.4005 (0.0204)	0.4294 (0.0512)	0.6419 (0.0172)				
Auxiliary Lane	0.8075 (0.0400)	0.7658 (0.0554)	0.7495 (0.0658)	0.6827 (0.0989)				
Direction	-0.3443 (0.3620)	-0.3292 (0.3935)	-0.3347 (0.3961)	-0.2657 (0.5152)	-0.03619 (0.9360)	-0.04407 (0.9218)	-0.07008 (0.8764)	-0.02456 (0.9574)
SD	0.5119 (0.0048)	0.5370 (0.0041)	0.5530 (0.0036)	0.5963 (0.0027)	0.3548 (0.2587)	0.3144 (0.3651)	0.2987 (0.2145)	0.3478 (0.3255)
Gamma	-0.2282 (0.6577)	0.8390 (0.1317)	1.3117 (0.0290)	2.1726 (0.0048)	1.0236 (0.4973)	2.6944 (0.4210)	3.6040 (0.5544)	2.8326 (0.2675)
-2 Log Likelihood	1176.96	916.96	762.72	599.2	591.2	450.56	365.12	286.4
AIC	1199.36	939.36	785.12	621.6	610.4	469.76	384.32	305.6
BIC	1210.08	950.08	795.68	632.32	614.08	473.6	388	309.44

4.4.2 Performance Measure for the Road

To evaluate the suggested approach, Potential for Safety Improvement (PSI) with Empirical Bays (EB) adjustment has been used as a performance measure for the new approach. This performance measure has been explained in Chapter 3.

4.4.3 Comparison Between the Original Data Based PSI and Suggested Data Based PSI

Tables 4-10 show the results of two different goodness of fit. The table presents the Mean Absolute Deviation (MAD) (i.e., equation 6 in chapter 3) and the Mean Square Prediction Error (MSPE) (i.e., equation 7 in chapter 3). For the weekdays, the results show that using the data that has segments with one or more crashes in the modeling process and the 15 minutes time interval gives the best results in term of both MAD and MSPE. While for the weekends, the results show using data that has one or more crashes and the 60 minutes time interval give the best results in term of

MSPE, while using the same time interval with the original data gives the best results in term of MAD.

Table 4-10: The Results of Two Different Goodness of Fit.

	Weekdays															
	5 minutes				**15 minutes**				**30 minutes**				**60 minutes**			
	Original	**>=1**	**>=2**	**>=3**	**Original**	**>=1**	**>=2**	**>=3**	**Original**	**>=1**	**>=2**	**>=3**	**Original**	**>=1**	**>=2**	**>=3**
MAD	1.742	1.743	2.389	3.066	1.632	1.603	2.684	3.086	1.751	2.092	2.666	3.118	1.774	2.101	2.682	3.271
MSPE	10.909	5.030	8.734	16.933	8.791	3.819	14.110	17.266	10.892	10.931	13.982	18.388	10.935	11.042	14.260	24.439
	Weekends															
	5 minutes				**15 minutes**				**30 minutes**				**60 minutes**			
	Original	**>=1**	**>=2**	**>=3**	**Original**	**>=1**	**>=2**	**>=3**	**Original**	**>=1**	**>=2**	**>=3**	**Original**	**>=1**	**>=2**	**>=3**
MAD	0.639	0.766	1.121	1.376	0.637	0.787	1.252	1.456	0.636	0.774	1.252	1.498	0.635	0.765	1.258	1.839
MSPE	1.025	0.814	1.784	3.342	1.014	0.878	3.272	4.074	1.011	0.841	3.257	4.535	1.006	0.812	3.339	9.252

CHAPTER 5: DUAL-STATE MODELS VS SINGLE-STATE MODELS

5.1 Introduction

As stated in the previous chapters, as the adopted time intervals get smaller, two major methodological challenges arise (i.e., more zero observations and repeated measures problems). Two relaxations methods have been proposed for the single state count models (Poisson and negative binomial) to account for the excessive zero observations. The first proposed approach is the zero-inflated model. Several transportation safety studies have been conducted using this approach (Shankar et al., 1997; Chin and Quddus, 2003). The second approach is the Hurdle model. The two models differ in their implementation to deal with the excessive zero observations. On the other hand, to account for the unobserved heterogeneity, the random parameter has been proposed. However, for transportation safety studies, no study has been tested whether the random parameters have to be used in the first part or the second part of the zero-inflated models and in case of using two random parameters in both parts, a correlation between them should be considered or not. In this chapter, the most common models with different random parameters implemented are adopted. Overall, 32 models are developed (NB/Hurdle models with random effects in the count part, NB/Hurdle models with random effects in the count part, NB/Hurdle models with random effects with no correlation between them in two parts, and NB/Hurdle models with random effects with a correlation between them in two parts) for different time periods. The best models have been suggested based on the Bayesian information criterion (BIC).

5.2 Methodology

The term Zero inflation is referring to a data which has a higher number of zero observations than what is expected from a standard Poisson distribution and often results to an over-dispersion. When analyzing any count data, the starting point is usually Poisson distribution. However, when the data is zero-inflated, this distribution cannot be used since its conditional mean cannot vary independently with its corresponding variance, and this may lead to a biased estimate. Other models have been developed to overcome these problems such as negative binomial, hurdle, and zero-inflated models. Negative binomial can account for the over-dispersion, zero-inflated Poisson and zero-inflated hurdle models can account for excesses zeros but not the over-dispersion whereas the zero-inflated negative binomial and hurdle models have shown more reliable ability to account for both excess zeros and over-dispersion issues.

To account for the correlations between repeated measures, these models were extended by including random effects (Hall, 2000; Yau et al. 2004) There are different suggested ways to extend these models with random effects. One way is to add random effects to the second part of the zero-inflated Poisson/negative binomial models. Another way which has been proven to be more efficient is to add a pair of uncorrelated or correlated random effects for both parts of the model (Min and Agresti, 2005; Yau et al. 2004). Models for zero-inflated count data: to account for the zero-inflated observation, different models have been used such as negative binomial, zero-inflated Poisson, zero-inflated negative binomial, hurdle Poisson, and hurdle negative binomial. However, in the current version of SAS 9.4, there are no straightforward methods to fit these models. One of the most commonly used methods for integral approximation of maximum likelihood is the adaptive Gaussian Quadrature center at the conditional mode of the random effects. This method with one quadrature point is equivalent to Laplace approximation (Aitkin, 1999; Pinherio and Bates, 1995). As the number of quadrature increase the accuracy of the estimation increase and the computational time increase. Literature for the adaptive Gaussian Quadrature suggested at least five quadrature points to increase the accuracy of the estimation (Liu and Pierce, 1994; Rabe-Hesheth et al., 2002).

