2012

Multi-level Safety Performance Functions For High Speed Facilities

Mohamed Ahmed
University of Central Florida

MULTI-LEVEL SAFETY PERFORMANCE FUNCTIONS FOR HIGH SPEED FACILITIES

by

MOHAMED MOSTAFA ABDUL-RAHMAN AHMED

B.Sc. Al-Azhar University, Cairo, Egypt, 2001
M.Sc. University of Central Florida, USA, 2009

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 the University of Central Florida
Orlando, Florida

Spring Term
2012

Major Professor: Mohamed A. Abdel-Aty

ABSTRACT

High speed facilities are considered the backbone of any successful transportation system; Interstates, freeways, and expressways carry the majority of daily trips on the transportation network. Although these types of roads are relatively considered the safest among other types of roads, they still experience many crashes, many of which are severe, which not only affect human lives but also can have tremendous economical and social impacts. These facts signify the necessity of enhancing the safety of these high speed facilities to ensure better and efficient operation. Safety problems could be assessed through several approaches that can help in mitigating the crash risk on long and short term basis. Therefore, the main focus of the research in this dissertation is to provide a framework of risk assessment to promote safety and enhance mobility on freeways and expressways. Multi-level Safety Performance Functions (SPFs) were developed at the aggregate level using historical crash data and the corresponding exposure and risk factors to identify and rank sites with promise (hot-spots). Additionally, SPFs were developed at the disaggregate level utilizing real-time weather data collected from meteorological stations located at the freeway section as well as traffic flow parameters collected from different detection systems such as Automatic Vehicle Identification (AVI) and Remote Traffic Microwave Sensors (RTMS). These disaggregate SPFs can identify real-time risks due to turbulent traffic conditions and their interactions with other risk factors.

In this study, two main datasets were obtained from two different regions. Those datasets comprise historical crash data, roadway geometrical characteristics, aggregate weather and traffic parameters as well as real-time weather and traffic data.

At the aggregate level, Bayesian hierarchical models with spatial and random effects were compared to Poisson models to examine the safety effects of roadway geometrics on crash occurrence along freeway sections that feature mountainous terrain and adverse weather. At the disaggregate level; a main framework of a proactive safety management system using traffic data collected from AVI and RTMS, real-time weather and geometrical characteristics was provided. Different statistical techniques were implemented. These techniques ranged from classical frequentist classification approaches to explain the relationship between an event (crash) occurring at a given time and a set of risk factors in real time to other more advanced models. Bayesian statistics with updating approach to update beliefs about the behavior of the parameter with prior knowledge in order to achieve more reliable estimation was implemented. Also a relatively recent and promising Machine Learning technique (Stochastic Gradient Boosting) was utilized to calibrate several models utilizing different datasets collected from mixed detection systems as well as real-time meteorological stations.

The results from this study suggest that both levels of analyses are important, the aggregate level helps in providing good understanding of different safety problems, and developing policies and countermeasures to reduce the number of crashes in total. At the disaggregate level, real-time safety functions help toward more proactive traffic management system that will not only enhance the performance of the high speed facilities and the whole traffic network but also provide safer mobility for people and goods. In general, the proposed multi-level analyses are useful in providing roadway authorities with detailed information on where countermeasures must be implemented and when resources should be devoted. The study also proves that traffic data collected from different detection systems could be a useful asset that should be utilized appropriately not only to alleviate traffic congestion but also to mitigate increased safety risks. The overall proposed framework can maximize the benefit of the existing archived data for freeway authorities as well as for road users.

To My Beloved Family

ACKNOWLEDGMENTS

All praise is due to Allah who guided me to complete this work, without His help this research would not have been possible. This dissertation is dedicated to my parents who taught me the value of success and gave me the moral support to accomplish it in all aspects of my life. Particularly, I am deeply indebted to my mother for her untiring prayer and persistent faith in me.

I would like to express my sincere gratitude to my advisor, Dr. Mohamed Abdel-Aty, for his continuous support and valuable guidance throughout my research. I am grateful to work with such a sincere and hard working advisor. I also want to acknowledge the support of the other committee members, Dr's. Ahmed Radwan, Haitham Al-Deek, Nizam Uddin, Kevin Mackie, and Anurag Pande. Additionally, I want to express my appreciation to my colleagues for their friendship.

In addition, I am thankful for my mother in law for her unyielding love and care; without her help, my family and I would not have been able to enjoy our life while studying here at UCF.

Words fail me to express my appreciation to my wife Mona Zahra whose help, love and continuous believe in me, has lightened the load off my shoulders. I owe her for her patience, passions and sacrifice of her priceless time to make more time for me to succeed. Last but not the least I would also like to acknowledge the precious gift that was given to me by almighty Allah; my two adorable daughters, Jumana and Malak who brought the joy and happiness to our small family during this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

AADT	Annual Average Daily Traffic
ADT	Average Daily Traffic
AVI	Automated Vehicle Identification
CAR	Crash Analysis Reporting System
CART	Classification and Regression Trees
CDOT	Colorado Department of Transportation
DHSMV	Department of Highway Safety and Motor Vehicles
DIC	Deviance Information Criterion
FB	Full Bayesian
FDOT	Florida Department of Transportation
NB	Negative Binomial
OOCEA	Orlando-Orange County Expressway Authority
RCI	Roadway Characteristics Inventory
SAS	Statistical Analysis Software
SGB	Stochastic Gradient Boosting
SR	State Road
SSE	Sum of Square Errors
VMT	Vehicle Miles Traveled

CHAPTER 1. INTRODUCTION

1.1 Overview

Transportation is unquestionably one of the most important elements in any successful economy. It is the science of safe and efficient movement of people and goods. The safety comes first in the definition of the transportation and in reality safety should always be the first to be considered in all aspects of life. Traffic safety is one of the most growing researched topics in transportation because of not only lives of people are priceless but also because of tremendous delays and loss in operation performance that these crashes can cause. According to the National Highway Traffic Safety Administration, 37,261 people were killed in 2008 and more than 2.3 million were injured in traffic crashes on the U.S. roads (NHTSA, 2008). Although the crash-related fatalities and the total number of crashes seem to be decreasing in the United States in the recent years, it is not acceptable that more than thirty seven thousand people are still losing their lives every year on roadways of the U.S. and more than two millions are injured. Traffic safety research is still in need of great effort to see crashes, deaths and injuries significantly decline in the years to come.

There were 243,342 traffic crashes in Florida in 2008 compared to 256,207 in 2007, other than pedestrian and motorcycle crashes Florida saw a decrease in overall crash and injuries from 2007 and 2008. Also, the number of fatalities on Florida roadways decreased from 2007 by 7.4% going from 3,221 to 2,983 in 2008 (1.5 deaths per 100 million vehicle miles traveled) (FHSMV, 2008). Despite the positive trends of 2008 crash statistics in decreases in fatalities and injuries in

Florida, the influence of the slowing economy, increased safety of vehicles, and fewer vehicles on the road (Vehicle Mile Traveled (VMT) per registered vehicle is going down) might be the reason behind these trends.

Roadway network comprises different road types, among these types; interstates, freeways, and expressways are considered the principal arterial systems that daily carry the majority of people and goods on the transportation network. Despite the fact that the frequency of crash occurrence is typically lower on interstates, freeways and expressways when compared to other types of roads, the highest traffic volume corridors, the longest continuous trips, and the highest proportion of vehicle miles traveled take place on these roads. Therefore, crashes on these types of roads are significant, where they can affect traffic conditions for hours. These facts signify the importance of improving the safety on these high speed facilities of interstates, freeways and expressways in order to insure better and efficient mobility.

Previous effort in freeway safety studies are categorized into two types; 1) aggregate analysis in which the frequency of crashes is the number of crashes occurring in some geographical space (road segments, intersections, or network) over specific time period (months, seasons or years), and 2) disaggregate analysis focusing on relating real-time traffic data and crash occurrence on freeways in a proactive safety management framework (Golob et al., 2004).

Regarding aggregate analysis; although, many researchers have put great effort in innovative methodological approaches to account for the formidable problems in data characteristics to improve the understanding of the factors that affect crash-frequencies, there is still a room for statistical methodologies that can be introduced to provide superior statistical fit and predictive capabilities and overcome these problems (Lord and Mannering, 2010).

Both aggregate and disaggregate studies are useful, aggregate studies help in providing direction for policies and countermeasures to reduce the number of crashes while disaggregate studies dealing with real-time data help in efficient, smart and proactive traffic management that will not only enhance the performance of the network but also provide safe movement for people and goods.

1.2 Research Objectives

The work in this study focuses on different levels to assess traffic safety on high speed facilities by developing Safety Performance Functions (SPFs) on aggregate and disaggregate levels. To develop this multi-level procedure, the following main objectives were achieved;

1. Identification of main contributing factors of crash frequencies on mountainous freeways as well as identification of sites with promise ("hot-spots") using advanced Bayesian statistical technique.

   This objective was achieved by the following:

   a) Modeling crash frequency of freeway to identify the confounding factors of geometry, traffic, and weather conditions using Full Bayesian (FB) hierarchical approach.

   b) A system for identifying and ranking the sites with promise (“hot-spots”) along the freeways was developed. Bayesian approaches Full Bayesian (FB) hierarchical approach was used to accomplish this task.

2. Investigating the viability of using the automatic vehicle identification (AVI) traffic data in identification of freeway real-time “hot-spots” in a proactive traffic management framework. Developing and comparing between single generic models for all crashes and specific models for rear-end crashes using AVI data.

   The following tasks were implemented to achieve the second objective:

   a) Utilizing classical (frequentist) matched case-control logistic regression to examine the viability of using traffic data collected from Automatic Vehicle Identification Systems (AVI) on Orlando Orange County Expressway Authority (OOCEA) expressway network for real-time safety risk analysis.

   b) Applying Bayesian updating approach in order to achieve reliable crash identification.

   c) Calibrating separate models to identify specific crash types (All crashes vs. Rear-End crashes).

3. Assessing the interaction between crash occurrence, mountainous freeway geometry, real-time weather and AVI traffic data in real-time risk assessment.

Objective three was achieved by the following:

a) Using Bayesian logistic regression technique to link crashes on Interstate 70 in Colorado with real-time space mean speed collected from AVI system, real-time weather and roadway geometry data.

b) Investigating whether the inclusion of roadway and weather factors in real-time crash prediction models is required for freeway sections that feature challenging geometry and adverse weather.

c) Investigating various factors affecting crashes in real-time during different seasons by estimating separate models for distinctive seasons.

4. Developing a framework for real-time risk assessment using data from multiple sources (i.e. remote traffic microwave sensors, automatic vehicle identification, and real-time weather) to achieve reliable and robust prediction performance under different scenarios of data availability.

Final objective was achieved by the following:

a) Utilizing Stochastic Gradient Boosting (SGB), a relatively recent and promising machine learning technique was used to calibrate several models using different datasets collected from mixed detection systems as well as real-time weather stations.

b) Data from different sources were fused to provide the ultimate set of predictors, a full model using the whole data was estimated.

c) The prediction performance of each model was compared. Depending on on-line data availability, a framework was provided for real life application.

1.3 Dissertation Organization

The dissertation is organized as follows: following this chapter, a summarized literature review on previous studies of aggregate freeway crash analysis highlighting the important factors that affect crash frequency as well as discussing the different statistical methodologies used in that area, followed by a detailed review of the real-time crash prediction literature. Chapter 3 presents data, methodology, and finding of the analysis of crash frequency of 20-mile mountainous section in Colorado using Full Bayesian Hierarchical approach. The preparation of OOCEA expressways automatic vehicle identification traffic data and crash data, methodologies, and viability of using this data in real-time safety risk analysis are provided in chapter 4. Followed by chapter 5, estimating separate models for specific crash type (rear-end) and compare it to single generic model for all crashes using Bayesian updating approach as well as. Chapter 6 discusses the inclusion of geometrical characteristics and weather information in real-time risk assessment. A framework for real-time risk assessment using traffic data from mixed detection systems, real-time weather and geometry is illustrated in chapter 7. The final chapter of this dissertation, chapter 8 concludes the findings, and discusses future recommendations.

CHAPTER 2. LITERATURE REVIEW

2.1 General

The literature review is divided into two main sections. First section summarizes the studies of aggregate crash analysis on freeway in which the frequency of crashes is the number of crashes occurring in some geographical space (road segments, intersections, or network) over specific time period (months, seasons or years), in these studies the traffic flow parameters are represented by aggregated measures (e.g. AADT and speed limit). This section also shed the light on important factors that affect crash frequency as well as discussing the different statistical methodologies used in that area. Second section provides a comprehensive review of previous disaggregate studies focused on relating real-time traffic data and crash occurrence on freeways in a proactive safety management framework. In these studies the units of analysis are the disaggregate crash events and the traffic flow is represented by the corresponding real-time traffic data at the same time and location of each crash.

2.2 Aggregate Analysis of Crashes

2.2.1 Overview

The aggregate crash frequency analysis has been an effective way to gain better understanding of the contributing factors that affect the likelihood of crashes and identify locations with high crash risk potential for many decades. These studies are important to provide directions to officials for policies and countermeasures to reduce number of crashes. Crash performance functions were conventionally used to establish relationships between the traffic characteristics (e.g. speed limit, ADT, and VMT), roadway geometry (e.g. number of lanes, curvatures, grades, etc.) and environmental factors (weather), driver characteristics and behavior (e.g. gender, age, acceleration, braking and steering information, driver response to stimuli, etc.) and crash occurrence.

Ceder (1982), Garber and Ehrhat (2000), and Yan et al. (2009) established relationships between these variables and crash frequency while Abdelwahab and Abdel-Aty (2002), Al-Ghamdi (2002) and Srinivasan (2002) related these variables to the severity of crashes.

2.2.2 Factors Affecting Crash Frequency

There are many factors that contribute to crash occurrence, two main categories of these factors that affect crash frequency on freeways are; 1) behavioral factors, and 2) non-behavioral factors. The data about behavioral factors are typically not available and hence they are less reported in the literature. Traffic flow characteristics, weather, and geometry were extensively reported in many studies as the main contributing factors that affect crash frequency on freeways.

The association between roadway geometry and crash occurrence is well documented in the literature, Wong and Nicholson (1992), Boughton (1975), National Cooperative Highway Research Program (1997), and the Federal Highway Administration (1982) showed strong association between adverse geometric elements and high crash frequency.

Milton and Mannering (1998) reported that the increase in section length tends to increase crash frequency and the effect of section length on expected crash frequency has an exponential non-linear form. Their study revealed that vertical grades greater than 1 percent produced higher crash frequencies. Upgrade slopes were found to slow trucks by 15 km/h for significantly long grades and this reduction in speed found to be associated with increased passing and risk taking by faster passenger vehicles and increase in crashes. In contrary, downgrades have the effect of increasing speeds and this increase in speed results in increase in crash rates. Sharp horizontal curves with radii less than 868 m were found to decrease the crash frequency and they explained that by the fact that the drivers may be more likely to drive cautiously.

Chang and Chen (2005) established empirical relationship between freeway crash frequency and highway geometric variables, traffic characteristics, and environmental factors. They compared between Classification and Regression Tree (CART) and Negative Binomial (NB). They concluded that CART model relies more on traffic and environmental variables than geometric and location variables to classify crash frequencies on the freeway sections. According to CART, the study showed that ADT is the best single variable to classify the crash frequency on the freeway having the initial split in node 1 based on the ADT of 20,622 vehicles/lane. This indicates that the increase in ADT over 20,622 may increase crash frequency, this finding also confirmed from their NB model that the increase in ADT tend to increase crash frequency because of the increase of exposure. The second important variable to classify crash frequency was the number of rainy days, more crashes was expected with segments with rainy days more than 81 days, and even more crashes are expected with bus ADT more than 4,677 buses/day. In general, freeway sections with higher traffic volume (ADT/lanes, bus volume, truck volume, and semi-tractor volume), higher precipitation (number of days and amount of rain) are found to be more prone to be classified with higher crash rates. Regarding geometric alignment, they found that grade greater than 3.85% and degree of horizontal curvature greater than 0.4° have greater tendency to be classified with higher crash frequencies. It was indicated from NB model which relied more on geometric variables that the presence of degree of horizontal curvature greater than 8° can significantly reduce the crash likelihood.

Carson and Mannering (2001) estimated three separate models for interstate freeways, principal arterial, and minor arterial state highways to examine the effect of warning signs on ice-accident frequency. They found that spatial factors (e.g. urban), traffic characteristics (e.g. AADT, truck percentage), and geometry (e.g. shoulder width, grade) have significant effect on crash frequency while ice-warning signs do not have a statistically significant impact on the frequency or severity of crashes that involve ice.

Chang (2005) compared the predication performance of NB model and Artificial Neural Network (ANN), the study showed that ANN model is a consistent alternative method to analyze the frequency of freeway crash. From both models, it was concluded that ADT, number of lanes, vertical and horizontal alignments are significantly influence the freeway crash frequency. Accident likelihood increase by increase of each of ADT, number of lanes, sections with steep upgrades (3% or more), and sections with steep downgrades. Sections with level grades, severe horizontal curve (degree of horizontal curve greater than 6°) have reduced crash likelihood.

Moreover, the study showed an increase in likelihood of crashes at ramps area because of the impact of merging and diverging maneuvers on crash risk.

In a recent study by Park et al. (2010), the safety effect of geometric design elements for various highway facilities was evaluated. The study revealed that crash frequencies on freeway segments were associated with ADT, on-ramp density, degree of curvature, median width, number of urban freeways lanes, and spatial factors (urban/rural).

2.2.3 Statistical Techniques of Analyzing Crash Frequency

Recently, researches have put many efforts using different statistical techniques in trials of revealing the contributing factors that are associated with crash frequency on roadway segments over certain period of time. Different modeling techniques that have been ranged from conventional regression to data mining techniques such as Artificial Neural Network (ANN) and Classification and Regression Trees (CART), and Bayesian statistical techniques such as Empirical Bayes (EB) and Full Bayesian (FB) were used to analyze crash frequency data.

Lord and Mannering (2010) provided a detailed review of the key issues associated with crash-frequency data as well as an assessment of the strengths and weaknesses of the various methodological approaches that have been used to address these problems. They concluded that despite the fact that many researchers have put great effort in innovative methodological approaches to account for these formidable problems in data characteristics to improve the understanding of the factors that affect crash-frequencies, there is still a room for statistical methodologies that can be introduced to overcome these problems.

The nature of the crash-frequencies of being non-negative count data and the randomness discrete distributional property led to use Poisson and negative binomial models (NB) extensively. Poisson and NB models known also for their easy estimation (Shankar et al. 1995; Hadi et al., 1995; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000; Savolainen and Tarko, 2005).

However, Poisson and NB models have their own restrict assumptions, Poisson model for example cannot handle over- and under- dispersion while NB can only deal with over- dispersed data. In order to overcome different statistical problems in the count data associated with Poisson and negative binomial models, other alternations were applied to these models such as using zero-inflated (Poisson and negative binomial), and random effect negative binomial (Shankar et al., 1997; Carson and Mannering, 2001; Lee and Mannering, 2002; Shankar et al., 1998; Lord and Mannering, 2010).

Moreover, other non parametric models have been used such as Classification and Regression Tree (CART) and Hierarchical Tree-Based Regression (HTBR) to predict and classify the crash occurrence on freeway (Chang and Chen, 2005; Karlaftis and Golias, 2002).

Unlike Poisson and NB models, CART and HTBR have an advantage of not requiring a specified functional form. However, the CART & HTBR models have their own disadvantages of the risk of over-fitting because of the lack of formal statistical inference procedures and they also lack of handling the interactions between risk factors as explained by Harrel (2001).

Chang (2005) concluded that Artificial Neural Network (ANN) is a consistent alternative of NB for analyzing crash frequency on freeway. Similar to CART and HTBR, ANN does not require assumptions to relate risk factors to crash frequency and it features additional ability of handling the interactions between the predictors. CART, HTBR, and ANN all share another drawback of the difficulty of performing elasticity and sensitivity analyses which is important to provide the marginal effects of the variables on crash frequency.

The Full Bayesian (FB) hierarchical approach has gained momentum recently to better account for spatial correlation between observations (e.g. crashes) among locations (e.g. roadways segments or intersections). The Full Bayesian (FB) has become very common in modeling crash frequency because its capability to account for uncertainty in crash data and to provide more detailed causal inferences and more flexibility in selecting crash count distributions. Moreover, random effects can be easily included with the Full Bayesian (FB) formulation to help address individual site differences and prevent regression to the mean bias. It is concluded also that this methodology is extendable to any type of crash and different roadways.

Tunaru (2002) developed a multiple response FB hierarchical model that could support complex correlation structure. Two different ranking criteria were used to identify hazardous sites using the developed model; ranking by the posterior probability that a site is the worst and ranking by posterior distributions of ranks. He concluded that the first criteria can be used for long term projects while the second can be used for short term projects.

Aguero-Valverde and Jovanis (2007) used Full Bayesian Hierarchical Models with random effects to identify road segments with elevated weather-related crash risk. They examined two different ranking criteria; “the expected excess crash frequency (compared to similar sites) and the relative risk (the ratio of the expected number of crashes at a site divided by the number expected for similar sites)”, they found that the results were consistent from the two methods.

Huang and Chin (2009) applied Full Bays (FB) hierarchical approach to identify crash hotspot on Singapore intersection crash data (1997-2006), they showed that the FB hierarchical models have better goodness-of-fit than non-hierarchical models and even more, the hierarchical models perform significantly better in safety ranking than the naïve approach using raw crash count.

2.3 Disaggregate Crash Analysis

2.3.1 Applications of ITS-archived Data in Traffic Safety

Safety performance of a transportation facility can be assessed by crash data analysis as one of the most frequent used tool (Abdel-Aty, and Pande, 2007). Crash performance functions were conventionally used to establish relationships between the traffic characteristics, roadway and environmental conditions, driver behavior and crash occurrence. Although these models are useful to some extent, the aggregate nature of traffic parameters is not capable to identify the real-time locations with high probability of crashes.

On the other hand, real-time crash analysis had the researchers' interest recently in the last one decade since it has the capability of identifying crashes in real time and hence being more proactive in safety management rather being reactive.

Madanat and Liu (1995) used traffic flow and environmental conditions measured by surveillance sensors to estimate the incident likelihood for two types of incidents related to crashes and overheating vehicles. The incident likelihood was estimated to enhance existing incident detection algorithms. Using binary logit model, it was concluded that merging section, visibility and rain are the most significant factors affecting crash likelihood prediction.

Loop detectors data were used by Hughes and Council (1999) to explore the relationship between freeway safety and peak period operations. They found that the variability in vehicle speeds was the most significant measure that affects crash occurrence while macroscopic measures as AADT and hourly volume were poor measures in the analysis of safety. They used data from single milepost location during the peak periods of the day with assistant of snapshots provided by cameras installed on the freeway to examine the changes in system performance as it approaches the time of the crash. They concluded that "design inconsistency" is one of the most important factors of crash causation, they also suggested that "traffic flow consistency" should be considered in future research as perceived by the driver as an important variable that affect human. Moreover, they call for determining of the exact time of crash in order to avoid "cause and effect" fallacy. Also, Feng (2001) suggested that the reduction of speed variance may help in reducing crash occurrence.

Oh et al. (2001) was the first to statistically link real-time traffic conditions and crashes. A Bayesian model was used with traffic data containing average and standard deviation of flow, occupancy, and speed for 10-seconds intervals. It was concluded that the five minutes standard deviation of speed contributes the most in differentiating between pre-crash and non-crash condition. Although there sample size of 53 crashes is small, they showed the potential capability of establishing the statistical relationship. Moreover, the practical application of their finding is questionable, since five minutes before the crash is not adequate time for any remedy actions.

"Crash precursors" were first introduced by Lee et al. (2002), they hypothesized that short-term turbulence of traffic flow is significantly affecting the likelihood of crash occurrence. They used the log-linear approach to model traffic conditions leading to crashes "precursor", spatial dimension was added by using data from upstream and downstream detectors of the crash location as well as data across the three lanes at the crash location to represent factors such as speed variation along a specific section of the crash location along the roadway and between lanes. Also, traffic density was considered at the instant of the crash in addition to other external controlling factors such as weather, road geometry and time of crash. Moreover, they used speed profile captured by the detectors to estimate the actual crash time instead of using the reported crash time. They refined their analysis in a later study (Lee et al., 2003) and the coefficient of temporal variation in speed was found to have a relatively longer-term effect on crash potential than density while the effect of average variation of speed across adjacent lanes was found to be insignificant.

Golob et al. (2003) in later study developed a software tool FITS (Flow Impacts on Traffic Safety) to predict type of crashes based on the flow conditions being monitored. They used data for more than 1000 crashes from six major freeways in Orange County in California to develop the model and applied the tool in a case study on a section of SR55.

Hourdos et al. (2006) developed on-line crash-prone condition model using 110 live crashes, crash-related traffic events, and other contributing factors visualized from video traffic surveillance system (e.g., individual vehicle speeds and headways) over each lane in different places of the study area. They were able to detect 58% of the crashes successfully with a 6.8 false decision rate (where 6.8% of the crash cases were detected as non-crash cases).

Kockelman and Ma (2004) conducted a study using 55 severe crashes that occurred during January 1998 for the same area analyzed by Golob et al. (2003). Unlike all previous studies that have indicated a relationship between speed variability and crash occurrence, they concluded that speeds measured as 30-second time series and their variations are not capable of predicting crash occurrence. However, their conclusion is suspected due to the small sample size.

Similarly, Ishak and Alecsandru (2005) used data for 116 crashes occurred on Interstate 4 in Orlando, Florida. They found that it is not possible to separate pre-incident, post-incident, and non-incident traffic regimes from each other. Moreover, they indicated that traffic conditions that lead to crash might not be discernible in real-time.

Abdel-Aty and Pande (2005) were able to capture 70% of the crashes using the Bayesian classifier based methodology, probabilistic neural network (PNN) using different parameters of the speed only. They found that the likelihood of a crash is significantly affected by the logarithms of the coefficient of variation in speed at the nearest crash station and two stations immediately preceding it in the upstream direction measured in the 5 minute time slice of 10-15 minutes prior to the crash time.

Park and Ritchie (2004) used individual vehicle trajectories obtained from a state-of-the-art vehicle-signature based traffic monitoring technology to relate the lane-changing behavior and presence of long vehicles within a freeway section and speed variation. They claimed that using section speed variance rather than the point speed variance usually obtained from loop detectors data is more efficient in representing traffic changes. They concluded that these factors are significantly affecting the section speed variability.

2.3.2 Real-Time Analysis Based on Traffic Regimes

Golob et al. (2004) related different traffic regimes to crash occurrence. They used data from the six freeways in Orange County in California. They found that about 76% of all crashes occurred in four regimes out of total eight regimes of traffic flow that exist on these freeways. This indicates that specific regimes of the traffic flow is more correlated with crash occurrence than others and hence the key of crash prediction on urban freeways is distinguishing these patterns of traffic flow in real-time.

Zhang et al. (2005) also established a link between traffic congestion and freeway crashes in different weather conditions. They concluded a U-shaped curve relationship between the “Relative Risk Ratio” (a measure of crash probability) and congestion. Moderate congestion resulted in high relative risk ratio while free flow and heavy congestion found to be related with low relative risk ratio.

Matched case-control was used by Abdel-Aty et al. (2004) to link real-time traffic flow variables collected by loop detectors and crash likelihood. Matched case-control was selected because it has the capability of eliminating the influence of location, time and weather condition. They concluded that the average occupancy at the upstream station along with the coefficient of variation in speed at the downstream station, both during 5-10 minutes prior to the crash, were the most significant factors affecting crash likelihood prediction.

They extended their work in later study (Abdel-Aty et al. 2005); multi-vehicle freeway crashes under high- and low-speed traffic regimes were found to differ not only in terms of severity but also in their mechanism. Therefore, these two different distributions of 5-minute average speeds obtained from the closest station to the location of the crash suggested using different models depending on the freeway operation characteristics. Although, they used similar procedure to build low and high-speed models, the parameters entered in the two models are different. They concluded that low speed crashes mostly occur in persisting congested conditions where queues form and dissipate quite frequently. In contrary, freeway operation was found to be smooth at the high-speed crash location before the crash while they argued that some disruptive conditions originating downstream and the propagating backwards were the causes of drivers' errors and hence increasing the crash potential. Also, they found that more parameters came out to be significant from the downstream stations in high-speed model at time duration 5-15 minutes prior to the time of the crash.

2.3.3 Identification of Type of Crash Using Real-Time Data

A detailed study carried out by Golob and Recker (2001) to analyze patterns in crash characteristics as a function of real-time traffic flow, non-liner canonical correlation analysis (NLCCA) and principal component analysis were used with three different sets of variables. The first set defined lighting and weather condition, the second set defined crash characteristics of collision type, location and severity and the third set consisted of real-time traffic flow variables. It was concluded that some collision types are more common under certain traffic conditions; they found that median speed and variation in speed between the left- and interior lanes is related to the collision type. In addition, the inverse of the traffic volume has more influence than the speed in determining the severity of the crash. Although, the established statistical links between environmental factors, traffic flow, and crash occurrence is sound, their findings are limited by the fact that the speed was estimated using a proportional variable (volume/occupancy) from traffic data that were obtained from single loop detectors. Moreover, their findings are not applicable in a real-time proactive management to separate traffic conditions leading to crash from normal traffic conditions since non-crash data were not included.

Modeling crash types was argued by Kim et al. (2006) to be useful for at least three reasons:

1) Identification of sites with high crash risk of specific crash types that may not be identifiable using total crash types.

2) Countermeasures are likely suitable for only a subset of all crashes.

3) Traffic, road geometry, and environmental factors are usually associated with different crash types.

The importance of crash-type analysis was also highlighted by Pande and Abdel-Aty (2006a), they suggested that the traffic conditions preceding crashes are expected to differ by type of crash and therefore the proactive traffic management should be type-specific. They proposed a step by step approach to analyze loop detector data to identify real-time traffic conditions prone to rear-end crashes. They found that rear-end crashes may be grouped into two distinct cluster based on the average speeds prevailing within 2-mile section around the crash location 5-10 minutes prior to the crash time.

Pande and Abdel-Aty (2006b) continued their analysis with different type of crashes on freeway, they investigated lane-change related crashes on a freeway using classification tree procedure, it was concluded that all sideswipe collisions and the angle crashes that occur on the inner lanes (left most and center lanes) of the freeway may be attributed to lane-changing maneuvers. The results also revealed that average speeds upstream and downstream of the crash location, difference in occupancy on adjacent lanes and standard deviation of volumes and speed downstream of the crash location were the significant variables affecting crash occurrence.

Chris Lee et al. (2006) investigated the real-time traffic factors related to sideswipe crashes using a surrogate measure of lane change called “overall average flow ratio (OAFR)” which accounts for imbalance of lane flow across neighboring lanes during short time periods (5-10 minutes) and compared conditions for sideswipe and rear-end crashes based on those factors. They modified the original expression of “average flow ratio (AFR)” between adjacent lanes that was developed in previous experimental study of lane change by Chang and Kao (1991) by suggesting that a geometric mean of ratios of flows between adjacent lanes can be used to indicate the likelihood of sideswipe crashes. Four year loop detector data from 36.3-mile on I-4 in Orlando were used. They conducted t-test to identify the factors that are contributing more to sideswipe than rear-end crashes by comparing the average values (or percentages) of traffic related factors included average speed, flow and occupancy – lane average of 30-second speed, coefficient of variation of speed, coefficient of variation of flow, and peak/off-peak period and road geometric factor included only the curvature of road section. They found that the OAFR is a good surrogate measure of lane change as they found that the OAFR is generally higher for sideswipe than rear-end crashes at a 95% confidence level in addition to coefficient of variation of flow and peak/off-peak period. Simple logistic regression was used to quantify the relationship between these potential indicators and sideswipe, and rear-end crashes. They concluded that the odds of sideswipe relative to rear-end crashes increases as value of OAFR and coefficient of variation of flow increase and when the time period is off-peak period.

2.3.4 Crash Prediction Using Archived Weather and ITS Traffic Data

Many studies showed strong relationship between weather and speed and safety, the effect of weather may include reduced visibility, stability, and controllability. However very few studies have investigated crash occurrence using real-time traffic data while controlling for environmental and weather conditions. The study by Golob and Recker (2001) was one of the earlier studies that examined the relationship between the types of freeway crashes and the traffic flow parameters while controlling for weather and ambient lighting conditions.

Abdel-Aty and Pemmanaboina (2006) used Principal Component Analysis (PCA) and logistic regression (LR) to estimate a weather model that determines a rain index based on the rain readings at the weather station in the proximity of the I-4 corridor in Orlando. The archived rain index was used with real-time traffic loop data to model the crash potential using matched case-control logit model. They concluded that the 5-minute average occupancy and standard deviation of volume observed at the downstream station and the 5-minute coefficient of variation in speed at the station closest to the crash, all during 5-10 minutes prior to the crash occurrence along with the rain index were found to be the most significant factors to affect crash occurrence.

Hassan et al. (2010) used real-time traffic data to explore visibility related crashes on I-4 and I-95 freeways in Orlando; the main hypothetical testing was to compare between traffic flow characteristics that lead to visibility related crash with non-crash cases at reduced visibility conditions. Random Forest (RF) was used to identify significant traffic flow factors affecting visibility related crash occurrence. The identified factors were then used to examine the effects of traffic flow characteristics on visibility related crashes using matched case-control logistic regression to control for the effect of other confounding variables such as the geometric design and crash time. They found that the 5-minutes average occupancy observed at the nearest downstream station during 10-15 minutes before the crash along with the average speed measured at the downstream and upstream stations 5-10 minutes before the crash increase the probability of having visibility related crash.

2.3.5 Transferability of Real-Time Crash Potential Models

Although many studies have been conducted to statistically link real-time crash risk and traffic data collected from loop detectors, few studies addressed how the results from one freeway might transfer to another. Abdel-Aty et al. (2008) used Random Forests and multilayer perception neural network (MPNN) to test the transferability between different freeway corridors. Their model was successfully transferable from I4 in Orlando to Dutch motorways.

Pande et al. (2010) tried to explicitly address the transferability issue in a recent study, using MPNN on loop detector data collected from I-4 and I-95 in Orlando they found that while the model developed for one direction of I-4 eastbound worked reasonably for the I-4 westbound the performance was not acceptable for the I-95 sections concluding that the same model for crash risk prediction may only work for corridors with very similar travel patterns.

2.3.6 Real-Time Crash Risk Prevention

Variable Speed Limit (VSL), ramp metering, and route diversion are the main ITS and traffic management strategies that were used to increase the capacity of freeways and alleviate the congestions without costly lane additions or major redesigns of the geometry. These management strategies have also a potential application in the field of traffic safety for example; using VSL in speed harmonization by reducing speed limits at congested downstream areas helps to maintain better traffic flow and reduce the risk of mainly rear-end collisions.

Park and Yadlapati (2003) used the minimum safe distance equation as a measure of safety to compare the actual following distances with minimum recommended following distance at work zone area; they found that implementing VSL reduces the speed variation between successive vehicles throughout the work zone area and the number of rear-end crashes should be reduced as well.

Lee et al., (2004) proposed the application of the developed log-linear models by estimating real-time crash potential. They focused in this study on how to reduce the crash potential using Advanced Travel Management (ATM) systems through different strategies of variable speed limits (VSL). Microscopic simulation tool PARAMICS was used to mimic responses from the drivers to changes in speed limits. VSL was found to significantly reduce the crash potential of the simulated data.

Abdel-Aty et al. (2006) showed that using VSL helps to reduce the real-time crash risk on freeway when the freeway was operating at high speed conditions.