The variance of the zero-inflated Poisson or the zero-inflated negative binomial models can never exceed the mean, and for this reason, they cannot accommodate for under-dispersion, while hurdle model can account for both under-dispersion and over-dispersion.

In this study, SAS 9.4 has been used to develop the suggested models. Adding random effects to zero-inflated or hurdle models make the statistical modeling process more complex than the commonly used SAS procedures. A procedure that can implement random effects in the modeling process in SAS is the PROC NLMIXE; however, this procedure needs to specify the initial values for the coefficient to reduce the modeling process time and to avoid overflow and arithmetic exceptions in the process of computation of the objective function and its derivatives. To get the initial values to be used in the zero-inflated negative binomial and Hurdle modeling procedure, Generalized Linear Mixed Model (GLMM) has been implemented first, and the coefficients from this model were considered the initial values for zero-inflated negative binomial and Hurdle models. Since the random effects variance may be hard to be approximate, grid search has been used to choose the optimal value based on a range of values. The range was set from 0.1 to 20.1 by 1. The number of the quadrature points has been set to 20 for all models.

5.3 Data preparation

The whole filtered traffic data that has been prepared in the previous chapter and the geometric data has been used here. Crashes that occurred on SR408 in the period of 2011-2015 were compiled with the traffic and geometric data for the analyses.

5.4 Models for Zero Inflated Count Data

Two well-known zero inflated count data models have been selected and used with different implementation of random effects.

5.4.1 Zero Inflated Negative Binomial

The zero-inflated negative binomial model consists of two parts, the first part is the zero part and the second part is the normal negative binomial models conditional on the first part. The zero-inflated negative binomial model can be considered as an extension of the traditional negative binomial as:

$$y_i = \begin{cases} 0 & with\ probability\ p_i \\ Negative\ Binomial & with\ probability\ 1 - p_i \end{cases} \tag{5-1}$$

Where p_i represents the logistic regression and is estimated as:

$$p_i = \frac{exp(\grave{\beta_i} x_i)}{1 + exp(\grave{\beta_i} x_i)} \tag{5-2}$$

Where $\grave{\beta_i}$ is the corresponding coefficient parameter and x_i is the parameter.

Zero inflated negative binomial can be defined as:

$$P(y_i) = \begin{cases} p_i + (1-p_i)\left(\dfrac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}} & y_i = 0 \\ (1-p_i)\dfrac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\dfrac{(\alpha\lambda_i)^{y_i}}{(1+\alpha\lambda_i)^{\left(y_i+\frac{1}{\alpha}\right)}} & y_i > 0 \end{cases} \tag{5-3}$$

5.4.2 Hurdle Model

The Hurdle models consist two parts too. The difference between it and the zero-inflated negative binomial is that the zero part is a binary model and the count part is truncated at zero count model. The hurdle negative binomial models having the following form:

$$P(y_i) = \begin{cases} p_i & y_i = 0 \\ (1-p_i)\left(1-\dfrac{1}{(1+\alpha\lambda_i)^{\frac{1}{\alpha}}}\right)\dfrac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma(y_i+1)\left(\frac{1}{\alpha}\right)}\dfrac{(\alpha\lambda_i)^{y_i}}{(1+\alpha\lambda_i)^{\left(y_i+\frac{1}{\alpha}\right)}} & y_i = 0 \end{cases} \tag{5-4}$$

5.5 Modeling Results and Discussion

The results of the 32 models (zero-inflated negative binomial using random effect in the count part, zero-inflated negative binomial using random effect in the zero part, zero-inflated negative binomial using a pair of uncorrelated random effects in both part, zero-inflated negative binomial using a pair of correlated random effects in both parts, and same has been applied for hurdle negative binomial for different time periods) are presented in Tables 5-1 to 5-6.

As it is shown in the tables, ten variables were used in each part of the models. Log of volume/ per lane ‘log (volume/lane’, log standard deviation of volume ‘SD(Volume)’, log average of speed ‘log(Speed)’, log standard deviation of speed ‘SD(Speed)’, log speed limit ‘Speed limit’, dummy variable for auxiliary lane existence (1 when there is an auxiliary lane) ‘Auxiliary lane’, dummy variable for the existence of a horizontal curve ‘Horizontal curve’, log of the segment length ‘log(length)’, and dummy variable for the direction (1 when the direction is eastbound). The modeling results for the both zero-inflated negative binomial and Hurdle with a correlated pair of random effects for "evening peak and night time" time period were not converted.

Table 5-1: Zero-inflated Negative Binomial and Hurdle Negative Binomial Results (Morning Peak).