Allaby et al. (2006) showed that VSL is more beneficial for traffic scenarios that experiencing higher congestion on freeway corridors since VSL helps in reduction in the frequency and severity of shockwaves in the congested traffic (i.e. damping of the stop and go oscillations). However, they concluded that for less congested conditions, areas upstream of VSL response zones are more likely to experience negative relative safety benefits.

Ramp metering is widely used in the U.S. states and European countries to reduce the turbulence caused at on-ramp merge areas where slower moving vehicles try to enter into faster moving traffic stream (Bohenberger and May 1999) and hence helps to reduce speed variation and the length of queues on the mainline which has remarkable safety potential as well (Abdel-Aty and Dhindsa 2007).

Lee et al. (2006) investigated the potential of using ramp metering on an urban freeway to reduce crashes. Although their study was limited to only single ramp and the network used was not calibrated using real traffic flow data, they showed that crash prevention could be achieved using ramp metering.

Dhindsa (2006) examined larger network calibrated with real traffic data. The study found that ramp metering used on seven ramps was successful in lowering the overall real-time crash risk along the freeway corridor when operating at low speed conditions and that the safety performance was increased with the number of ramps that were metered.

Abdel-Aty et al. (2007) compared the effects of VSL and ramp metering on traffic safety, concluded that variable speed limit strategies reduced the crash potential under moderate to high speed conditions while ramp metering were found to be effective in reducing the crash potential during the low-speed conditions.

Abdel-Aty and Gayah (2010) showed that ramp metering successfully reduce both rear-end and lane change crash risks along the freeway. They examined two ramp metering strategies to reduce real-time crash risk along urban freeway. Both uncoordinated ALINEA and the coordinated Zone ramp metering algorithms successfully reduced the real-time crash risk and provided good overall safety benefits.

The main idea of route diversion in proactive traffic safety management is diverting vehicles from areas that have a high real-time likelihood of crash occurrence. The diversion will result in reduction in traffic demand in these areas and hence reduce the real-time crash risk.

Abdel-Aty and Gayah (2008) examined the ability of route diversion for reducing the real-time crash risk along urban freeway. On one hand they found that route diversion is an effective active crash prevention strategy during uncongested conditions on freeway which helped to decrease the crash risk between the locations where vehicles were diverted from and where the diverted vehicles re-enter the freeway. However, the crash risk was increased near location where vehicles re-enter the freeway due to the additional volume of merging vehicles. On the other hand route diversion found to be not effective during heavy congestion situations due to excessive crash risk migration to the locations where the diverted vehicles re-enter the freeway.

Although a great effort has been performed in analyzing real-time data collected from inductive loop detectors in safety framework, no safety analysis has been carried out using traffic data from one of the most growing surveillance system; the tag readers on toll roads (AVI). In this study, for the first time, the identification of freeway locations with high real-time crash potential has been examined using real-time speed data collected from AVIs. Various issues related to the viability of using AVI data in real-time crash prediction are discussed and presented in chapters 4, 5, 6, and 7.

CHAPTER 3. AGGREGATE ANALYSIS: SAFETY PERFORMANCE FUNCTIONS FOR MOUNTAINOUS FREEWAY

3.1 Introduction

While rural freeways generally have lower crash rates, interactions between driver behavior, traffic and geometric characteristics, and adverse weather conditions may increase the crash risk along some freeway sections. The analysis presented in this chapter is exploring the safety effects of roadway geometrics on crash occurrence along a 20-mile freeway section (Interstate 70 in Colorado) that features mountainous terrain and adverse weather.

The main objective of this analysis was to gain more understanding of the effects of roadway geometrics and weather on crash frequencies of mountainous freeways. The results from this analysis represents an essential step preceding the disaggregate crash analysis.

This research attempted an exploratory safety analysis on this section of the freeway by; 1) examining the effect of mountainous highway geometrics and traffic characteristics in adverse weather on the frequency of crashes, 2) identifying hazardous road segments and crash-prone time periods for more focus within an Advanced Traffic Management strategy.

The section of interest features mountainous road geometry and frequent severe weather. As a result of this mountainous terrain, this section of the interstate highway features steep slopes up to 7%. Moreover, climate with all its aspects of temperature, humidity, precipitation and wind is dramatically impacted by the considerable high elevations. This section experienced relatively higher fatality rate, a 0.48 per 100 million vehicle miles traveled (MVMT), compared to the entire interstate system in 2004 (fhwa.dot). In order to come up with an effective ITS upgrade, it is vital for a preliminary evaluation of the contributing factors to crash occurrence and identification of hot-spots.

To achieve the abovementioned objectives, vehicle crash data from I-70 in the state of Colorado were obtained for 6 years (2000-2005) together with roadway geometry, traffic characteristics, and adverse weather represented in the snow and dry season. A series of Negative Binomial (NB) models were fitted as a preliminary analysis to examine the significant factors that contribute to crash occurrence; the grades and weather were found to significantly affect the crash occurrence on this mountainous freeway. Full Bayesian Hierarchical models with random effect were used to fully account for the uncertainty associated with parameter estimates and provide exact measures of uncertainty on the posterior distributions of these parameters and hence overcome the maximum likelihood methods' problem of overestimating precision because of ignoring this uncertainty (Goldstein, 2003; Rao, 2003). Application of random effects models will help also in pooling strength across sets of related units and hence improve the parameter estimation in spare data (i.e. crash frequency models) (Aguero-Valverde and Jovanis, 2007). Moreover, since the crash risk might be spatially correlated among adjacent roadway segments, Bayesian spatial models were also examined. Finally, Bayesian ranking techniques were used to effectively rank the hazard levels associated with the roadway segments of analysis.

3.2 Description of Roadway Section

3.2.1 General Description

The freeway section under consideration is a 20.13 miles long of I-70 starting at Mile Marker (MM) 205.673 at Silverthorne and ends at MM 225.80 at Silver Plume in Colorado. The section encompasses three main parts; the Eisenhower Memorial Tunnel of 1.69 miles long starting at MM 213.18 and ending at MM 214.87, about 7.5 miles of the west side of the tunnel and 11.60 miles of the east side. The Eisenhower Tunnel is a twin bore tunnel with 26 feet of travel width (two lanes of 13 feet each). The tunnel is the highest point along the interstate highway system with an elevation of 11,158 ft and an average grade of 1.7 percent rising toward the west (Coloradodot.info).

3.2.2 Road Alignment

The section passes through extreme mountainous terrain. The horizontal alignment of this section has relatively several sharp horizontal curves' radii. In addition to the steep grades on the west and east sides of the tunnel, as shown in Figure 3-1, the west side has grades up to about 7% while the east side has grades that vary from 1.3% to 6%.

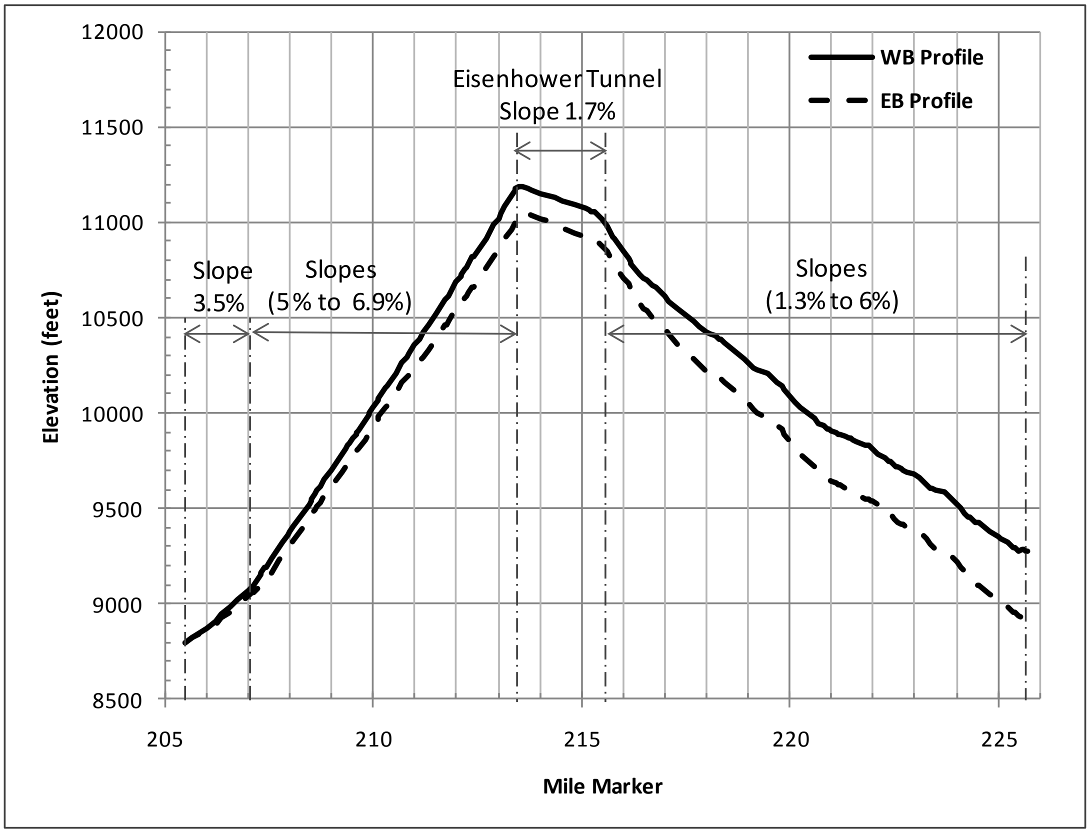


Figure 3-1: Longitudinal Profile

3.2.3 Climate

The section has a quite complex climate compared to most of the U.S. highways. The elevations in the vicinity of the area vary from 8,700 feet to more than 14,000 feet on the highest peaks above the Eisenhower tunnel. The climate within this section is affected by the high altitudes and typically results in variations of all aspect of climate such as temperature, humidity, precipitation and, wind within short distance and time. The crash report identifies the weather and pavement conditions when a crash occurs. The plots of crash frequencies vs. weather and road conditions (see Figure 3-2) conform to the metrological data (climate.colostate.edu), suggesting that there are two main seasons: snow season from October through April and the dry season from May through September which experience small amount of rain, this can explain the small percentage of rain related crashes of 6% that occurred almost exclusively within the dry season. Regarding the distribution of weather-related crashes over the 6 years, 47% of the total crashes occurred within snowy weather where the pavement condition was icy, snowy or slushy, about 6% of the total crashes occurred in rain where the pavement was wet while all other 47% occurred within clear weather and dry pavement conditions. It is worth mentioning that small percentage of snow related crashes occurred within the defined dry season (about 2%) while a negligible number of rain related crashes occurred within the defined snow season (only 2 crashes on WB in the month of October within the 6 years). Classifying the climate into two main seasons will help us understand if there is a significant difference between crashes occurring within seasons that feature snow versus dry and the underlying seasonal effect on the roadway segments. Careful examination of the trends depicted in Figure 3-2 produced these two main seasons. Although, all crashes related to weather and pavement conditions are aggregated within the two seasons to develop the data structure needed for the modeling effort of this study, the likelihood of crash occurrence in normal weather and dry pavement conditions remains constant in both seasons. Moreover, modeling the crash frequency of each specific weather condition (to account for a third rain season) would result in zero inflated problems associated with the short segments of the mountainous road section and the low crash frequency. Thus we were constrained by the data to use 2 main seasons, although more seasons might be possible on other freeways with higher crash frequencies and more distributed crashes per season.

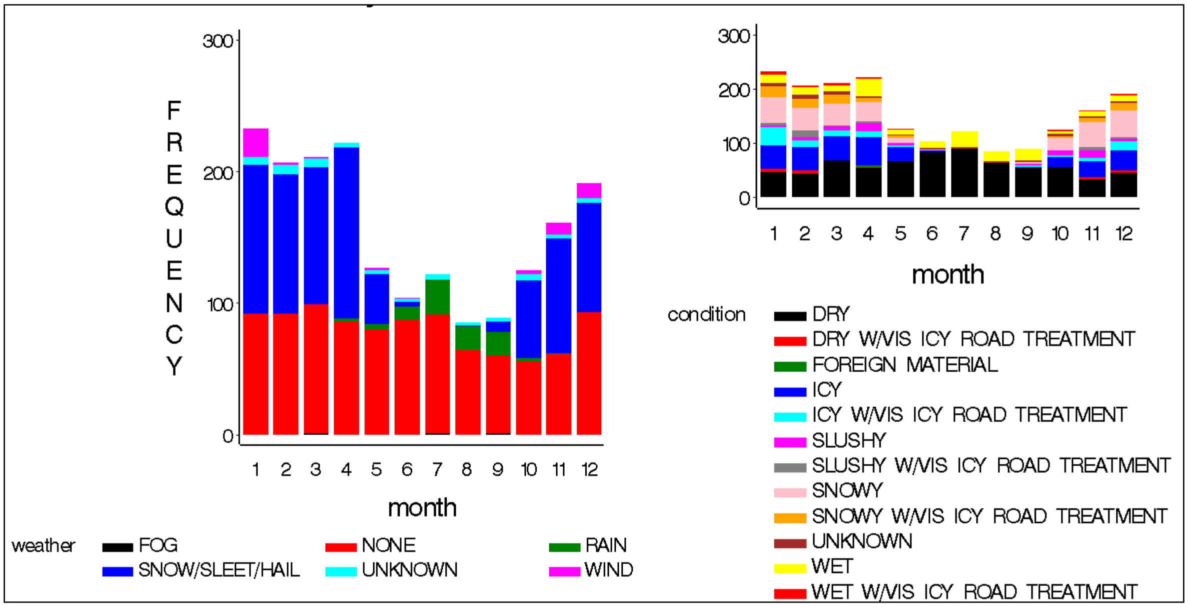


Figure 3-2: Distribution of the Monthly Crashes by Weather and Pavement Conditions for Aggregated 6 Years

3.3 Data Preparation and Preliminary Crash Analysis

There are many factors that contribute to crash occurrence, including driver behavior, traffic and geometric characteristics, weather conditions and interrelationships between these different factors. Unfortunately, the driver behavior factors are usually not available. Therefore, the available roadway, traffic and weather conditions factors were used in this study. There were two sets of data used in the study; roadway data and crash data. The roadway data were collected from CDOT, Roadway Characteristics Inventory (RCI) and Single Line Diagrams (SLD). The crash data were obtained from the road crash database maintained by CDOT.

A first but essential step in data preparation is road segmentation. Given the variation of road geometry, a major criterion employed for segmentation in this study was homogeneity in roadway alignment. According to the RCI data, both horizontal and vertical alignments were scrutinized. Moreover, a minimum-length criterion was set to 0.1 mile to avoid the low exposure problem and the large statistical uncertainty of the crash rate per short segment (Miaou 1994). Segments shorter than 0.1 mile were combined with adjacent segment with similar geometrical characteristics as much as possible. For example, a 0.021 mile long straight segment was combined with the preceding segment with smooth curve of 39755 feet radius, rather than the subsequent sharp-curved segment with 1813 feet radius. With this approach, 20 less-than-0.1mile segments from 104 homogeneous segments were combined with their adjacent segments, resulting in 84 segments for each direction. Table 3-1 illustrates the definitions and descriptive statistics of traffic, road geometrics, and weather characteristics for the segments.

Segment length and AADT are multiplied to estimate daily VMT to reflect the crash exposure for each segment. Among risk factors, of most interest are road alignment factors. The longitudinal grades are defined as a categorical variable with 8 categories gradually from upgrade (being positive) to downgrade (being negative), categorizing grades within 2% according to the American Association of State Highway and Transportation Officials (AASHTO 2004) classification would help in reducing the number of short segments by combining the segments that share all other geometrical characteristics and fall within the same grade range and hence avoiding excessive zero frequency within short segments without losing interpretable useful information about grades. For segments with multiple grades, the equivalent grade for those segments was calculated in accordance with the Highway Capacity Manual (HCM 2000) (Highway Capacity Manual 2000). Specifically, an overall average grade was calculated in case of no single portion of the grade is steeper than 4 percent or the total length of the grade is less than 0.75 mile. For some sub segments steeper than 4 percent, the HCM (2000) (Highway Capacity Manual 2000) composite grade procedure was used to determine an equivalent grade.

Defining variables for horizontal alignment is more complicated. The basic parameters, including curve radius, deflection angle, and degree of curvature, are parameterized for the curve contained in each segment. The curve direction is also monitored as safety effect may be different between left-side and right-side curves. Other variables speed limit, median width, shoulder width, number of lanes, and truck percentage, are also included as control variable although there are no much variation for these factors at the 20-mile freeway section.

Table 3-1: Summary of Variables Descriptive Statistics

Variables	Description	Mean	Stdev	Min	Max
Response Variable					
Crash Frequency	Frequency of all crashes per segment	5.45	7.37	0	55
Exposure Variables					
Segment Length	Length of the road segment (mile)	0.24	0.16	0.099	0.92
AADT	Average Annual Daily Traffic	27626	1889	25500	29300
Daily_VMT	Daily Vehicle Mile Traveled	6582	4419	2267	23409
Risk Factors					
Season	Rainy = 0, Snowy = 1	-	-	-	-
Grade	Longitudinal grade, eight categories: Upgrade: 0-2%=1, 2-4%=2, 4-6%=3, 6-8%=4; Downgrade: 0-(-2)%=5, (-2)-(-4)% =6, (-4)-(-6)% =7, (-6)-(-8)% =8	-	-	-	-
Curve Radius	Curve radius (ft)	4396	6356	1348	39755
Deflection Angle	Deflection angle of curve	21.07	13.43	1.02	48.90
Degree of Curvature	Degree of the curve per segment with curves	2.39	1.13	0.14	4.25
Curve Length	Length of the curve per segment with curves	0.17	0.09	0.01	0.48
Curve Length Ratio	Percentage of curve length to total segment length	0.53	0.46	0	1
No of Lanes	Number of lanes: 2 lanes=0, 3 lanes =1	-	-	-	-
Median Width	Width of median (ft)	20.67	15.88	2	50
Outside Shoulder	Outside shoulder width (ft)	6.80	3.20	1	20
Inside Shoulder	Inside shoulder width (ft)	3.99	1.83	0	12
Speed Limit	Posted speed limit	60.95	4.8547	50	65
Truck Percentage	Percentage of Trucks	10.35	0.39	10	10.8

In the study area, a total of 1877 crashes were reported over 6 years of the study period (2000-2005), 804 and 1057 crashes occurred on the East and West bounds, respectively. Sixteen crashes were not assigned to any of the East or West directions and they were excluded from this study. Four Hundred were rear end crashes, 234 turn over crashes and 370 were collision with guard rail or median barrier while the side swipe crashes were 223 on the mainline. Twenty five percent of the crashes occurred on curves with steep grades, about 60% occurred on straight segments with steep grades and the remaining 15% occurred on either curve or straight with flat grades.