	Weekday Models (Morning Peak)							
Parameters	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)
Count Part								
Intercept	**17.0135 (0.0017)**	**5.7353 (0.0735)**	**16.2377 (<.0001)**	**12.4927 (0.0056)**	**-8.4952 (<.0001)**	**5.7239 (0.0745)**	**-7.7427 (<.0001)**	**6.1264 (0.0600)**
Log(Volume/lane)	**1.3818 (<.0001)**	**1.1098 (<.0001)**	**1.2445 (<.0001)**	**1.1754 (<.0001)**	**1.6049 (<.0001)**	**1.1129 (<.0001)**	**1.3930 (<.0001)**	**0.9753 (<.0001)**
SD(Volume)	-	-	-	-	-	-	-	-
Log(Speed)	**-2.1931 (0.0060)**	**-2.8918 (<.0001)**	**-1.7945 (0.0002)**	**-2.2401 (0.0002)**	-	**-2.8936 (<.0001)**	-	**-2.9441 (<.0001)**
SD(Speed)	**0.2299 (0.0789)**	-	-	-	-	-	-	-
Speed limit	**-3.813 (0.0026)**		**-3.7622 (<.0001)**	**-2.3664 (0.0455)**	-	-	-	-
Auxiliary Lane	-	**-0.6855 (0.0033)**	-	**-0.7109 (0.0001)**	-	**-0.6854 (0.0034)**	**-0.5951 (0.0216)**	-
Horizontal Curve	**-0.295 (0.0179)**	-	**-0.4301 (0.0021)**	-	-	-	-	-
% of Heavy Vehicle	-	-	-	-	-	-	-	-
Log(length)	**0.7031 (<.0001)**	-	**0.2240 (0.0238)**	-	**0.6947 (0.0002)**	-	-	-
Direction	-	-	-	-	-	-	-	-
Zero Part								
Intercept	**8.8977 (0.0214)**	**-15.7697 (0.0004)**	**-2.7220 (0.0372)**	**-16.0155 (0.0097)**	**-30.0073 (0.1074)**	**-14.8880 (0.0171)**	**-33.9049 (0.0213)**	**11.2410 (<.0001)**
Log(Volume/lane)	-	**-0.7867 (0.0038)**	-	**-0.8054 (0.0231)**	-	**-0.9101 (0.0111)**	-	**-1.3695 (<.0001)**
SD(Volume)	**-2.525 (0.0231)**	**-0.5770 (0.0090)**	-	**-0.6110 (0.0409)**	**-1.3547 (0.0256)**	**-0.6097 (0.0441)**	**-0.8342 (0.2290)**	**-0.7031 (0.0203)**
Speed	-	-	-	-	-	-	-	-
SD(Speed)	-	**-0.5339 (<.0001)**	**-1.0782 (0.0015)**	**-0.5967 (0.0003)**	**-0.7539 (0.0931)**	**-0.5728 (0.0006)**	**-0.9702 (0.0239)**	**-0.4641 (0.0029)**
Speed limit	-	**5.4938 (<.0001)**	-	**5.6381 (<.0001)**	**8.5544 (0.0470)**	**5.5087 (0.0001)**	**8.4526 (0.0105)**	-
Auxiliary Lane	-	-	-	-	-	-	**-1.5417 (0.0701)**	-
Horizontal Curve	-	**0.2525 (0.0109)**	**0.7482 (0.0083)**	**0.3135 (0.0204)**	**0.6526 (0.0440)**	-	**0.5752 (0.0277)**	-
% of HV	-	-	-	-	-	-	-	-
Log(length)	-	**-0.8233 (<.0001)**	**-1.4342 (0.0271)**	**-0.8462 (<.0001)**	-	**-0.9207 (<.0001)**	**-2.0346 (0.0032)**	**-0.6743 (0.0033)**
Direction	-	-	**2.1612 (0.0003)**	**0.9978 (0.0140)**	-	**0.9107 (0.0247)**	**1.2883 (0.0679)**	-
Alpha	**0.0569 (0.4339)**	-	**0.1142 (0.1789)**	**0.1515 (0.1905)**	**0.05579 (0.4219)**	**0.04131 (0.5814)**	**0.05397 (0.4255)**	**0.04153 (0.5872)**
Variance I	**0.5484 (0.0002)**	-	-	-	**0.5990 (0.0009)**	**0.1789 (0.0936)**	**0.2789 (0.0168)**	**0.2947 (0.0338)**
Variance II	-	-	**2.9680 (0.0256)**	**0.5595 (0.0061)**	**0.4894 (0.6418)**	**0.6003 (0.0053)**	**1.5546 (0.2362)**	**0.8920 (0.0012)**
Correlation	-	-	-	-	-	-	**-0.1846 (0.6190)**	**-0.5127 (0.0011)**
-2 Log Likelihood	**1684.7**	**1688.7**	**1673.1**	**1671.6**	**1658.7**	**1674.1**	**1675.7**	**1687.4**
AIC	**1706.7**	**1716.7**	**1699.1**	**1701.6**	**1680.7**	**1702.1**	**1705.7**	**1711.4**
BIC	**1738.8**	**1757.6**	**1737.0**	**1745.4**	**1712.8**	**1743.0**	**1749.5**	**1746.4**

Table 5-2: Zero-inflated Negative Binomial and Hurdle Negative Binomial Results (Off-Peak).

	Weekday Models (Off Peak)							
Parameters	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)
Count Part								
Intercept	**14.0535 (0.1093)**	**20.3367 (0.0482)**	**14.0603 (0.0343)**	**16.3251 (0.0818)**	**-8.2813 (<.0001)**	**16.0304 (0.1127)**	**-8.9077 (<.0001)**	**-16.0099 (<.0001)**
Log(Volume/lane)	**1.2326 (<.0001)**	**1.3752 (0.0287)**	**1.3847 (<.0001)**	**1.2267 (0.0782)**	**0.7516 (0.0018)**	**1.5308 (0.0137)**	**0.9059 (0.0092)**	**2.7696 (<.0001)**
SD(Volume)	-	-	-	**0.6190 (0.0828)**	**0.8095 (<.0001)**	-	**0.6388 (0.0016)**	-
Log(Speed)	**-5.2596 (0.0107)**	-	**-5.2454 (0.0006)**	-	-	-	-	-
SD(Speed)	-	**0.7035 (0.0028**	-	-	-	**0.6075 (0.0109)**	-	-
Speed limit	-	**-7.1101 (0.0011)**	-	**-6.9455 (0.0005)**	-	**-6.5893 (0.0019)**	-	-
Auxiliary Lane	-	-	-	-	-	-	-	-
Horizontal Curve	-	**0.3526 (0.0283)**	-	**0.4305 (0.0033)**	-	**0.4118 (0.0088)**	-	-
% of Heavy Vehicle	-	-	-	**0.7862 (0.0324)**	-	**0.6126 (0.0982)**	-	-
Log(length)	**0.4282 (0.0582)**	**0.4265 (0.0938)**	**0.2792 (0.0191)**	**0.5857 (0.0204)**	**0.6974 (<.0001)**	**0.5730 (0.0283)**	-	**0.6033 (0.0220)**
Direction	-	-	-	**-0.5681 (0.0620)**	-	-	**-0.2913 (0.1451)**	-
Zero Part								
Intercept	**2.4342 (0.2104)**	**-7.4866 (0.0240)**	**2.5171 (0.0756)**	**-6.2673 (0.2725)**	**-2.5843(<.0001)**	**-3.7067 (0.5176)**	**-42.2112 (0.0797)**	**9.7432 (<.0001)**
Log(Volume/lane)	**-1.0874 (0.0195)**	**-0.5306 (0.0317)**	**-0.7910 (0.0145)**	**-0.6331 (0.0860)**	-	**-0.6978 (0.0591)**	-	**-1.6940 (<.0001)**
SD(Volume)	-	**-0.7227 (<.0001)**	-	**-0.8478 (<.0001)**	-	**-0.8421 (<.0001)**	-	-
Speed	-	-	-	-	-	-	**9.2734 (0.0979)**	-
SD(Speed)	-	-	-	-	**-0.8628 (0.0063)**	-	-	-
Speed limit	-	**3.4063 (<.0001)**	-	**3.4073 (0.0073)**	-	**2.8623 (0.0241)**	-	-
Auxiliary Lane	-	-	-	-	-	-	-	-
Horizontal Curve	-	-	-	-	-	-	-	-
% of HV	-	-	-	-	-	-	-	-
Log(length)	**-0.8582 (0.0929)**	**-0.7817 (<.0001)**	**-1.0391 (0.0090)**	**-0.9135 (<.0001)**	-	**-0.9064 (<.0001)**	**-2.1604 (0.0057)**	**-0.9395 (<.0001)**
Direction	-	**0.2482 (0.0228)**	-	-	-	-	-	-
Alpha	**0.006210 (0.9604)**	**0.1780 (0.5801)**	**0.006210 (0.9506)**	**0.6483 (0.3662)**	**0.1013 (0.2745)**	**0.1878 (0.5839)**	**0.006210 (0.9585)**	**0.2585 (0.4847)**
Variance I	**1.0448 (0.0007)**	**0.2642 (0.2545)**	-	-	**0.5633 (<.0001)**	**0.1600 (0.4474)**	**0.5077 (0.0064)**	**0.9981 (0.0512)**
Variance II	-	-	**3.0412 (0.0176)**	**0.7295 (<.0001)**	**8.9683 (0.0006)**	**0.7331 (<.0001)**	**1.0834 (0.5545)**	**0.9456 (<.0001)**
Correlation	-	-	-	-	-	-	**-0.3498 (0.3270)**	**-0.9715 (0.0005)**
-2 Log Likelihood	**2747.7**	**2840.1**	**2804.3**	**2734.9**	**2745.4**	**2735.3**	**2742.1**	**2755.9**
AIC	**2765.7**	**2868.1**	**2822.3**	**2764.9**	**2763.4**	**2765.3**	**2764.1**	**2775.9**
BIC	**2792.0**	**2909.0**	**2848.5**	**2808.7**	**2789.7**	**2805.1**	**2796.3**	**2809.1**

Table 5-3: Zero-inflated Negative Binomial and Hurdle Negative Binomial Results (Evening Peak).

	Weekday Models (Evening Peak)							
Parameters	**15 Min. (ZINB)**	**15 Min. (HNB)**	**15 Min. (ZINB)**	**15 Min. (HNB)**	**15 Min. (ZINB)**	**15 Min. (HNB)**	**15 Min. (ZINB)**	**15 Min. (HNB)**
Count Part								
Intercept	-6.8881 (0.0021)	-10.1539 (0.0274)	-7.3097 (<.0001)	-6.3250 (0.0284)	-6.8791 (0.0005)	-10.1433 (0.0096)		
Log(Volume/lane)	1.1493 (0.0032)	1.7199 (0.0286)	1.3361 (<.0001)	0.9625 (0.0643)	1.2115 (0.0004)	1.7443 (0.0102)		
SD(Volume)	-	-	-	-	-	-		
Log(Speed)	-	-	-	-	-	-		
SD(Speed)	-	-	-	0.7644 (0.0007)	-	-		
Speed limit	-	-	-	-	-	-		
Auxiliary Lane	-	-	-	-	-	-		
Horizontal Curve	-	-	-	-	-	-		
% of Heavy Vehicle	-	-	-	-	-	-		
Log(length)	0.5671 (0.0269)	0.5929 (0.0743)	0.4898 (0.0013)	0.4980 (0.0068)	0.5705 (0.0099)	0.5610 (0.0529)		
Direction	-	-	-	-	-	-		
Zero Part								
Intercept	-37.2850 (0.0359)	6.9120 (<.0001)	-37.2096 (0.0260)	5.6220 (0.0168)	-37.2718 (0.0666)	6.9606 (0.0011)		
Log(Volume/lane)	-	-1.1983 (<.0001)	-	-0.8374 (0.0540)	-	-1.2164 (0.0016)		
SD(Volume)	-	-	-	-	-	-		
Speed	8.6443 (0.0451)	-	8.7792 (0.0289)	-	8.7029 (0.0764)	-		
SD(Speed)	-	-	-	-0.6842 (0.0006)	-	-		
Speed limit	-	-	-	-	-	-		
Auxiliary Lane	-	-	-	-	-	-		
Horizontal Curve	-	-	-	-	-	-		
% of HV	-	-	-	-	-	-		
Log(length)	-0.6415 (0.0989)	-0.8346 (<.0001)	-0.8432 (0.0571)	-0.9399 (<.0001)	-0.6686 (0.0918)	-0.9976 (<.0001)		
Direction	-	-	-	-	-	-		
Alpha	0.006172 (0.9885)	0.006172 (0.9892)	0.06963 (0.6402)	0.2223 (0.4961)	0.006135 (0.9892)	0.006134 (0.9816)		
Variance I	0.9144 (0.0043)	1.0565 (0.1130)	-	-	0.9892 (0.0574)	0.7118 (0.0952)		
Variance II	-	-	2.1713 (0.0318)	0.8727 (0.0016)	2.0959 (0.5506)	0.6838 (0.0018)		
Correlation	-	-	-	-	-	-		
-2 Log Likelihood	1274.3	1314.2	1296.6	1286.0	1268.8	1273.0		
AIC	1292.3	1330.2	1312.6	1304.0	1284.8	1293.0		
BIC	1318.6	1353.5	1336.0	1330.3	1308.1	1322.2		

Table 5-4: Zero-inflated Negative Binomial and Hurdle Negative Binomial Results (Night time).