Figures 3-3 and 3-4 depict a preliminary crash distribution for east and west bound respectively. In the figures, each of the east and west bound sections are divided into 3 miles long sub-sections. Each of these sub-sections has different number of homogenous segments according to roadway geometry as explained above (e.g. first section at MM 207 has 13 homogenous segments, starts at MM 206 and ends at MM 208).

As shown in Figure 3-3, although the section that starts at MM 215 and ends at MM218 at the east bound has the second least number of 9 segments, it has the highest mean of the crash frequency of 6 and 18 for dry and snowy seasons, respectively. It is worth mentioning that the sub-section at MM 216 on east bound is located after the tunnel with average downgrade of 6.5%.

Generally, west bound has higher crash frequency within the 3 miles sub-sections than the east bound in both seasons. Similarly, the 3 miles section centered at MM 216 has the highest mean of the crash frequency of 5.56 followed by the sub-section at MM 213 having 5.30 in dry season while the sub-section at MM 213 experienced a mean of the crash frequency of 18 in the snow season.

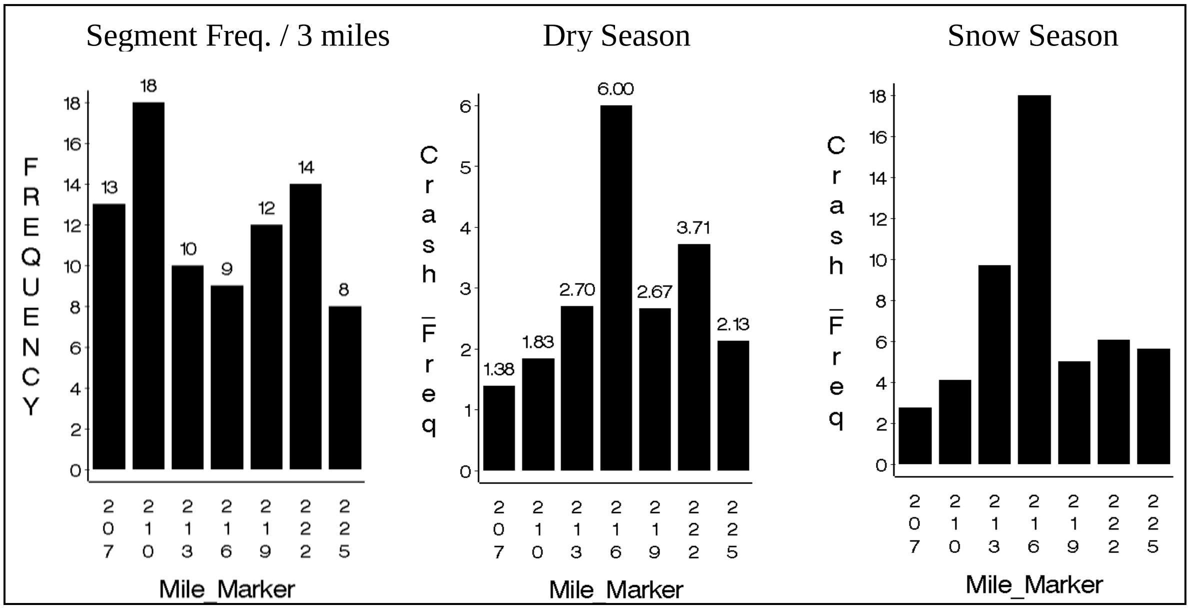


Figure 3-3: East Bound Crash Frequencies in Dry and Snowy Seasons

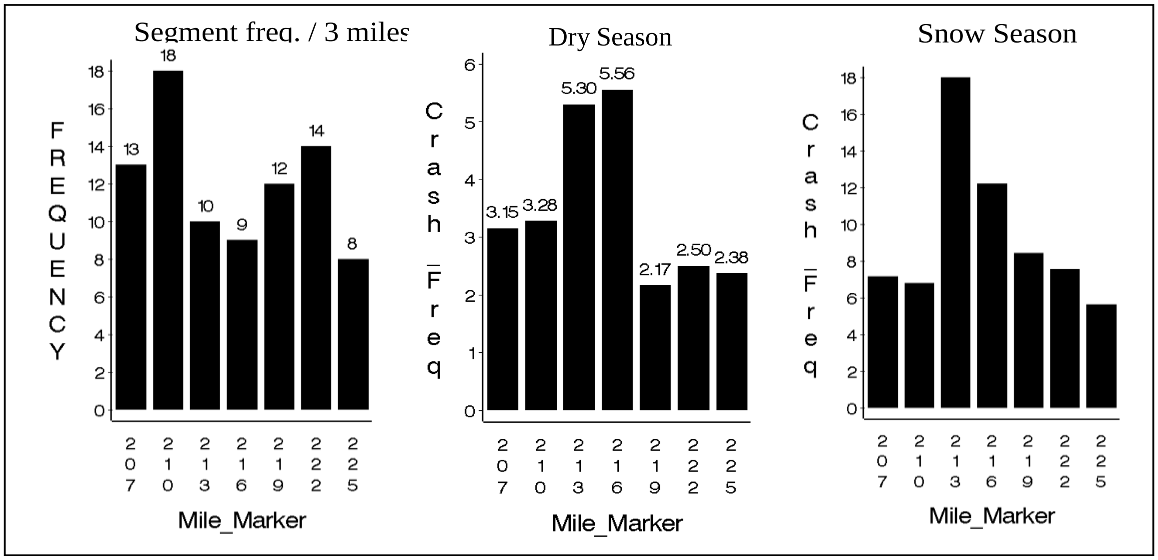


Figure 3-4: West Bound Crash Frequencies in Dry and Snowy Seasons

3.4 Bayesian Hierarchical Approach

The factors affecting the occurrence of crashes could be conceptually categorized into two groups, associated with crash exposure and crash risk, respectively.

$$\text{Crash occurrence} \sim \text{Crash exposure} \times \text{Crash risk}$$

While exposure factors account for the amount of opportunities for crashes which traffic systems or drivers experience, the risk factors reflect the conditional probability that a crash occurs given unit crash exposure. Statistically, the stochastic crash occurrence is rationally assumed to be Poisson process, which justifies the popular use of the Poisson distribution to model crash frequencies (Jovanis and Chang, 1986).

$$\begin{aligned} \mathrm{y}_{it} \mid \lambda_{it} &\sim \text{Poisson } (\lambda_{it}) = \text{Poisson } (\mu_{it} e_{it}) \\ \log \lambda_{it} &= \log e_{it} + \mathbf{X}_{\mathbf{it}}^{'} \boldsymbol{\beta} \end{aligned} \tag{3.1}$$

in which, y_{it} is the crash count at segment i (i = 1,…,168(84 segments on each direction)) during season t (t = 1 for dry season, 2 for snow season) with the underlying Poisson mean λ_{it}. μ_{it} and e_{it}, contributing to λ_{it}, denote risk factors (covariates $\mathbf{X}_{\mathbf{it}}$ and the coefficients $\boldsymbol{\beta}$) and exposure factors, respectively. Based on parameter estimation, the Incidence Rate Ratio (IRR) is generally computed to more conveniently understand the impact of covariates, say k, on the expected crash frequency for one unit change of continuous variables or binary effect for dummy variables (Haque et al., 2010).

$$IRR_k = \frac{E(y_{it} \mid \mathbf{X_{it}}, x_k + 1)}{E(y_{it} \mid \mathbf{X_{it}}, x_k)} = \exp(\beta_k) \tag{3.2}$$

In this current study, daily VMT, the product of AADT and length of road segment, is employed to reflect crash exposure associated with each road segment. Moreover, a time exposure coefficient (1 for dry season, log(5/7) for snow season) is used to offset the unbalanced design of seasons (5 month for dry season and 7 month for snow season). As shown in Table 1, risk factors include road alignment (grade and curve), road design (number of lanes, median width, and shoulders), traffic characteristics (speed limit and truck percentage), and the environmental factor (season).

In regard to model structure, given the “variance = mean” constraint of Poisson model, the Negative Binomial model (NB), a parent model of Poisson model, has been extensively employed to deal with the over-dispersion problem, which is generally observed in crash data (Miaou and Song, 2005; Persaud et al., 1997, 2001; Harwood et al., 2000; Hauer et al., 2002; Hovey and Chowdhury, 2005; Shankar et al., 1995). Nevertheless, as ordinary NB models only provides a blind account for individual heterogeneity, numerous techniques have recently been proposed to more specifically accommodate for various crash data features, for example, zero-inflation model for excess zeros (Shankar et al., 1997; Carson and Mannering, 2001; Lee and Mannering, 2002; Lord et al., 2005,2007), a two-state Markov switching count-data model to overcome the drawbacks of the traditional zero-inflated Poison (ZIP) and zero-inflated negative binomial (ZINB) (Malyshkina et al., 2009), spatial and time series model for spatiotemporal data (Aguero-Valverde and Jovanis, 2006; Quddus, 2008a, 2008b, Huang et al., 2010), hierarchical model for multilevel data structure (Huang and Abdel-Aty, 2010). Furthermore, the use of variable dispersion parameters in negative binomial models have been reported useful to improve the model-fitting (Heydecker and Wu, 2001; Miaou and Lord, 2003; Miranda-Moreno et al., 2005; El-Basyouny and Sayed, 2006; Mitra and Washington, 2007; Lord and Park, 2008). Multivariate count models have also been applied to jointly model crash frequency at different levels of injury severity (Tunaru, 2002; Park and Lord, 2007; Ma et al., 2008; Ye et al., 2009; Aguero-Valverde and Jovanis, 2009; El-Basyouny and Sayed, 2009a). More recently, a more flexible random parameter modeling approach, including random intercept and/or random slope, is emerging in the literature, in which model parameters are allowed to vary from site to site (Li et al., 2008; Anastasopoulos and Mannering, 2009; Huang et al., 2008, 2009; El-Basyouny and Sayed, 2009b; Huang and Chin, 2010). Lord and Mannering (2010) provided a detailed review of the key issues associated with crash-frequency data as well as an assessment of the strengths and weaknesses of the various methodological approaches that have been used to address these problems.

Despite the availability of various statistical model selection measures, selection of appropriate crash prediction models should be dependent on the characteristics of the specific crash data. Specifically, we have three basic observations for the current crash data: (a) Over-dispersion: the data may be highly over dispersed as the overall mean and variance equal to 5.45 and 54.32, respectively, as shown in Table 3-1; (b) Site-specific structure: each segment has two observations; crash count during each of the dry and the snow seasons. Hence, random effects may be appropriate to account for the global site-specific effects; (c) Spatial distribution: as road segments are mutually connected, spatial heterogeneities, resulting from spatial confounding factors, may exist for adjacent segments.

Based on these observations, two alternative models are suggested, i.e. random effect model (also called hierarchical Poisson model) and spatial model, both of which are modified from the basic Poisson model.

Random effect model:
$$\begin{aligned}&\log \lambda_{it} = \log e_{it} + \mathbf{X}_{\mathbf{it}}^{'}\boldsymbol{\beta} + \theta_i \\ &\exp(\theta_i) \sim \text{gamma}\,(a, a) \\ &\text{overdispersion parameter}: k = 1/a\end{aligned} \tag{3.3}$$

Spatial model:
$$\begin{aligned}&\log \lambda_{it} = \log e_{it} + \mathbf{X}_{\mathbf{it}}^{'}\boldsymbol{\beta} + \theta_i + \phi_i \\ &\theta_i \sim \text{normal}\,(0, 1/\tau_h) \\ &\phi_i \sim \text{normal}(\bar{\phi}_i, 1/\tau_i) \text{ with } \bar{\phi}_i = \frac{\sum_{i \neq j} \phi_j \omega_{ij}}{\sum_{i \neq j} \omega_{ij}} \text{ and } \tau_i = \frac{\tau_c}{\sum_{i \neq j} \omega_{ij}} \\ &\alpha = \frac{sd(\phi)}{sd(\theta) + sd(\phi)}\end{aligned} \tag{3.4}$$

Clearly, the random effect model is actually a slight modification of the ordinary NB model, in which the two observations associated with one same segment share an equal extra error component. In the spatial model, the extra variance component consists of two parts, θ_i for site-specific random effects, denoting the global extra-Poisson variability, and ϕ_i for spatial correlation with the Gaussian Conditionally Autoregressive prior (CAR model, Besag, 1974). It is noted that θ_i is assumed to be Normal distribution rather than the Gamma distribution in the random effects model. This is because the multivariate normal distribution is more convenient computationally while combining with the Gaussian spatial component (ϕ_i) than the multivariate version of Gamma distribution (Huang et al. 2010), This also is suggested by the literature that Poisson Lognormal PLN was found to provide the best statistical fit for the spatial model (Milton et al. 2008; Anastasopoulos and Mannering, 2009; Li et al., 2008; El-Basyouny and Sayed 2009a). Regarding ω, the proximity matrix, a 0-1 adjacency weight is employed. In other words, each segment is specified an equal weight to its adjacent segment(s). With the model specification, α denotes the proportion of variability in the random effects that is due to spatial heterogeneity, in which, sd is the empirical marginal standard deviation function.

Although the most common CAR model is employed in this study to model spatial effects, there are other techniques available in the literature such as Simultaneous Autoregressive (SAR), Moving Average (MA) (Congdon, 2007), and Multiple Membership (MM) (Goldstein, 1995; Goldstein et al., 1998; Langford et al., 1999). El-Basyouny and Sayed (2009c) compared CAR, MM and Extended Multiple Membership (EMM) to the traditional PLN model, they concluded that EMM provided the best fit with a little better performance than CAR and both EMM and CAR outperformed the MM and PLN.

The candidate models could be estimated conveniently by Bayesian inference using the freeware WinBUGS package (Lunn et al., 2000). The CAR model is embedded in the function "car.normal" in GeoBUGS, an add-on to WinBUGS that fits spatial models. The DIC, a Bayesian generalization of AIC, is used to measure the model complexity and fit (Spiegelhalter et al., 2003). DIC is a combination of the deviance for the model and a penalty for the complexity of the model. The deviance is defined as $-2\log(\mathrm{likelihood})$. The effective number of parameters, pD, is used as a measure of the complexity of the model, $\mathrm{pD} = \mathrm{Dbar} - \mathrm{Dhat}$, where Dbar is the posterior mean of the deviance, and Dhat is a point estimate of the deviance for the posterior mean of the parameters. DIC is given by DIC = Dhat + 2 pD. In addition, a R^2 -type Bayesian measure is developed to evaluate the model fitting,

$$R^2_{Bayes} = 1 - \frac{\sum_{\forall i,t}(y_{it} - \lambda_{it})^2}{\sum_{\forall i,t}(y_{it} - \bar{y})^2} \qquad (3.5)$$

which estimates the proportion of explained sum of squares to total sum of squares. It could be regarded as a global model-fitting measurement.

3.5 Results and Discussion

3.5.1 Model Estimation and Diagnostics

In model estimation, with no prior knowledge of the likely range of values of the parameters for mountainous freeway section, non-informative priors were specified for parameters. For each model, three chains of 20,000 iterations were set up in WinBUGS based on the convergence speed and the magnitude of the dataset. All the models were converged reasonably through visual inspection on the history plots and confirmed by the Brooks-Gelman-Rubin (BGR) convergence diagnostics (Brooks and Gelman, 1998). After ensuring the convergence, first 10,000 samples were discarded as adaptation and burn-in. To reduce autocorrelation, only every tenth samples of the rest were retained for parameter estimation, calculation of DIC and Bayesian R^2, as well as site rankings.

Exploratory modeling indicated that the crash frequencies are not significantly associated with *Speed Limit*, *Truck Percentage*, *Percentage of Curve Length* in all the three models. This was expected since there is a little variation in those variables between segments; the speed limit and the truck percentage are almost identical along the considered section and hence they were excluded from the final models. Results of model estimation with the remaining factors are summarized in Table 3-2.

Comparisons among the three candidate models imply very interesting findings. On one hand, the over-dispersion observed in crash data is confirmed by the extra variance components of the random effect model and the spatial model. Specifically, significant dispersion parameter is identified in the random effect model (k = 0.418, 95%CI (0.305, 0.561)). In the spatial model, variance components from spatial correlation and site-specific random effects are 0.469 (95%CI(0.297, 0.710)) and 0.584 (95%CI(0.481, 0.686)), respectively, which apparently indicate the proportion of the over-dispersion accounted by the spatial clustering is 44.1% (α=0.441, 95%CI(0.330, 0.560)). Moreover, model diagnostic measures confirmed that the random effect and spatial models outperform the Poisson model by accounting for over-dispersion. Specifically, DIC is substantially reduced from 1903 in Poisson to 1456 in the random effect model and 1468 in the spatial model. The Bayesian R^2 is increased from 0.61 to 0.88.

On the other hand, however, while all the parameters are significant in the Poisson model except of *Degree of curvature,* some of them come out to be insignificant in the random effect model (*Grade(4),* and *Median Width*). This phenomenon becomes more remarkable especially in the spatial model where almost all the variables turn out to be insignificant despite having the same sign as in the basic Poisson model. Another interesting observation from the parameter coefficients is that the safety effects of most of the geometry-dependant factors fade away gradually from Poisson through the other two, e.g. *Grade*, *Degree of curvature*, and *Percentage of Curve Length* etc. But the non-geometry-dependant factor (*Season*) remains constant (0.600 in Poisson, random effect model and spatial model).