	Weekday Models (Night Time)							
Parameters	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)	15 Min. (ZINB)	15 Min. (HNB)
Count Part								
Intercept	16.2995 (0.0074)	20.3696 (0.0591)	19.6445 (0.0005)	21.2356 (0.0214)	12.4837 (0.2457)	20.4066 (0.0587)		
Log(Volume/lane)	0.5868 (<.0001)	0.5825 (0.0025)	0.5846 (<.0001)	0.4479 (0.0237)	0.5780 (0.0024)	0.5824 (0.0025)		
SD(Volume)								
Log(Speed)	-4.8394 (0.0011)		-2.9283 (0.0240)					
SD(Speed)				0.6462 (0.0597)				
Speed limit		-6.3035 (0.0189)	-2.7014 (0.0005)	7.1081 (0.0006)	-4.5387 (0.0842)	-6.3126 (0.0187)		
Auxiliary Lane	0.4196 (0.0125)							
Horizontal Curve								
% of Heavy Vehicle								
Log(length)	0.9805 (<.0001)		0.7459 (<.0001)					
Direction								
Zero Part								
Intercept	-0.09190 (0.8045)	-11.7060 (0.0183)	0.1076 (0.7985)	-12.0680 (0.0627)	5.6252 (<.0001)	5.6181 (<.0001)		
Log(Volume/lane)		-0.6269 (<.0001)		-0.5827 (<.0001)	-0.5563 (<.0001)	-0.5553 (<.0001)		
SD(Volume)								
Speed		3.7959 (0.0017)		3.9288 (0.0123)				
SD(Speed)	-1.7443 (0.0011)		-1.3980 (0.0006)	-0.4720 (0.0003)	-0.6940 (<.0001)	-0.6936 (<.0001)		
Speed limit								
Auxiliary Lane				-0.4433 (0.0049)				
Horizontal Curve								
% of HV								
Log(length)		-0.9857 (<.0001)		-1.0570 (<.0001)				
Direction								
Alpha	0.2355 (0.2394)	1.4231 (0.5968)	0.4401 (0.0532)	60.9015 (0.8092)	1.0501 (0.5483)	1.4228 (0.5954)		
Variance I	0.3506 (0.0002)	0.9588 (0.0969)			1.7081 (0.0780)	0.9593 (0.0969)		
Variance II			1.8142 (0.0527)	0.2329 (0.0087)	0.6896 (<.0001)	0.6796 (<.0001)		
Correlation								
-2 Log Likelihood	3171.5	3162.8	3156.2	3182.8	3128.2	3126.5		
AIC	3191.5	3180.8	3176.2	3200.8	3146.2	3150.5		
BIC	3220.7	3207.0	3205.4	3227.1	3172.5	3185.5		

The modeling computational time is dramatically increased when a correlated pair of random effect was included in the model process. Also, in this modeling process, the adaptive Gaussian Quadrature with 20 points was considered to increase the accuracy of the modeling results. The results also show that using a pair of random effects in both the zero part and the count part improve the models. The Bayesian Information Criterion (BIC) of these models were always lower than the (BIC) value for other models. It is also found that zero-inflated negative binomial performs better than the hurdle negative binomial. The results validate other researches studies regarding zero-inflated count data. Different significant variables were found for different models, in general, the log (volume per lane) and the log (segment length) were the most common significant variables. Log of average speed and speed limit variables were found to have a negative impact on the number of crashes (i.e., as the average speed or the speed limit increase, less crashes occurs). Comparing the best zero-inflated models (zero-inflated negative models and hurdle negative binomial models) with the modeling results of the negative binomial models in Chapter 4 shows that the zero-inflated negative binomial is better than the negative binomial models and it is the best.

CHAPTER 6: SUMMARY AND CONCLUSIONS

6.1 Summary

This dissertation focuses on determining the viability of using less aggregated traffic data to find a better relationship between the traffic status and the crash occurrence, find the crash contribution factors for the expressway (SR 408) based on currently available traffic detection data for different time periods, and then hotspot identification on this road.

In Chapter 3, the viability of using less aggregated traffic data to improve the crash frequency and hotspot identification were considered. In this chapter, several SPFs for different time intervals (5, 15, 30 and 60) minutes and time periods (Morning peak, off-peak, evening peak, and night time for the weekdays and heavy traffic and low traffic for the weekend) were developed. Also, the hotspot locations based on the proposed SPFs and the traditional SPF (AADT based SPF) were conducted. The comparison results clearly highlighted that the proposed SPFs improve the accuracy of hotspot identification in term of error. It was shown that the 15 minutes time interval is the best in term of error (reducing the error term) time interval for the weekdays, while the 60 minutes interval is the best interval for the weekends.

In Chapter 4, two difficulties were noticed when we considered shorter time intervals. These difficulties were the excessive number of zeros as the time interval gets smaller, and the repeated measures problem.

Several Safety Performance Functions (SPFs) were developed with different scenarios: first, the whole data was used to develop several full SPFs, in these models, a random effect was considered in the modeling process to account for the repetition in the data. Then, to reduce the number of zero observations, several Full SPFs were developed using different data sets. Three different data sets were prepared and used in addition to the whole data set. These three-data set reduced the number of zero observation by including only either segments that have one or more crashes or two or more crashes or three or more crashes. A comparison between the hotspot identification between the full data and the proposed data set impose that the data set that has only the segments with one or more crashes give the best results in reducing the error term (i.e., the Mean Absolute Error "MAD" and the Mean Square Predicted Error "MSPE").

In Chapter 5, two zero-inflated count data models were used to develop several full SPFs with the consideration of implementing the random effect in 4 suggested ways (use the random effect either with the zero part, or count part, or uncorrelated (or correlated) pair in both parts). Additionally, to increase the accuracy of the models, the adaptive Gaussian Quadrature with 20 quadrature points were used to improve the accuracy of the estimation with the consideration of using grid search for the variance of the random effects to choose the optimal value. The results show that using two uncorrelated random effects in both parts or the zero-inflated models improve the estimations. Also, the results show that the zero-inflated negative binomial was better than hurdle negative binomial in term of AIC and BIC.

6.2 Conclusion and Implications

The findings from Chapter 3 show that using less aggregated traffic data is more appropriate and accurate to identify hotspot locations. Using less aggregated traffic data present better relationship between the contributing factors of crashes and the crash occurrence. We show the importance of treating the weekdays and the weekends separately and dividing the weekdays into four time periods (morning peak, off-peak, evening peak, and night time) and the weekends into two time periods (high volume and low volume) in addition to separating the two directions. We revealed that different variables affect the occurrence of crashes by different time periods for different weekdays or weekends. This approach with the availability of such a detailed data is important to implement, and it allows practitioners to understand the influence of the traffic pattern on the crash frequency.