Furthermore, based on estimation of pD (the number of effective variables in Bayesian model) and R^2, we found that, compared to the random effect model, the spatial model has equal R^2 (0.88) and has only an increase of 5 effective variables (pD from 117.3 to 122.3). With all these observations, we argue that the spatial model does not actually outperform the random effect model. This may be reasoned that the spatial heterogeneity mostly depends on road geometries among adjacent segments, which have been accommodated for by the well-defined geometry-dependent factors in the models. In other words, with explicit consideration for various road geometric factors in the model, the specification for spatial effect becomes redundant and hence, may reduce the significance of the geometric factors instead. We further confirmed this argument by calculating an R^2 which does not include residual terms for crash expectations (i.e. λ_{it}), as shown by R^2 (without error terms) in Table 3-2. Clearly, results indicate that the inclusion of error terms reduced the model-fitting proportion explained by the risk factors, especially in the spatial model.

In summary, the over-dispersion problem in Poisson model is effectively addressed by the random effect and spatial models, but the spatial model may have the problem of redundantly accounting for geometry-dependant effect. Therefore, the random effect model, which has the least DIC, is selected for further model inference and site ranking. The adequacy of the random effects assumption may be assessed with lack-of-fit statistics, although these statistics test the fit of the model as a whole rather than the specific random effects assumption. This random effects assumption may be made less restrictive if θ is allowed to vary with specific site effects.

Season was found to significantly affect crash occurrence (β = 0.600, 95%CI (0.499, 0.702)), the Incident Rate Ratios (IRR) are obtained by exponentiation of the regression coefficients exp[β]. IRR value shows that the risk of crashes during snow season was approximately 82% higher than the crash risk in dry season, given all other variables constant. The increased crash risk within the snow season may be explained by the confounding effect of the snowy, icy, or slushy pavement conditions during the snow season, and exacerbated by the steep slopes. This finding is important for officials to pay more attention and devote more resources during snow season than in dry season for traffic management.

Table 3-2: Parameters Estimates

Model	**Poisson**			**Random Effect**			**Spatial**		
		Credible interval			Credible interval			Credible interval	
	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
Season [snow]	0.600	0.501	0.698	0.600	0.499	0.702	0.600	0.498	0.710
Season [dry] (reference)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Grade[1]	-1.302	-1.538	-1.072	-1.287	-1.797	-0.778	-1.041	-1.950	-0.097
Grade[2]	-0.855	-1.026	-0.685	-0.870	-1.322	-0.422	-0.458	-1.400	0.534
Grade[3]	-0.786	-0.949	-0.617	-0.907	-1.285	-0.516	-0.316	-1.251	0.679
Grade[4]	-0.530	-0.735	-0.328	-0.297	-0.845	0.277	0.237	-0.745	1.286
Grade[5]	-1.193	-1.421	-0.981	-1.167	-1.674	-0.657	-0.663	-1.374	0.047
Grade[6]	-0.888	-1.084	-0.704	-0.857	-1.322	-0.386	-0.434	-1.095	0.244
Grade[7]	-0.698	-0.884	-0.515	-0.672	-1.175	-0.185	-0.281	-0.886	0.342
Grade[8] (reference)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Degree of curvature	-0.032	-0.066	0.003	-0.048	-0.131	0.035	-0.050	-0.132	0.029
Three road lanes	-0.484	-0.620	-0.346	-0.509	-0.846	-0.157	-0.435	-1.119	0.321
Median width	-0.007	-0.010	-0.003	-0.006	-0.015	0.003	-0.012	-0.027	0.002
k (dispersion parameter)	-	-	-	0.418	0.305	0.561	-	-	-
Sd(Φi): Spatial correlation	-	-	-	-	-	-	0.469	0.297	0.710
Sd(θi): site-specific random effect	-	-	-	-	-	-	0.584	0.481	0.686
α	-	-	-	-	-	-	0.441	0.330	0.560
pD: no of effective variables	11.9	-	-	117.3	-	-	122.3	-	-
DIC	1903	-	-	1456	-	-	1468	-	-
R^2 (with error terms)	0.61	0.59	0.62	0.88	0.86	0.90	0.88	0.86	0.90
R^2 (without error terms)	-	-	-	0.52	0.32	0.60	0.39	0.02	0.56

3.5.2 Interpretation of Risk Factors

Road alignment factors, i.e. slope and curve, are the other key variables of interest. Preliminary analysis on the data indicates that more than 85% of the total crashes occurred on steep grades (Grade <-2% or >2%). Steep grades are often considered implausible in design, and all design manuals recommend avoiding or keeping minimal the use of steep slopes. Nevertheless, this is not the case with mountainous terrain highways since the steep grades cannot be easily avoided. Longitudinal slope comes out to be significant as indicated in Table 3-2. The effects of various slopes are compared to *Grade[8]* (reference condition, steep slope ranges from -6% to -8%). Figure 3-5 shows the slope coefficients and their 95% credible intervals, it can be noted that in order, *Grade[8]* is the most hazardous slope followed by *Grade[4]*, *Grade[7]*, *Grade[2]*, *Grade[6]*, *Grade[3]*, *Grade[5]* *then Grade[1]*.Generally, trends in the results indicate that the steeper the slope, the higher the crash risk; and segments with upgrade slope are safer than corresponding downgrades in the same slope range. These results are consistent with the preliminary analysis and complementary to existing findings that the steep grades may increase the likelihood of crash occurrence (Shankar et al., 1995; Chan and Chen 2005).

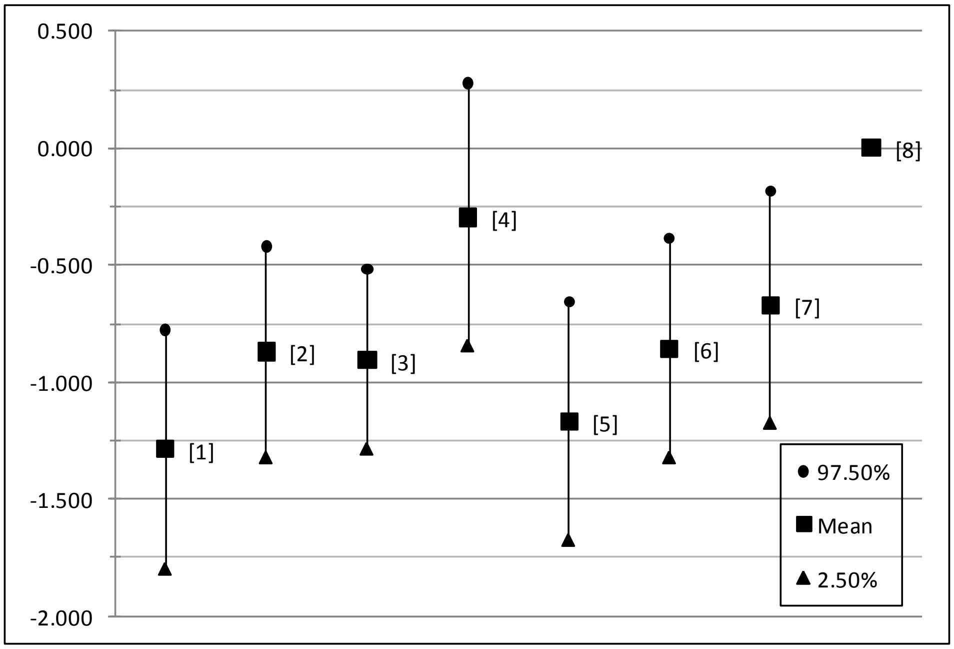


Figure 3-5: Grade Coefficients

In regard to the curve effect, although not statistically significant, the result implies that a unit increase in *Degree of Curvature* (β = -0.048, 95%CI(-0.131,0.035), IRR = 0.95) is associated with a 5% decrease in the crash risk, with all other factors equal. Actually, it is not uncommon that high degree of curvature was found to be associated with decrease in crash likelihood (Shankar et al., 1995; Anastasopoulos et al., 2008; Change and Chen, 2005). Previous studies argued that the feeling of danger along sharp curves might make the drivers compensate by driving more cautiously, leading to lower crash rate instead.

Other variables included in the models are *Number of Lanes* and *Median Width.* Results revealed that segments with three lanes (β =-0.509, 95%CI(-0.846, -0.157), IRR = 0.6) are 40% less in crash risk than two-lane segments, with all other factor being equal. This finding conforms to the study by Park et al. (2010). The increase of safety due to the increase in number of lanes is plausible since this freeway has a high percentage of trucks which could be confined to the 2 right lanes providing more space for other vehicles, contributing to easier maneuvers and less speed variance. Median width is associated with a tiny positive effect (β = -0.006, 95% CI (-0.015, 0.003), IRR = 0.99), which is only significant in the Poisson model. The increasing safety associated with wide median is well known as median works as division for traffic in opposite directions and a recovery area for out-of-control vehicles (Anastasopoulos et al., 2008; Shankar et al., 1998).

3.5.3 Ranking of Sites

The ranking of sites is important to enable officials to pay more attention to those sites with high crash risk. Sites can be ranked by the probability that a site is the worst or by posterior distribution of ranks (Tanaru, 2002). The separate rankings for dry and snow seasons were produced based on the estimation on λ_{it} , the estimated rankings are presented graphically in Figures 3-6 and 3-7. The results confirmed that sites with steep grades are drastically affected during snow season and those segments received significantly higher risk ranks than in the dry season. Moreover, an overall site ranking is developed by rating the weighted average of crash expectations in the two seasons (λ_{i1} for dry season and λ_{i2} for snow season), i.e.,

$Safety\_Site_i = 0.42 \times \lambda_{i1} + 0.58 \times \lambda_{i2}$ to offset the unbalanced design of seasons (5 month for dry season and 7 month for snow season) as explained in the model specification section.

For illustration, the overall site rankings for the 84 segments are plotted on the longitudinal profile for eastbound and westbound, as shown in Figures 3-6 and 3-7, respectively. Sites with high rank values are more dangerous while sites with low rank values are safer. The results appear to be in good agreement with results from the preliminary analysis that the steep downgrade sections received the high risk ranks in general. The segments at Eisenhower tunnel seem to be safer in both east and west bounds. However, the segments just before and after the tunnel received relatively high rank on the eastbound. On the westbound, the downgrade segments received most of the high ranks.

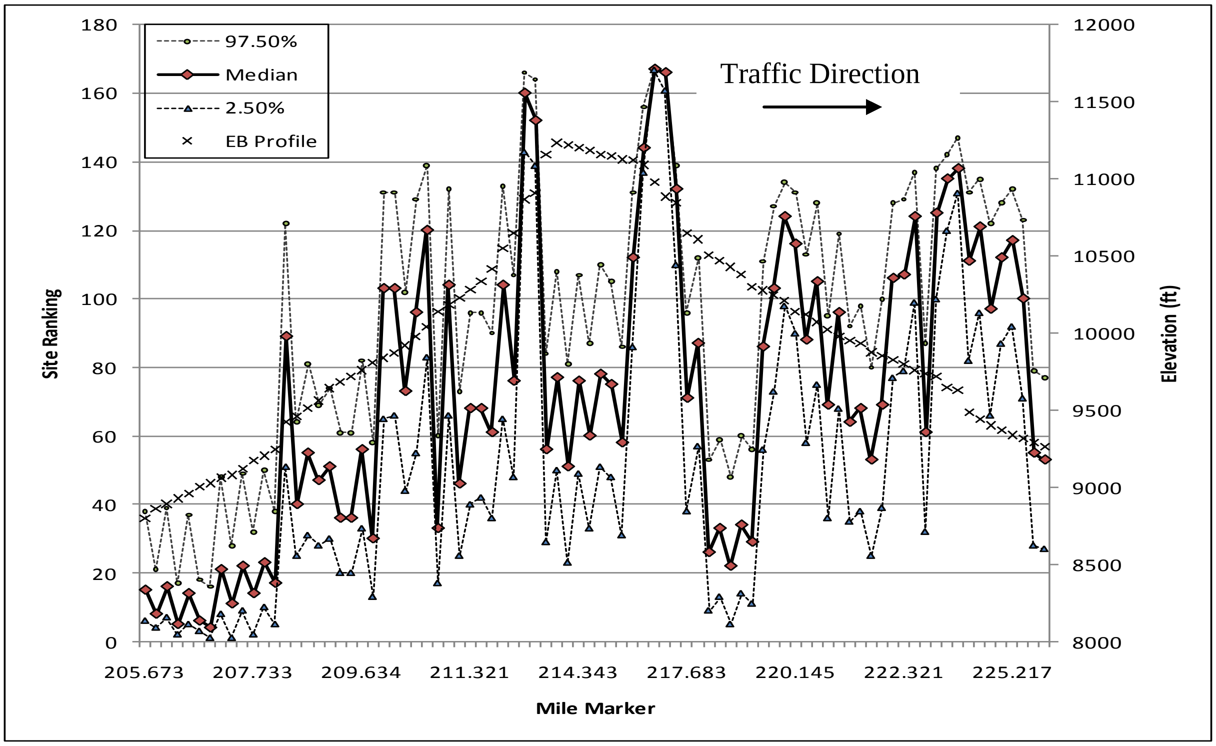


Figure 3-6: East-Bound Segment Ranking

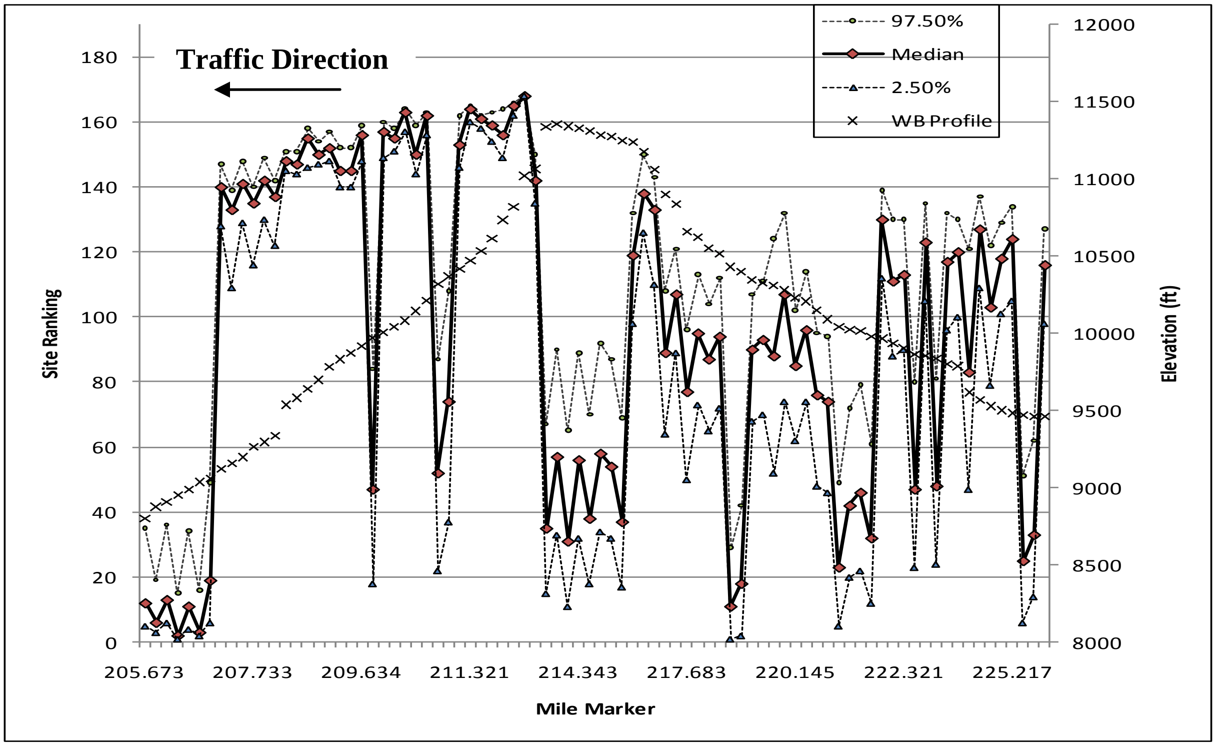


Figure 3-7: West-Bound Segment Ranking

3.6 Conclusion

This chapter presents an exploratory investigation of the safety problems of a mountainous freeway section of unique weather condition. Hierarchical Full Bayesian models were developed to relate crash frequencies with various risk factors associated with adverse weather, road alignments and traffic characteristics. Using the calibrated model, the sites were ranked in term of crash risk for further safety diagnostics and mitigation.

In modeling, it was found that while the random effect and spatial models outperform the Poisson model, the spatial model may have the problem of redundantly accounting for the geometry-dependant effect. Therefore the random effect model is selected for model inference.

Crash risk during snow season was estimated to be approximately 82% higher than the crash risk in dry seasons. Results also identified clear trends associated with the effect of slopes, i.e. the steeper the slope, the higher the crash risk; and segments with upgrade slope are safer than downgrades in the same slope range. The degree of curvature is negatively correlated with crash risk, which is consistent with previous studies that some visual variation of the road alignment may help with drivers' alertness increase and hence decrease crash risk. Median width and number of lanes also showed to be effective in affecting crash risk. Segments with three lanes are 40% less in crash risk than two-lane roads.

Based on site ranking, segments succeeding the tunnel in both east and west bounds received the highest rank of hazardous sites. These segments feature steep slopes and reduction in number of lanes for the east bound. In particular sites with steep slopes should receive more attention from officials and decision makers during snow season to control the excess of crash rate during this season. Also, the identified sites could be included in the strategy for choosing the location of future Variable Speed Limits.

CHAPTER 4. THE VIABILITY OF USING AVI DATA IN REAL-TIME RISK ASSESSMENT

4.1 Introduction

The main objective of this chapter is examining the viability of using Automatic Vehicle Identification (AVI) data for real-time safety risk assessment.

Traffic detection technology is the main spine of any Intelligent Transportation System (ITS); there are a wider range of vehicle detection devices in use than ever before on highways, starting from the popular inductive loops and magnetometers to video and radar-based detectors. It is known that the history of loop detector extends to 50 years ago when it was first developed in 1960s, the inductive loop detectors have become the most widely utilized sensors in traffic management systems.

The inductive loop detector remained unchallenged for more than 30 years because of its simple design, until less intrusive detection options became technologically advanced enough to offer a relief from some of the inherent challenges of the loop detectors.

The main problem of the loops is the reliability, since loop detectors tend to fail due to the very hard environment of the pavement, the temperature variation, and the resulted shifts in the pavement which can break the wires and the loop detector would no longer be functioning. According to the Traffic Detector Handbook (2006), the actual loop detectors failure rates differ from agency to agency because of the large number of variables that contributes to the failure.

This failure rate found to be consistent with failure rate literature for different states and it varies from between 24% and 29% at any given time. The secondary problem of the loop detectors is the maintenance, since cutting into pavement to repair the defective loops may shorten the lifetime of the pavement or result in pavement damage. Moreover, maintenance sometimes is limited or not possible on congested roadways.

During the last decade, new non-intrusive detection devices were deployed as alternatives to inductive-loop detectors such as video, microwave and laser radar, passive infrared and ultrasonic and acoustic sensors. Nowadays, non-intrusive detection devices improved in terms of accuracy, cost and ease of use. The installation and maintenance are relatively easy than the loop detectors since the non-intrusive detection devices can be mounted above or alongside the roadway and hence enhance and increase the reliability. While the inductive loops are expected to continue to function for several years, many transportation agencies seem to be shifting attention to non-intrusive alternatives.

The AVI is among other systems such as satellite positioning and mobile communications using GSM/GPRS that contributed in the advancement of the Electronic Toll Collection (ETC) systems by first introducing the dedicated ETC lanes where the vehicles slow down into channeled toll lanes and recently the express ETC lanes operated at highway speeds also known as Open Road Tolling (ORT). Open Road Tolling with ETC technology nowadays are widely utilized worldwide to automate the payment process, increase system throughput and reduce congestion, improve customer service, enhance safety, apply congestion pricing, increase toll revenues and reduce environmental impacts. ETC systems are composed of Automatic Vehicle Identification (AVI) that determines the ownership of the vehicle to be charged to the corresponding customer, Automatic Vehicle Classification to charge different fair rates to different vehicle types, and Video Enforcement Systems to capture images of the violator and/or license plate that pass through the ETC lanes without a valid transponder. The structure of the ETC systems depends on two main factors; 1) the tolling system and 2) the number of access points on the freeway in case of travel time estimation is incorporated within an ATIS system. It is worth mentioning that the spacing between access points is about 1 mile or less for urban freeways and can exceed 3 miles for rural ones. Before ETC systems, there were three main tolling systems; 1) the "closed ticket system," 2) the "closed barrier system" and 3) the "open barrier system". The advent of the new ETC systems changed the way toll roads are designed and operated. ETC systems have the ability to easily support other value-added services on the same technology platform. These services might include but not limited to fleet and engine management systems, emergency response services, congestion pricing, pay-as-you-drive insurance services and navigation capabilities. The aspect of tolling (a distance-base, a flat-rate or a congestion-base) and the type of facility and access (freeway, expressway, or conventional road) play an important role in the structure and the spacing of the tag readers.

The Central Florida Expressway System utilizes Automatic Vehicle Identification (AVI) system for Electronic Toll Collection (ETC) as well as for the provision of real time information to motorists within the ATIS. This system estimates the segment travel time by monitoring the successive passage times of vehicles equipped with E-Pass, O-Pass or Sun-Pass, electronic Radio

Frequency Identification (RFID) tags at expressway Open Road Tolling (ORT) plazas as well as at exits. Data are gathered using AVI tag readers that are installed for the purpose of toll collection and additional tag readers installed solely for the purpose of estimating travel times. It is worth to mention that there are no specific guidelines for the design of the ETC systems in the U.S.

Commonly deployed inductive loop detectors (ILDs) measure time-mean-speed (TMS), whereas AVIs measure space-mean-speed (SMS). TMS is defined as the arithmetic mean of the speed of vehicles passing a point during a given time interval. Hence, TMS only reflects the traffic condition at one specific point. On the other hand, SMS which is defined by Gerlough and Huber, 1975 as "the mean of the speeds of the vehicles traveling over a given length of road and weighted according to the time spent traveling that length" (there are several definitions of SMS depending on how it is calculated; the mentioned definition is the best to describe the AVI's SMS). Since not all the vehicles are equipped with the transponders, the accuracy of travel time estimation would depend on the percentage of the vehicles that are equipped with the transponders. The penetration of E-Pass users reached above 80% on Central Florida's expressway system. While traffic flow data collected from ILDs were a good safety measure in real-time proactive safety management, data collected from AVI have not been investigated before in any safety related study.

As discussed in the review of literature chapter that a great effort has been performed in analyzing real-time data collected from inductive loop detectors in safety framework, there are no safety analysis studies have been carried out using traffic data from one of the most growing surveillance system; the tag readers on toll roads (AVI). In this study, for the first time, the identification of freeway locations with high real-time crash potential has been examined using real-time speed data collected from AVIs. A stratified matched case-control logistic regression is used to classify the real-time traffic conditions measured by AVI into either leading or not leading to a crash. Matched case-control is used to control for the variability of different factors such as crash site, time, season, day of the week, etc. To select significant variables associated with the crash vs. no-crash target variable, Random Forest (RF) is utilized. Random Forest showed robustness in variable selections recently in transportation studies due to its stability over using single decision tree (Abdel-Aty et al. 2008 and Harb et al. 2008)

4.2 Description of Roadway Network

4.2.1 General Description

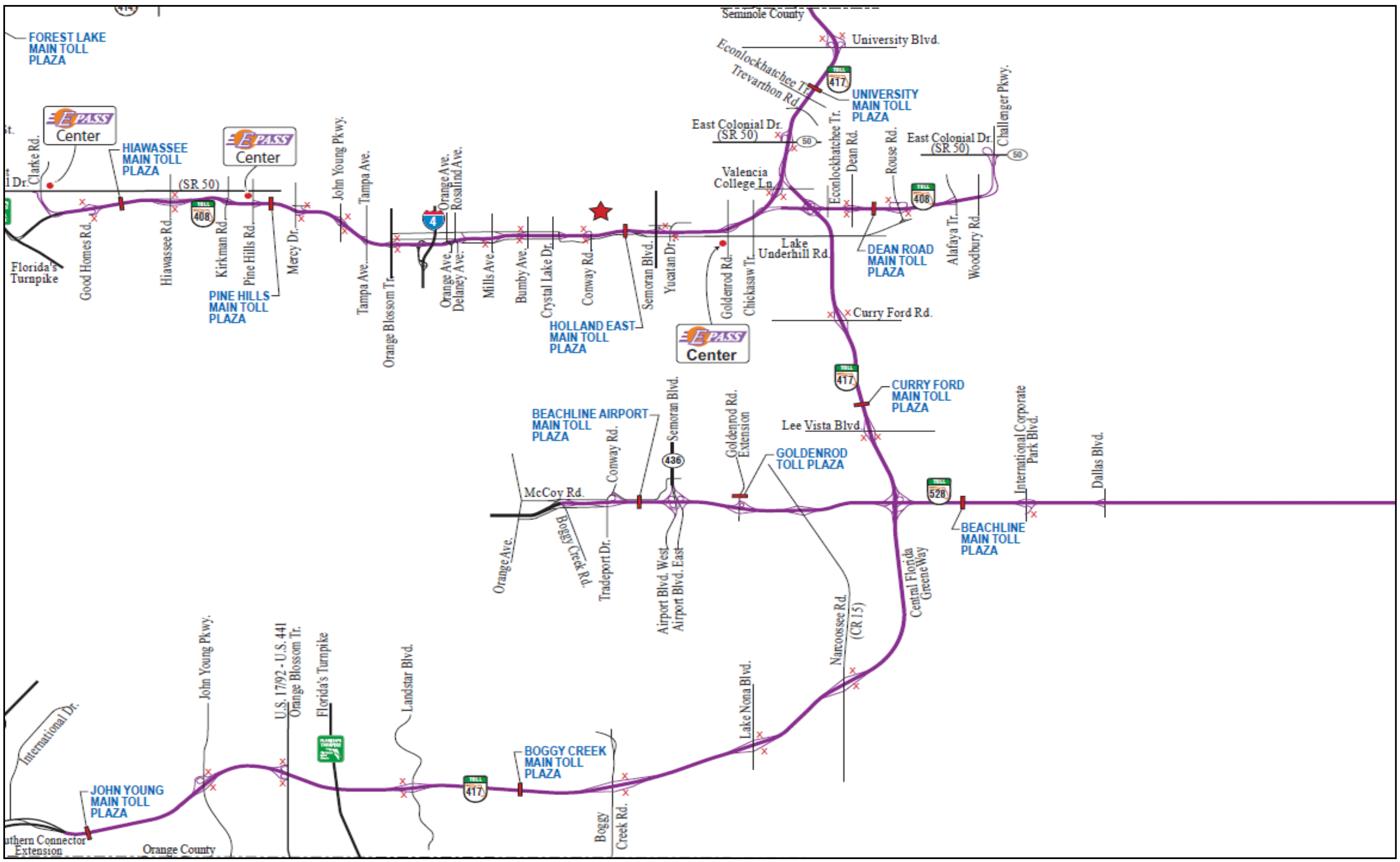
The network studied is about 78 miles of freeways consisting of three toll roads in Orlando, Florida. State Road 408 (SR408), SR417 and SR528. SR408 is nearly 23-mile that extends from Florida’s Turnpike in west Orlando to Challenger Parkway in the east. Traffic on SR408 is almost commute traffic since it connects the east and the west of Central Florida, and passes through the down town area. SR417 and SR528are 33-mile and 22-mile, respectively. SR417 connects Sanford to East Orlando with high percentage of non-commuters travelling between the Orlando-Sanford International Airport, the Orlando International Airport and the attraction areas, however it also includes many commuters from North Orlando State Road 528 provides a crucial connection for residents and tourists between the attractions area, the Orlando International

Airport and the East Coast beaches and Cape Canaveral. As mentioned earlier Central Florida's expressways are equipped with an AVI system for toll collection and travel time estimation, in the study, Figure 4-1 illustrates the expressway network as well as the AVI segments, the AVI segment tag readers are spaced according to toll plazas locations and location of exits of interest to provide the travel time.

Table 4-1 provides summary statistics of the AVI segments on each of the studied freeways, SR408 has 23 AVI segments on the eastbound and 24 on the westbound of average length of 0.9-mile, SR417 has 21 AVI segments on both directions while SR528 has 8 and 9 AVI segments on the eastbound and westbound, respectively, SR528 has longer AVI segments that vary from 1.07-mile to 7.56-mile with an average length of approximately 3 miles.

Table 4-1: Summary Statistics for AVI Segments

Freeway			Automated Vehicle Identification Segments				
State Road ID		Length (mile)	Number of AVI Segments	Length (mile)			
				Min.	Mean	Max.	S.D.
SR408	EB	23	23	0.15	0.92	2.31	0.56
	WB		24	0.14	0.88	2.28	0.55
SR417	NB	33	21	0.21	1.49	2.98	0.75
	SB		21	0.25	1.46	2.87	0.70
SR528	EB	22	8	1.27	2.96	7.56	2.24
	WB		9	1.07	2.80	7.56	2.20



(Source: OOCEA System's Toll Facility Reference Manual)
Figure 4-1: Expressway Network in Orlando.

4.3 Data Description and Preparation

There were two sets of data used in the study; expressway AVI archived data from SR408, SR417 and SR528 in Orlando and the corresponding crash data for year 2008. The Orlando-Orange County Expressway Authority (OOCEA) archives and maintains only the processed 1-minute space mean speed and the estimated average travel time along the defined road segments. The unprocessed original time stamps of the tag readings are not available; this data is typically discarded after the travel time is processed due to privacy issues. The crash data were obtained from the road crash database maintained by FDOT for year 2008.

The crashes have been assigned on each segment; three upstream segments and three downstream segments were identified to be considered in the preliminary analysis. The first upstream and downstream segments were named US1 and DS1, respectively. The subsequent upstream segments were named US2 and US3, respectively while the subsequent segments in downstream direction were named DS2 and DS3, respectively. The data structure is illustrated in Figure 4-2.

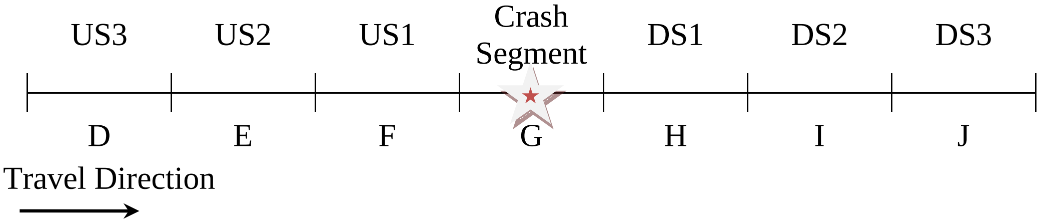


Figure 4-2: AVI Segment Scheme

AVI data corresponding to each crash case were extracted in the following process; for example a crash occurred on February 7, 2008 (Thursday) at 2:00PM, SR408 eastbound, the crash segment G was identified using Geographic Information System (GIS) software, in addition to other six segments (three in the upstream and three in the downstream directions) from 1:30PM to 2:00PM (30 minutes). Five randomly non-crash cases were also determined for the same location and time for different Thursdays where no crashes were observed within 1 hour of the original crash time.

The extracted 1-minute speed data were aggregated to different aggregation level of 2, 3, 5, and 10 minutes to investigate the best aggregation level that will give better accuracy in the modeling part. Five-min aggregation level was found to be the best in terms of statistical fit and model accuracy. The 30 minutes speed data were divided into six time slices, time slice 1 represents the period between the crash time and 5-min prior to the crash time until time slice 6 which represents the interval between 25min and 30-min prior to the crash occurrence. Time slice 1 was discarded in the analysis since it will not provide enough time for successful intervention to reduce crash risk in a proactive safety management strategy. Moreover, the actual cash time might not precisely known, Golob and Recker (2004) discarded the 2.5 minutes of traffic data immediately preceding each crash’s reported time to avoid uncertainty of the actual crash time. In general with the proliferation of mobile phones and CCTV cameras on Freeways, crash time is almost usually immediately identified.

In the modeling part; letters were assigned to each segment in accordance with the crash location to define the location of the crash segment with respect to the upstream/downstream segments. The assigned letters are D, E, F, G, H, I, and J with G being the segment that the crash occurred on, segments F, E, and D are in order the closest segments to the crash segment in the upstream direction while segments H, I, and J are in order the closest segments to the crash segment in the downstream direction as illustrated in Figure 4-2.

Average speeds, standard deviations of the speed and logarithm of coefficient of variation of the speed were calculated over the 5-min time intervals. The nomenclature takes the following form XYS_Zβ. XY takes the value of AV, SD, or CV for average, standard deviation or coefficient of variation, respectively. S stands for speed. Z represents AVI segments and takes values of D to J while β takes the values from 2 to 6 which refer to the time slices.

Unlike ILDs data which are known to suffer from high percentage of missing observations or bad reading, AVI data have less than 5% missing observations with no unreasonable values of speeds. The missing data for the speed were imputed by preserving the distribution of the original data and then the coefficient of variation was calculated. The final data set had a total of 105 variables consisting of 3 speed parameters for each of the 7 AVI segments at 5 time intervals (time slices).

To examine the effect of short-term turbulence of traffic speed only; crashes involving driving under influence of alcohol or drugs and distraction related crashes were excluded from crash data set. A total number of 670 crashes were considered in the analysis and 2680 non-crash cases; Table 4-2 provides the number of crash/non-crash cases used in the study for the studied freeways.

Table 4-2: Number of Crashes on Freeway Corridors

State Road ID		**Number of crash cases**	**Number of non-crash cases**
SR408	EB	180	720
	WB	160	640
	Both Directions	340	1360
	Total	1700	
SR417	NB	96	384
	SB	69	276
	Both Directions	165	660
	Total	825	
SR528	EB	82	328
	WB	83	332
	Both Directions	165	660
	Total	825	
Sub Total		670	2680
Total Observation		3350	

4.4 Methodology

4.4.1 Random Forest and Important Variable Selection

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho (1998), and Amit and Geman (1997) in order to construct a collection of decision trees with controlled variation. RF has the capability of handling thousands of variables without deletion or deterioration of accuracy. Using ensembles of predictors for classification has proved to give more accurate results than the use of a single predictor. Moreover, RF has an advantage over the traditional classification trees of obtaining unbiased error estimates with no need for a separate cross-validation-test data set, when a particular tree is grown from a bootstrap sample, one third of the training cases are left out and not used in the growing of the tree, the left-out cases are called out-of-bag (OOB) data. Abdel-Aty et al., and Harb et al. showed that RF may be used as a robust data mining technique to determine important variables in the transportation field.

The basis of the RF algorithm is first to choose the number of the trees to grow and the number of m variables that would be selected to split each node to produce stable results and minimum out-of-bag (OOB) error rate. The OOB error rate depends on two main components; the correlation between any two trees in the forest and the strength of each individual tree in the forest. The correlation between any two trees in the forest increases the error rate; where as increasing the strength of the individual trees decreases the forest error rate. Reducing m reduces both the correlation and the strength and increasing it increases both. Somewhere in between is an optimal range of m that can be found using OOB. Alternatively, a default value of the number of the candidate variables that will be selected randomly at each split m can be used for classification $m=(p)^{1/2}$ where p is total number of variables. RF is monitoring the error rate for observations left out of the bootstrap sample out-of-bag (OOB) for each grown tree on a bootstrap sample. Fig. 3 shows the OOB error rate against different tree numbers, it is noted that 1000 trees is enough to achieve a constant minimum error rate and hence produce stable estimates.