Chapter 4 presents an important implication for traffic safety researchers and practitioners, as we consider shorter time intervals to present more accurately the influence of the traffic condition on the crash occurrence, two important difficulties arising. First, as the considered time interval get shorter, more zero observations occur. Second, the repeated measurement problem. Thus we suggest using a data set that has only the data of the segments that have one or more crashes within the study period. This suggested way gives better results than the previous chapter.

Chapter 5 conducted crash analysis using two most commonly used models that can account for the excess zeros in the data, and suggested different ways to implement the random effect to account for the repeated measurement and unobserved heterogeneity and to improve the accuracy of the models. The results clearly suggested that using zero-inflated negative binomial with a pair of random effects give the best results, suggesting the implementation of using this model to analyze such traffic data.

For the traffic safety practitioners, several important implications can be done based on this study. This study examined different contributing factors on a crash occurrence for different (time intervals, time periods, directions) that can give a good understanding of the significant contributing factors that effect on the crash occurrence. These finding could help the practitioner to wisely utilize the traffic safety sources and focus on one time period or direction with some regulations that will reduce the crash occurrence and increase the level of service of the road. Using less aggregated traffic data also provide engineers the ability to apply different countermeasure at different times to improve traffic safety. ITS technologies can provide several promising countermeasures that could be used in real time based on the location and time such as ramp-metering, Variable Speed Limit (VSL), Dynamic Message Signs (DMS), and High-Occupancy Toll (HOT) lanes.

REFERENCES

Aarts, L., & Van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention, 38*(2), 215-224.

Abdel-Aty, M., & Pande, A. (2005). Identifying crash propensity using specific traffic speed conditions. *Journal of safety Research, 36*(1), 97-108.

Abdel-Aty, M., & Pande, A. (2007). Crash data analysis: Collective vs. individual crash level approach. *Journal of safety Research, 38*(5), 581-587.

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., & Santos, C. D. (2007). Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. *Journal of Intelligent Transportation Systems, 11*(3), 107-120.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*(1897), 88-95.

Abdel-Aty, M. A., Hassan, H. M., Ahmed, M., & Al-Ghamdi, A. S. (2012). Real-time prediction of visibility related crashes. *Transportation research part C: emerging technologies, 24*, 288-298.

Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention, 32*(5), 633-642.

Ahmed, M., & Abdel-Aty, M. (2013). A data fusion framework for real-time risk assessment on freeways. *Transportation research part C: emerging technologies, 26*, 203-213.

Ahmed, M., Abdel-Aty, M., & Yu, R. (2012). Bayesian Updating Approach for Real-Time Safety Evaluation with Automatic Vehicle Identification Data. *Transportation Research Record: Journal of the Transportation Research Board*(2280), 60-67.

Ahmed, M., Huang, H., Abdel-Aty, M., & Guevara, B. (2011). Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis & Prevention, 43*(4), 1581-1589.

Ahmed, M., Yu, R., & Abdel-Aty, M. (2011). *Safety applications of automatic vehicle identification and real-time weather data on freeways.* Paper presented at the 18th ITS World Congress.

Ahmed, M. M., & Abdel-Aty, M. (2012). The viability of using automatic vehicle identification data for real-time crash prediction. *Intelligent Transportation Systems, IEEE Transactions on, 13*(2), 459-468.

Aitkin, M., A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. Biometrics, 1999. 55(1): p. 117-128.

Akin, D., Sisiopiku, V. P., & Skabardonis, A. (2011). Impacts of weather on traffic flow characteristics of urban freeways in Istanbul. *Procedia-Social and Behavioral Sciences, 16*, 89-99.

Baruya, A. (1998). Speed-accident relationships on different kinds of European roads. *MASTER Deliverable, 7*.

Chang, J., Oh, C., & Chang, M. (2000). *Effects of traffic condition (v/c) on safety at freeway facility sections.* Paper presented at the Transportation Research, E-Circular, Fourth International Symposium on Highway Capacity Proceedings.

Cheng, W., & Washington, S. P. (2005). Experimental evaluation of hotspot identification methods. *Accident Analysis & Prevention, 37*(5), 870-881.

Chin, H. C., & Quddus, M. A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention, 35*(2), 253-259.

Christoforou, Z., Cohen, S., & Karlaftis, M. G. (2011). Identifying crash type propensity using real-time traffic data on freeways. *Journal of safety Research, 42*(1), 43-50.

Chung, K., & Ragland, D. R. (2009). *The continuous risk profile approach for the identification of high collision concentration locations on congested highways*: Springer.

El-Basyouny, K., & Sayed, T. (2010). Safety performance functions with measurement errors in traffic volume. *Safety science, 48*(10), 1339-1344.

El-Basyouny, K., & Sayed, T. (2013). Safety performance functions using traffic conflicts. *Safety science, 51*(1), 160-164.

Garber, N. J., & Gadirau, R. (1988). Speed Variance and Its Influence on Accidents.

Golob, T. F., Recker, W., & Pavlis, Y. (2008). Probabilistic models of freeway safety performance using traffic flow data as predictors. *Safety science, 46*(9), 1306-1333.

Golob, T. F., & Recker, W. W. (2003). Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of Transportation Engineering, 129*(4), 342-353.

Guo, F., Wang, X., & Abdel-Aty, M. A. (2010). Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention, 42*(1), 84-92.

Gwynn, D. W. (1967). Relationship of accident rates and accident involvements with hourly volumes. *Traffic Quarterly, 21*(3).

Hall, D.B., Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics, 2000. 56(4): p. 1030-9.

Hassan, H. M., & Abdel-Aty, M. A. (2013). Predicting reduced visibility related crashes on freeways using real-time traffic flow data. *Journal of safety Research, 45*, 29-36.

Hauer, E. (2009). Speed and safety. *Transportation Research Record: Journal of the Transportation Research Board*(2103), 10-17.

Hauer, E., & Hakkert, A. (1988). Extent and some implications of incomplete accident reporting. *Transportation Research Record, 1185*, 1-10.

Hauer, E., Harwood, D., Council, F., & Griffith, M. (2002). Estimating safety by the empirical Bayes method: a tutorial. *Transportation Research Record: Journal of the Transportation Research Board*(1784), 126-131.

Haynes, R., Jones, A., Kennedy, V., Harvey, I., & Jewell, T. (2007). District variations in road curvature in England and Wales and their association with road-traffic crashes. *Environment and Planning A, 39*(5), 1222.

Haynes, R., Lake, I. R., Kingham, S., Sabel, C. E., Pearce, J., & Barnett, R. (2008). The influence of road curvature on fatal crashes in New Zealand. *Accident Analysis & Prevention, 40*(3), 843-850.

Hossain, M., & Muromachi, Y. (2012). A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention, 45*, 373-381.

Hossain, M., & Muromachi, Y. (2013). Understanding crash mechanism on urban expressways using high-resolution traffic data. *Accident Analysis & Prevention, 57*, 17-29.

Hourdos, J., Garg, V., Michalopoulos, P., & Davis, G. (2006). Real-time detection of crash-prone conditions at freeway high-crash locations. *Transportation Research Record: Journal of the Transportation Research Board*(1968), 83-91.

Huang, H., & Abdel-Aty, M. (2010). Multilevel data and Bayesian analysis in traffic safety. *Accident Analysis & Prevention, 42*(6), 1556-1565.

Joshua, S. C., & Garber, N. J. (1990). Estimating truck accident rate and involvements using linear and Poisson regression models. *Transportation planning and Technology, 15*(1), 41-58.

Jovanis, P. P., & Chang, H.-L. (1986). Modeling the relationship of accidents to miles traveled. *Transportation Research Record, 1068*, 42-51.

Kockelman, K. M., & Kweon, Y.-J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis & Prevention, 34*(3), 313-321.

Kononov, J., Bailey, B., & Allery, B. (2008). Relationships between safety and both congestion and number of lanes on urban freeways. *Transportation Research Record: Journal of the Transportation Research Board*(2083), 26-39.

Kononov, J., Durso, C., Reeves, D., & Allery, B. (2012). Relationship between traffic density, speed, and safety and its implications for setting variable speed limits on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(2280), 1-9.

Kononov, J., Lyon, C., & Allery, B. (2011). Relation of flow, speed, and density of urban freeways to functional form of a safety performance function. *Transportation Research Record: Journal of the Transportation Research Board*(2236), 11-19.

Kwon, O. H., Park, M. J., Chung, K., & Yeo, H. (2012). *Comparing the Performance of Sliding Moving Window, Peak Searching, and Continuous Risk Profile Methods for Identifying High Collision Concentration Locations.* Paper presented at the Transportation Research Board 91st Annual Meeting.

Laureshyn, A., Svensson, Å., & Hydén, C. (2010). Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation. *Accident Analysis & Prevention, 42*(6), 1637-1646.

Lave, C. A. (1985). Speeding, coordination, and the 55 mph limit. *The American Economic Review*, 1159-1164.

Le, T., & Porter, R. (2012). Safety evaluation of geometric design criteria for spacing of entrance-exit ramp sequence and use of auxiliary lanes. *Transportation Research Record: Journal of the Transportation Research Board*(2309), 12-20.

Lee, C., & Abdel-Aty, M. (2008). Two-level nested Logit model to identify traffic flow parameters affecting crash occurrence on freeway ramps. *Transportation Research Record: Journal of the Transportation Research Board*(2083), 145-152.

Lee, C., Abdel-Aty, M., & Hsia, L. (2006). Potential real-time indicators of sideswipe crashes on freeways. *Transportation Research Record: Journal of the Transportation Research Board*(1953), 41-49.

Lee, C., Hellinga, B., & Saccomanno, F. (2003). Real-time crash prediction model for application to crash prevention in freeway traffic. *Transportation Research Record: Journal of the Transportation Research Board*(1840), 67-77.

Liu, Q. and D.A. Pierce, A Note on Gauss-Hermite Quadrature. Biometrika, 1994. 81(3): p. 624-629.

Lord, D., Manar, A., & Vizioli, A. (2005). Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. *Accident Analysis & Prevention, 37*(1), 185-199.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice, 44*(5), 291-305.

Madanat, S., & Liu, P.-C. (1995). A prototype system for real-time incident likelihood prediction. *ITS-IDEA Program Project Final Report*.

Manual, H. S. (2010). American association of state highway and transportation officials (AASHTO). *Washington, DC, 10*.

Martin, J.-L. (2002). Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis & Prevention, 34*(5), 619-629.

Mensah, A., & Hauer, E. (1998). Two problems of averaging arising in the estimation of the relationship between accidents and traffic flow. *Transportation Research Record: Journal of the Transportation Research Board*(1635), 37-43.

Miaou, S.-P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention, 26*(4), 471-482.

Miaou, S.-P., & Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record: Journal of the Transportation Research Board*(1840), 31-40.

Milton, J., & Mannering, F. (1998). The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation, 25*(4), 395-413.

Min, Y., & Agresti, A., Random effect models for repeated measures of zero-inflated count data. . Statistical Modelling, 2005. 5.

Monsere, C. M., Bosa, P. G., Bertini, R. L., & Board, T. R. (2007). *Exploring Methods for Ranking Speed and Winter-Weather Related Crash Locations in Oregon by Combining Climate, Crash, and Highway Data.* Paper presented at the Transportation Research Board 86th Annual Meeting.

Montella, A. (2010). A comparative analysis of hotspot identification methods. *Accident Analysis & Prevention, 42*(2), 571-581.

Nilsson, G. (2004). *Traffic safety dimensions and the power model to describe the effect of speed on safety.* Lund University.

Noland, R. B. (2003a). Medical treatment and traffic fatality reductions in industrialized countries. *Accident Analysis & Prevention, 35*(6), 877-883.

Noland, R. B. (2003b). Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. *Accident Analysis & Prevention, 35*(4), 599-611.

O'Donnell, C., & Connor, D. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis & Prevention, 28*(6), 739-753.

Pack, M. L., Bryan, J. R., & Steffes, A. (2008). *Overview and Status of Regional Integrated Transportation Information System in the National Capital Region.* Paper presented at the Transportation Research Board 87th Annual Meeting.

Pande, A., & Abdel-Aty, M. (2006). Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention, 38*(5), 936-948.

Park, B.-J., & Lord, D. (2009). Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention, 41*(4), 683-691.