Using the package “randomforest” in the “R Software”, the RF model was estimated; using m=6 variables that randomly sampled as candidates at each split, the OOB error rate was found to be minimum of 0.183 and 65.24 % of variance explained by the model. Important variables selection based on the mean decrease Gini ‘IncNodePurity’, as the node purity value increase the importance of the variable increase (Kuhn et al. 2008).

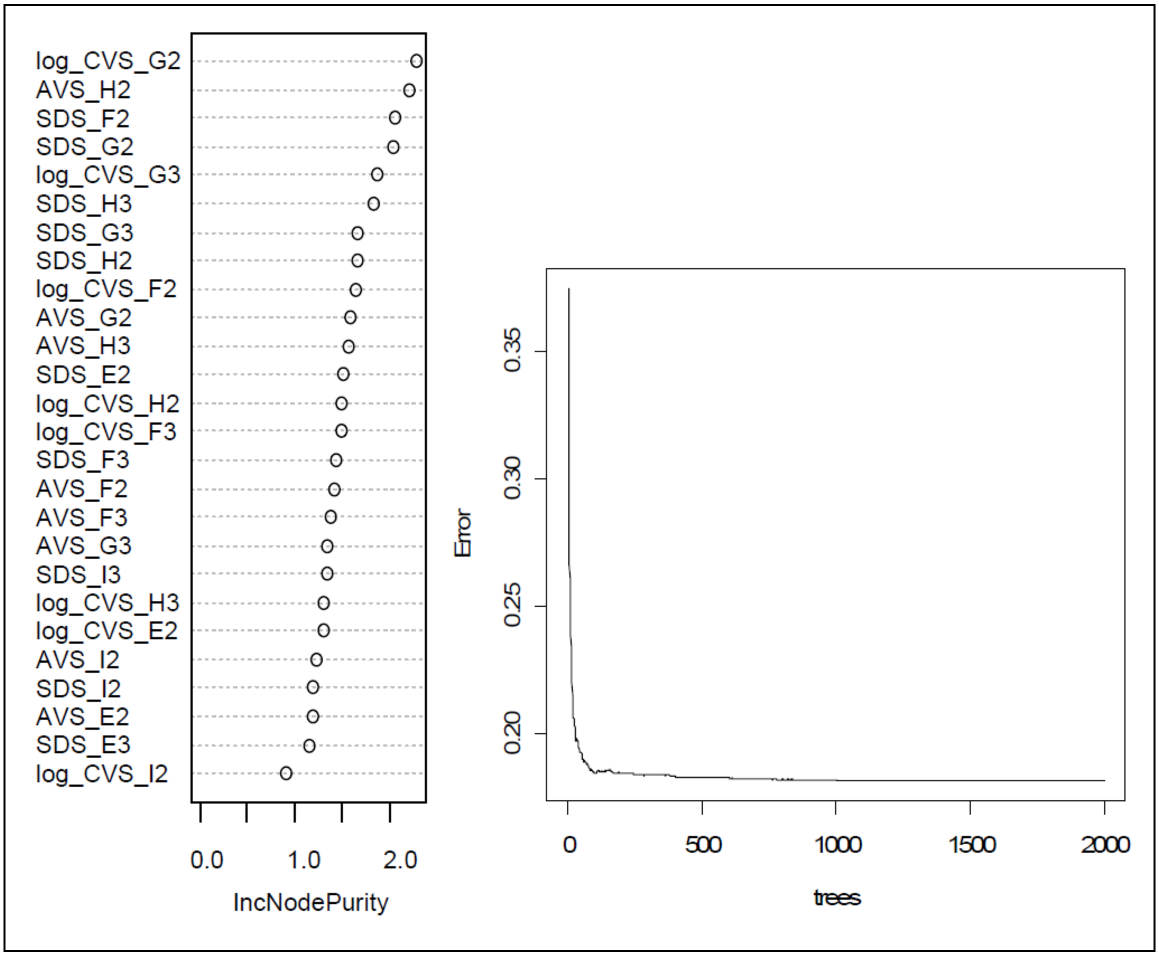


Figure 4-3: Variable Importance and OOB Error.

Examining RF with each data set for the three roadway corridors, most of the important variables were related to the segment that the crash occurred on, first upstream and downstream segments for SR408 and SR417, while SR528 did not return any reasonable results, SR408 and SR417 showed similar results in variable selection. Therefore, the combined data were considered in the final run.

Figure 4-3 shows the important variables from the RF produced for the combined data of SR408 and SR417 in both directions. The logarithm of coefficient of variation of the speed at the crash segment G at time slice 2 from 5 to 10 minutes before the crash time (log_CVS_G2), average speed on downstream segment H in time slice 2 (AVS_H2) and the standard deviation of speed of the upstream segment between 5 to 10 minutes before the crash (SDS_F2) were found to be the most important variables according to Node Purity.

Hence, only variables related to the crash segment and the nearest upstream and downstream segments were included in matched case-control modeling procedure.

4.4.2 Matched Crash-Non-Crash Analysis

The study design utilized a matched case-control methodology, a simple and robust way of examining the crash precursors accounting for confounding factors such as time of crash, seasonal effect and location including all related geometric characteristics. Case-control studies are expected to provide more accurate results as they eliminate confounding factors by matching (Breslow and Day, 1980). For each selected crash case, a randomly selected m controls (non-crash cases) were selected on account of matching factors of location, time of day, day of week, and season (Orlando has 2 distinct weather seasons and matched non crash cases are taken from the same season for each crash case). Different m: 1 ratios have been examined, m=4 was found to give slightly better results. Previous studies show that negligible power is gained through adding controls beyond 3 to 1 matching (Breslow and Day, 1980). Finally the matched set (stratum) was formed of m (4) +1 observations. The modeling is performed under the conditional likelihood principle of statistical theory accounting for within stratum differences between crash and non-crash speed parameters. Use of the conditional likelihood eliminates the parameters associated with the covariates used for matching (e.g. crash time and location).

Matched case-control studies are based upon the classical prospective logistic regression model, with binary outcome Y (case-control status), covariate (X), stratum level N. Suppose that there are N stratum with 1 crash and m non-crash cases in stratum j, where j = 1, 2, 3 …… N. The probability p_j (x_{ij}) that the ith observation in the jth stratum being a crash; where the vector of k speed parameters $x_1, x_2, \ldots\ldots, x_k$ can be noted as $x_{ij} = (x_{1ij}, x_{2ij}, \ldots x_{kij})$; i = 0, 1, 2 …….m and j =1, 2,….N. This crash probability may be modeled by the following linear logistic regression model as described in a study by Abdel-Aty et al. (2004):

$$Logit\{P_j(X_{ij})\} = \alpha_j + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \ldots\ldots\ldots + \beta_k X_{kij} \qquad (4.1)$$

The logistic regression model for matched case-control studies differs from unmatched studies in that it allows the intercept to vary among the matched units of cases and controls. The intercept α summarizes the effect of variables used to form strata on the crash probability and it is different for different stratum.

In order to account of the stratification in the analysis, a conditional likelihood is constructed. It should be noted that the crash probabilities cannot be estimated using Equation (4.1) since the conditional likelihood function L (β) is independent of the intercept terms α1, α2, ………., αN and hence, the effects of matching variables cannot be estimated. This conditional likelihood function is expressed as follows:

$$L(\beta) = \prod_{j=1}^{N}\left[1 + \sum_{i=1}^{m} \exp\left\{\sum_{u=1}^{k} \beta_u (x_{uij} - x_{u0j})\right\}\right]^{-1} \qquad (4.2)$$

However, the values of β parameters that maximize the conditional likelihood function given by Equation (4.2) are also the estimates of β coefficient in Equation (4.1). These estimates are log odds ratio and may be used to approximate the relative risk of a crash.

In this analysis, procedure PHREG in SAS 9.2 is utilized. PHREG provides the hazard ratio which is another term for relative risks used in SAS. In addition, a prediction model can be developed using the log odds ratios under this matched crash-non-crash analysis. This can be demonstrated by considering two observation vectors $x_{1j} = (x_{11j}, x_{21j}, x_{31j}\ldots\ldots., x_{k1j})$ and $x_{2j} = (x_{12j}, x_{22j}, x_{32j}\ldots\ldots., x_{k2j})$ from the jth strata on the k speed parameters. Using Equation (1), the log odds ratio of crash occurrence due to speed parameters vector x1j relative to traffic flow vector x2j will have the following form:

$$\log\left\{\frac{p(x_{1j})/\left[1-p(x_{1j)})\right]}{p(x_{2j})/\left[1-p(x_{2j)})\right]}\right\} = \beta_1(x_{11j}-x_{12j})+\beta_2(x_{21j}-x_{22j})+......+\beta_k(x_{k1j}-x_{k2j}) \qquad (4.3)$$

The right hand side of Equation (4.3) is independent of αj and can be calculated using the estimated β coefficients. Thus, the above relative log odds ratio (left hand side of Equation (4.3)) may be utilized for predicting crashes by replacing X_{2j} with the vector of values of the traffic flow variables in the jth stratum of non-crash cases. One may use simple average of all non-crash observations within the stratum for each variable. Let $\bar{x}_{2j} = (\bar{x}_{12j}, \bar{x}_{22j}, \bar{x}_{32j}, ..., \bar{x}_{k2j})$ denote the vector of mean values of non-crash cases of the k variables within the jth stratum. Then the log odds ratio of crash relative to non-crash cases may be approximated by the following equation:

$$\log\left\{\frac{p(x_{1j})/\left[1-p(x_{1j)})\right]}{p(\bar{x}_{2j})/\left[1-p(\bar{x}_{2j)})\right]}\right\} = \beta_1(x_{11j}-\bar{x}_{12j})+\beta_2(x_{21j}-\bar{x}_{22j})+......+\beta_p(x_{k1j}-\bar{x}_{k2j}) \qquad (4.4)$$

And hence, the log odds ratio can be used for predicting crashes by establishing a threshold value that attain the desirable crash classification accuracy.

As mentioned earlier, important variables were found to be related to the crash segment and two adjacent segments in the upstream and downstream directions at time slice 2 and 3 according to the results obtained in RF. These 18 variables only of AVS, SDS, and CVS were considered for further analysis using the matched case-control.

4.5 Results and Discussion

In the preliminary analysis, a model was built for the combined datasets for all freeway sections. A univariate analysis was conducted first to check the significance of each variable. Different automatic search techniques of stepwise, forward and backward were attempted to identify significant variables in multivariate analysis. These procedures were implemented to identify which terms were still statistically significant in the presence of other factors. Since variables not significant at 0.05 may still be associated with the response after adjusting for other covariates, any variable with $P < 0.25$ in the univariate results were considered eligible to enter into the multivariate model. There was an agreement between the three search techniques that the log of the coefficient of variation of speed of the crash segment at time slice 2 (Log_CVS_G2) is the only significant variable. This variable has positive beta coefficient, which mean that the odds of a crash increase as the variation in speed increase and the average speed decrease at the segment of the crash at 5-10 minutes before the crash occurrence. Table 4-3 shows the hazard ratio for the significant variable. Hazard ratio is the exponent of the beta coefficient and it represents an estimate of the expected change in the risk ratio of having crash versus non-crash per unit change in the corresponding factor, the hazard ratio of 1.234 means that the risk for a crash increases 1.234 times for each unit increase in Log_CVS_G2. It should be noted that the hazard ratio is multiplicative in nature for the continuous variables, this means that a two units increase in Log_CVS_G2 changes the risk by 1.234^2 or 1.52.

Table 4-3: Overall Model Estimates and Fit Statistics

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Log_CVS_G2	1	0.21018	0.08901	5.5763	0.0182	1.234

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	3255.499	3249.915
AIC	3255.499	3251.915
SBC	3255.499	3256.253

Since the combined data sets were collected from different populations, it was worth investigating each of the three freeway corridors separately. Therefore, other models were developed for each of the three freeways individually; univariate as well as multivariate analysis using automatic search techniques have been conducted.

All speed parameters related to SR528 were found to be statistically insignificant. It is worth mentioning that using toll tag readers to estimate travel times introduces a delay in generating observed travel times, for example if a travel time of T minutes is observed, then that travel time applies to a vehicle that entered the segment T minutes ago. The length of the AVI segment plays a significant role in the space mean speed estimation, for example if a number of vehicles entered a segment of 1 mile length, then it should be expected to have them exit the segment within 1 minute in a normal traffic condition given that the speed is 60 mph, on the other hand if the length of AVI segment is 7 miles then the estimated travel time applies to vehicles that entered the segment 7 minutes ago. Moreover, during times of rapid change in the segment travel time, this delay on long segments can reduce the usefulness of AVI data since the estimated measures will not be able to capture the variation in the space mean speed. In particular, this delay may mean that toll tag readers along long segments are ineffective tools for incident prediction.

Table 4-4: SR408 Model Estimates and Fit Statistics

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Log_CVS_G2	1	0.27305	0.11513	5.6254	0.0177	1.314

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1536.143	1530.482
AIC	1536.143	1532.482
SBC	1536.143	1536.310

The final model for SR408 resulted in one significant variable as shown in Table 4-4: LogCVS_G2 (log of the coefficient of variation of speed) from segment G (crash segment) at time slice 2 (5-10 minutes before the crash). The variable has positive beta coefficient, which means that the odds of a crash increase as the variation of the speed increase at the crash segment. This also could be explained that on average of one mile AVI segment, the increase of the standard deviation coupled with decrease of the average speed 5-10 minutes before the crash (since the coefficient of variation of speed includes the standard deviation as the nominator and the average speed as the denominator) may increase the likelihood of crash occurrence. This indicates an increase in the turbulence of traffic. The hazard ratio is found to be 1.314 which means that the crash risk increases 1.314 times for each unit increase in Log_CVS_G2. Moreover, the hazard ratio increased from 1.234 in the overall model to 1.314, this indicates that the risk for a crash increased by 8% for each unit increase in Log_CVS_G2 when SR528 and SR417 data sets were excluded from the model.

Table 4-5 provides the estimates and fit statistics for the model for SR417; two variables came out to be significant: SDS_G2 and AVS_H2. Standard deviation of speed of the crash segment at time slice 2 has a positive beta coefficient while the average speed of the adjacent downstream segment at time slice 2 has a negative beta coefficient. This means that high variation in the speed at the crash segment with decrease in the average speed in the downstream segment may increase the risk of having crash at this location. Decrease in speed downstream might represent queue build up.

The results from both models suggest that the real-time crash prediction models are not transferable from one road to another due to the differences in the driver population as well as the structure of the AVI system; it is noteworthy that both roads have different type of road users as stated before in the data description part. However, transferability might be possible for roadways with similar AVI system spacing and population, these findings were depicted by Pande et al. (2011), although the data they used were collected from very similar loop detector structure in Central Florida (I-4 and I-95), they found that it may not be advisable to use the same model for two freeways with different driver population or travel pattern.

Table 4-5: SR417 Model Estimates and Fit Statistics

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
SDS_G2	1	0.12163	0.05649	4.6357	0.0313	1.129
AVS_H2		-0.05683	0.02336	5.9183	0.0150	0.945

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	654.827	643.355
AIC	654.827	647.355
SBC	654.827	653.295

In order to implement the estimated model in real-time application, sensitivity analysis is conducted. Table 4-6 and 4-7 show sensitivity and the specificity for the final models. Sensitivity is the proportion of crashes that are correctly identified as crashes while specificity is the proportion of non-crashes that are correctly identified as non-crashes by the model (Agresti, 2001). The sensitivity and the specificity can be calculated using the odds ratio given by Equation (4.4). For example, the mean of the two variables SDS_G2 and AVS_H2 of all 4 non-crash cases for SR417 model were calculated within each matched set. The estimated vector of these non-crash means replaced the vector in Equation (4.4) for the j^{th}-matched set. The odds ratio can be estimated by utilizing the beta coefficients from the model in Equation 4 where the vector is the actual observation in the data set. The sensitivity was found to be 67.94% and 69.09% while the two models achieved specificity of 53.53% and 54.85% for SR408 and SR417, respectively at a threshold equal to 1. The classification accuracy is considered good for all crash types, and the accuracy would be expected to increase when evaluating specific crash types (Pande and Abdel-Aty, 2006).

Both models have relatively high false positive rates, at threshold of 1, about 46% were classified as crashes incorrectly while the false negative rates were low, about 32% of crashes were classified as non-crashes. Different classification accuracy can be obtained by changing the threshold depending on the management strategy. The threshold should be chosen carefully in the real-world application; large number of false alarms might affect the drivers' compliance with the system and hence reduce the effectiveness of the system. Nevertheless, Advanced Traffic Management (ATM) objectives of reducing turbulence to improve operation can still be achieved even with high percentage of false alarms. ITS strategies such as variable speed limits could be introduced without the drivers' knowledge of false alarm or not.

Table 4-6: Classification Results SR408

		SR408		
Frequency Percent Row % Column%		Predicted		Total
		0	1	
Actual	0	728 42.82 53.53 86.98	632 37.18 46.47 73.23	1360 80.00
	1	109 6.41 32.06 13.02	231 13.59 67.94 26.77	340 20.00
Total		837 49.24	863 50.76	1700 100.00

Table 4-7: Classification Results SR417

		SR417		
Frequency Percent Row % Column%		Predicted		Total
		0	1	
Actual	0	362 43.88 54.85 87.65	298 36.12 45.15 72.33	660 80.00
	1	51 6.18 30.91 12.35	114 13.82 69.09 27.67	165 20.00
Total		413 50.06	412 49.94	825 100.00

4.6 Conclusion and Recommendations

While the most common application of the AVI is electronic toll collection and travel time estimation, there is a promising traffic safety application in the context of ATM. This study implemented for the first time data collected from the AVI in a real-time traffic safety analysis. AVI data were found to be promising in providing a measure of crash risk in real-time. The operation-based management of expressways can benefit from the collected AVI traffic data not only to ease the congestion and enhance the operation but also by providing warnings of increase risk situation on the crash risk measures identified in this study to increase safety on freeways and expressways.