Persaud, B., & Dzbik, L. (1993). *Accident prediction models for freeways*.

Persaud, B. N., & Mucsi, K. (1995). Microscopic accident potential models for two-lane rural roads. *Transportation Research Record*(1485), 134-139.

Pinheiro, J.C. and D.M. Bates, Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of computational and Graphical Statistics, 1995. 4(1): p. 12-35.

Poch, M., & Mannering, F. (1996). Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering, 122*(2), 105-113.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles, Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal, 2002. 2(1): p. 1-21.

Shankar, V., Albin, R., Milton, J., & Mannering, F. (1998). Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. *Transportation Research Record: Journal of the Transportation Research Board*(1635), 44-48.

Shankar, V., & Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *Journal of safety Research, 27*(3), 183-194.

Shankar, V., Mannering, F., & Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention, 27*(3), 371-389.

Smith, B., & Venkatanarayana, R. (2007). New methodology for customizing quality assessment techniques for traffic data archives. *Transportation Research Record: Journal of the Transportation Research Board*(1993), 165-174.

Taylor, M., Baruya, A., & Kennedy, J. (2002). The relationship between speed and accidents on rural single-carriageway roads.

Venkataraman, N., Ulfarsson, G., Shankar, V., Oh, J., & Park, M. (2011). Model of relationship between interstate crash occurrence and geometrics: exploratory insights from random parameter negative binomial approach. *Transportation Research Record: Journal of the Transportation Research Board*(2236), 41-48.

Versavel, J. (1999). *Road safety through video detection.* Paper presented at the Intelligent Transportation Systems, 1999. Proceedings. 1999 IEEE/IEEJ/JSAI International Conference on.

Wang, C., Quddus, M. A., & Ison, S. G. (2009). Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention, 41*(4), 798-808.

Xu, C., Liu, P., Wang, W., & Li, Z. (2012). Evaluation of the impacts of traffic states on crash risks on freeways. *Accident Analysis & Prevention, 47*, 162-171.

Xu, C., Tarko, A. P., Wang, W., & Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention, 57*, 30-39.

Yau, K.K., A.H. Lee, and P.J. Carrivick, Modeling zero-inflated count series with application to occupational health. Comput Methods Programs Biomed, 2004. 74(1): p. 47-52.

Yu, R., & Abdel-Aty, M. (2013a). Investigating the different characteristics of weekday and weekend crashes. *Journal of safety Research, 46*, 91-97.

Yu, R., & Abdel-Aty, M. (2013b). Multi-level Bayesian analyses for single-and multi-vehicle freeway crashes. *Accident Analysis & Prevention, 58*, 97-105.

Yu, R., Abdel-Aty, M., & Ahmed, M. (2013). Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis & Prevention, 50*, 371-376.

Zhang, G., Avery, R., & Wang, Y. (2007). Video-based vehicle detection and classification system for real-time traffic data collection using uncalibrated video cameras. *Transportation Research Record: Journal of the Transportation Research Board*(1993), 138-147.

Zhang, Y., Xie, Y., & Li, L. (2012). Crash frequency analysis of different types of urban roadway segments using generalized additive model. *Journal of safety Research, 43*(2), 107-114.

Zheng, Z., Ahn, S., & Monsere, C. M. (2010). Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention, 42*(2), 626-636.

Zhou, M., & Sisiopiku, V. (1997). Relationship between volume-to-capacity ratios and accident rates. *Transportation Research Record: Journal of the Transportation Research Board*(1581), 47-52.

APPENDIX A: SR 408 DETECTORS

Table A-1: SR 408 Eastbound Detectors

		Number of lanes				Number of lanes	
ID	Milepost	Mainline (w/ TP Express)	TP Cash	ID	Milepost	Mainline (w/ TP Express)	TP Cash
1	1.2			30	11.5	5	
2	1.4	2		31	12.1	5	
3	1.7	2		32	12.5	5	
4	2.2	3		33	12.9	5	
5	2.4	3		34	13.3	5	
6	2.7	3	2	35	13.7	3	3
7	3.2	2	1	36	14.2	3	2
8	3.6	2		37	14.5	4	
9	4.3	3		38	14.7	4	
10	4.6	4		39	15	5	
11	4.9	3		40	15.7	4	
12	5.3	3		41	15.8	4	
13	6	3	2	42	16.1	4	
14	6.4	3	1	43	16.5	5	
15	6.8	3		44	17.3	3	
16	7	3		45	17.7	2	
17	7.4	3		46	18	2	
18	7.6	3		47	18.4	2	
19	8	3		48	18.8	2	
20	8.4	3		49	19	2	2
21	8.9	3		50	19.4	2	1
22	9.2	3		51	19.5	2	
23	9.4	4		52	20.1	2	
24	9.6	3		53	20.3	2	
25	9.7			54	20.8	2	

26	10.3	3		55	21.8	2	
27	10.6	4		56	22.3	2	
28	10.8	5		57	22.7	2	
29	11.2	5					

Table A-2: SR 408 Westbound Detectors.

		Number of lanes				Number of lanes	
ID	Milepost	Mainline (w/ TP Express)	TP Cash	ID	Milepost	Mainline (w/ TP Express)	TP Cash
1	1.2			29	11.6	4	
2	1.4	2		30	12.1	5	
3	1.6	3		31	12.6	5	
4	2	3		32	13	5	
5	2.4	3		33	13.3	3	2
6	2.7	2	1	34	13.6	3	4
7	3.2	2	2	35	14.2	5	
8	3.6	2		36	14.4	4	
9	4.3	3		37	14.5	5	
10	4.6	4		38	15.2	5	
11	4.9	3		39	15.7	5	
12	5.3	3		40	15.9	4	
13	5.9	3	2	41	16.1	4	
14	6.3	3	2	42	16.5	5	
15	6.8	3		43	17	3	
16	7.3	3		44	17.8	3	
17	7.4	4		45	18	3	
18	7.6	3		46	18.4	2	
19	8.1	3		47	18.8	2	
20	8.4	3		48	19	2	1
21	8.9	3		49	19.4	2	2
22	9.2	3		50	19.7	3	
23	9.7	3		51	19.9	2	
24	9.9	2		52	20.7	3	
25	10.3	3		53	20.8	2	

26	10.6	4		54	21.8	2	
27	10.9	4		55	22.3	2	
28	11.3	5		56	22.7	2	