Travel time and space mean speed data, collected from tag readers (AVI) of a total of 78-miles on the Central Florida expressway network in Orlando in 2008. Historical crash data were collected for the same period and study road sections. Utilizing random forest for significant variable selection and stratified matched case-control to account for the confounding effects of the location and time, the log odds of crash occurrence may be obtained and hence a proactive safety management system may be incorporated with existing ATIS.

The estimated speed collected from the AVIs is different than the one collected from ILDs, AVIs measure the arithmetic mean of the speed of all the vehicles occupying a given length at a given instant whereas ILDs measure the arithmetic mean of the speed of vehicles passing a point during a given time interval. Therefore, the AVI segment length plays an important role in estimating the space mean speed that will be used in any traffic safety management strategy. On one hand, the results suggest that the AVI data could only be useful if the AVI segments are within 1.5 mile on average, on the other hand, it has been found that the model is not easily transferrable from one road to another unless the AVI structure and driver population are similar. The coefficient of variation in speed at the crash segment during 5-10 minutes prior to the time of the crash is found to be the most significant factor affecting the crash likelihood on a freeway with tag readers spaced 1-mile on average and mostly commute drivers while the standard deviation of the speed at the crash segment and the average speed at the adjacent downstream segment were found to be the most significant on another freeway section with AVI segments length of an average of 1.5-mile with mixed type of road users.

All speed parameters obtained from AVIs spaced on average at 3-mile apart were found to be statistically insignificant to identify crash prone conditions. Although, this study shows that AVI segments within 1.5 mile may be useful in real-time crash analysis, further investigation is needed to determine the exact cut off and threshold values of the appropriate length of the AVI segment in order to be used as a guideline in ITS applications.

CHAPTER 5. REAL-TIME RISK ASSESSMENT FOR SPECIFIC CRASH TYPE (ALL CRASHES VS. REAR-END CRASHES)

5.1 Introduction

As discussed in the previous chapter and reviewed from the literature (Chapter 2) that Intelligent Transportation Systems (ITS) rely heavily on detection systems to collect data that are essential to manage traffic, ease congestion and provide motorists with travel time information. In the last decade, traffic safety studies showed that traffic safety could be incorporated in real-time traffic management systems as well as providing warnings of the increase in risk situation to promote safety on freeways and expressways (Madanat and Liu, 1995; Hughes and Council, 1999; Oh et al., 2001;Lee et al. 2002, 2003; Golob and Recker, 2001; Abdel-Aty et al. 2004, 2005, 2007, 2008; Pande and Abdel-Aty 2006a, 2006b; Hourdos et al. 2006). These efforts have been devoted to statistically link real-time traffic conditions to crash occurrence. Most of this real-time crash prediction research attempted the use of data collected from inductive loop detectors (ILDs) (Hughes and Council, 1999; Oh et al., 2001;Lee et al. 2002, 2003; Golob and Recker, 2001; Abdel-Aty et al. 2004, 2005, 2007, 2008; Pande and Abdel-Aty 2006a, 2006b), however, there is a lack of traffic safety studies that investigated data collected from Automatic Vehicle Identification (AVI) Systems (Ahmed and Abdel-Aty, 2011; Ahmed et al. 2011).

It is difficult to delineate from fundamental notions of time mean speed and space mean speed the measure of safety risk without detailed analyses and hence, better understanding of these systems is essential in the safety context. Key question therefore are whether AVI can be used to predict crash risk in real time, as demonstrated in the previous chapter and concluded by Ahmed and Abdel-Aty, 2011; and Ahmed et al., 2011 that AVI data are useful in real-time risk assessment, another questions arises of what level of accuracy could be achieved for specific type of crash vs. predicting all crashes, and if that prediction performance can be improved by targeting the specific single most frequent type of crashes, the rear-end collisions. Rear-end collisions are one of the frequently occurring types of crashes on freeways and expressways (National Highway Traffic Safety Administration, 2007). Their impact on operation is the most noticeable since most of them occur during congested time periods (Abdel-Aty et al., 2005).

In this chapter, a generic semi-parametric Bayesian matched case-control model was calibrated for all crash types and another model for rear-end crashes. We investigate also if prior knowledge about the covariates from previous years at the same location can provide better fit and enhance the capability of the model to predict crashes more accurately. In order to examine this approach as in real-life applications; one year of data (2007) were used to calibrate the model using classical (frequentist) matched case-control logistic regression, then the coefficients estimates were used as prior in Bayesian Matched Case-Control to update the coefficients using another year of data (2008) and different year of data from (2009) were used for validation.

Unlike other studies that were limited by the availability of data in which the sensitivity analysis was carried out using the same data that were used to calibrate the model, in this study we use a separate dataset for validation and scoring the model.

5.2 Data Collection and Preparation

The expressway section under consideration is 33 miles long of SR417 where AVI data were available. Central Florida’s expressways are equipped with an AVI system for toll collection and travel time estimation; there are 22 AVI tag readers along the 33-mile section on both directions with an average spacing of 1.47-mile.

There were two sets of data used in the study; expressway AVI archived data from SR417 in Orlando and the corresponding crash data for three years 2007 through 2009. The Orlando-Orange County Expressway Authority (OOCEA) archives and maintains only the processed 1-minute space mean speed and the estimated average travel time along the defined roadway segments. The unprocessed original time stamps of the tag readings are not available; this data are typically discarded after the travel time is processed due to privacy issues. The crash data were obtained from the Crash Analysis Resource (CAR) maintained by FDOT for the same years.

As shown in the previous chapter and illustrated by Ahmed and Abdel-Aty, 2011 that crash occurrence was mostly related to the AVI crash segment, one segment in the upstream and another segment in the downstream directions and therefore these segments were considered in the data extraction process and modeling parts. The crashes have been assigned on each segment; upstream and downstream segments were identified to extract their corresponding AVI data. The upstream, crash, and downstream segments were named U, C and D, respectively. The AVI segment scheme is illustrated in Figure 5-1.

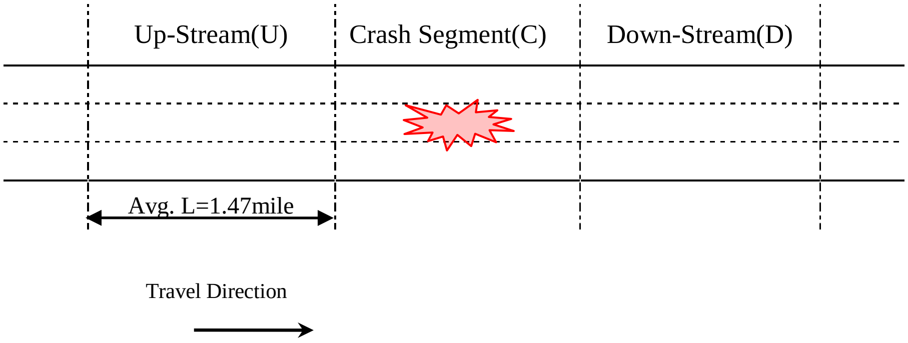


Figure 5-1: AVI Segment Scheme

AVI data corresponding to each crash case were extracted in the following process; for example a crash occurred on February 7, 2008 (Thursday) at 2:00PM, SR417 eastbound, the crash segment C was identified using Geographic Information System (GIS) software, in addition to other two segments (one in the upstream and one in the downstream directions) from 1:30PM to 2:00PM (30 minutes). Four non-crash cases were also determined for the same season (to control for weather conditions), location and time for different Thursdays. It is worth mentioning that the crash and the none-crash cases were only extracted where no crashes were observed within 1 hour of the original crash time at the same AVI segment. There were 4 crashes that occurred within the crash segment few minutes after the first crash; these crashes were not considered because all speed parameters would be affected by the first crash event.

As discussed earlier, the extracted 1-minute speed data were aggregated to different aggregation level of 2, 3, 5, and 10 minutes to investigate the best aggregation level that will provide better accuracy in the modeling part. Five-min aggregation level was found to provide better statistical fit (smaller DIC) and relatively higher classification accuracy. The 30 minutes speed data were divided into six time slices, time slice 1 represents the period between the crash time and 5-min prior to the crash time until time slice 6 which represents the interval between 25min and 30-min prior to the crash occurrence. Time slice 1 was discarded in the analysis since it would not provide enough time for successful intervention to reduce crash risk in a proactive safety management strategy. Moreover, the actual crash time might not be precisely known, Golob and Recker, 2001 discarded the 2.5 minutes of traffic data immediately preceding each crash reported time to avoid uncertainty of the actual crash time. In general with the proliferation of mobile phones and CCTV cameras on expressways, crash time is almost usually immediately identified.

Average speeds, standard deviations of the speed and logarithm of coefficient of variation of the speed (standard deviation of speed divided by the average speed) were calculated over the 5-min time intervals. The measure notations take the general form XY_Zβ. Where XY takes the value of AV, SD, or CV for average, standard deviation or coefficient of variation of speed, respectively. Z represents AVI segments and takes values of U, C, and D for upstream, crash, or downstream segments while β takes the values from 2 to 6 which refer to the time slices.

Unlike ILDs data which are known to suffer from high percentage of missing observations or bad reading, AVI data have less than 5% missing observations with no unreasonable values of speeds. The missing data for the speed were imputed by preserving the distribution of the original data and then the coefficient of variation was calculated. The final dataset had a total of 45 variables consisting of 3 speed parameters for each of the 3 AVI segments at 5 time intervals (time slices).

Although crashes involving driving under the influence of alcohol or drugs and distraction related crashes were less than 2% of total crashes, they were excluded from the crash dataset to examine the effect of short-term turbulence of traffic speed only. Hence, the analysis presented in this study is based on 447 total crashes in which 171 were rear-end crashes.

5.3 Methodology

5.4 Bayesian Updating Approach

This study utilizes the Bayesian semi-parametric Cox proportional hazards model (PHM) to explain the relationship between an event (crash) occurring at a given time and a set of risk factors in matched case-control design and to mainly control for the confounding factors of time, location, and season. Cox PHM model is used commonly for survival analysis; an important distinction in survival analysis is how the time-dependency in the event process (the baseline hazard in the absence of any covariate effects) is parameterized. Cox's semi-parametric model assumes a parametric form for the effects of the covariates, but it allows an unspecified form for the baseline hazard. Therefore, Cox PHM can be utilized regardless of whether the survival time is discrete or continuous. The Cox PHM is performed with the SAS® (BAYES PROC PHREG) (SAS Institute Inc., 2011) by forming a stratum for each matched set, a dummy variable for the survival time is created in the dataset such that all the crash cases in a matched set have the same event time value, and the corresponding non-crash cases (controls) are censored at the later times.

The classical Cox's semi-parametric model estimates the coefficients of parameters solely based on the information from the observed data whereas the Bayesian Cox's semi-parametric makes use of the combined information of the prior as well as the observed data to estimate the parameters' coefficients. In the Bayesian framework, the data is used to update beliefs about the behavior of the parameter to assess its distributional properties as well as possible. PROC PHREG with BAYES option generates a Markov chain that contains the approximate posterior distribution samples by Gibbs sampler, using the adaptive rejection sampling algorithm (Gilks et al. 1995 and Gilks and Wild, 1992). The DIC, a Bayesian generalization of AIC, is used to measure the model complexity and fit. The Deviance Information Criterion DIC, a Bayesian generalization of Akaike Information Criterion AIC, is used to measure the model complexity and fit. DIC is a combination of the deviance for the model and a penalty for the complexity of the model. The deviance is defined as $-2\log$ (likelihood). The effective number of parameters, pD, is used as a measure of the complexity of the model, $\mathrm{pD} = \mathrm{Dbar} - \mathrm{Dhat}$, where Dbar is the posterior mean of the deviance, and Dhat is a point estimate of the deviance for the posterior mean of the parameters. DIC is given by DIC = Dhat + 2 pD (Spiegelhalter et al. 2003). Moreover, receiver operating characteristic (ROC) curve analysis was used to assess the prediction performance. In addition, a sensitivity analysis is conducted to measure the accuracy of each of the estimated models using different validation dataset from year 2009.

5.5 Results and Discussion

5.5.1 Model Estimation and Diagnostics (All Crashes vs. Rear-End Crashes)

As mentioned earlier, frequentist matched case-control model was estimated for all crashes that occurred in 2007 on the expressway section, the dataset is comprised of 690 observations (138 crash cases and 552 non-crashes (control)). With prior knowledge of the likely range of values of the parameters from 2007, informative priors were specified for parameters for all crashes that occurred in 2008 (165 crashes and 660 non-crashes) to avoid using the same data in the updating process. It is noteworthy to mention that using non-informative prior in Bayesian estimation resulted in the same estimate obtained from frequentist model. In Bayesian update, one chain of 20,000 iterations were set up in SAS based on the convergence speed and the magnitude of the dataset, before drawing inferences from posterior sample, the trace, autocorrelation and density plots should be examined for each parameter to be content that the underlying Markov chain has converged. Following Brooks-Gelman-Rubin (BGR) convergence diagnostics (Brooks and Gelman, 1998), the trace, autocorrelation, and density plots for the two significant parameters shown in Figure 2 suggest that the mixing in the chain is acceptable with no correlation. After ensuring the convergence, the first 2,000 samples were discarded as adaptation and burn-in.

A univariate analysis was conducted first to check the significance of each variable. Different automatic search techniques of stepwise, forward and backward were attempted to identify significant variables in multivariate analysis. These procedures were implemented to identify which terms were still statistically significant in the presence of other factors. Since variables not significant at 0.05 may still be associated with the response after adjusting for other covariates, any variable with $P < 0.25$ in the univariate results were considered eligible to enter into the multivariate model (SAS Institute Inc., 2011). There was an agreement between the three search techniques that there are two significant variables associated with crash occurrence, Table 5-1 provides the estimates of beta coefficients, credible interval, hazard ratio and fit statistics for the (All Crashes) Model; two variables came out to be significant: SD_C2 and AV_D2. Standard deviation of speed of the crash segment at time slice 2 (5-10 minutes prior the crash time) has a positive beta coefficient while the average speed of the adjacent downstream segment at time slice 2 has a negative beta coefficient. This means that high variation in the speed at the crash segment with decrease in the average speed in the downstream segment may increase the risk of having crash at this location. Decrease in speed downstream might represent queue build up. Hazard ratio is the exponent of the beta coefficient and it represents an estimate of the expected change in the risk ratio of having crash versus non-crash per unit change in the corresponding factor, the hazard ratio of 1.13 means that the risk for a crash increases by 13% for each unit increase in SD_C2. It should be noted that the hazard ratio is multiplicative in nature for the continuous variables, this means that a two units increase in SD_C2 changes the risk by 1.132 =1.28 (28% increase).

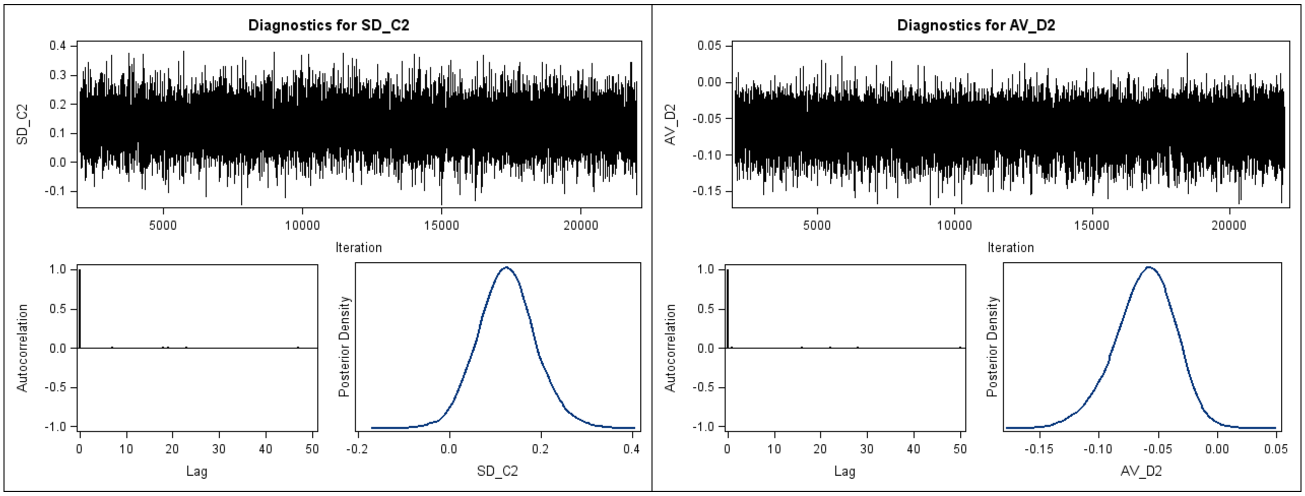


Figure 5-2: Diagnostics Plots (All Crashes Model)

Following the same methodological updating approach as explained before, a Bayesian matched case-control model was estimated for rear-end crashes only from 2008 using informative priors from the frequentist model that was estimated using data for rear-end crashes only from 2007. The dataset for 2007 have 280 observations (56 rear-end crash cases and 224 non-crashes (control)) while the 2008 dataset used to update the model coefficients have total of 305 observations (61 rear-end crashes and 244 non-crashes). Similarly, the convergence was assessed using plots for trace, autocorrelation and density, the model has converged reasonably. Table 5-2 shows the coefficient estimates, credible interval, hazard ratio and fit statistics. SD_C2 and AV_D2 came out to be significant, however, the hazard ratio increased for the standard deviation of speed of the crash segment at time slice 2 for rear-end crashes model by more than twice the hazard ratio for all crashes model while the hazard ratio decreased for the average speed of the downstream segment at time slice 2 by about 20 percent. This may indicate that the increase in variation of the speed at any given segment coupled with decrease in average speed in the downstream segment may result in rear-end crash more than any other type of crashes.

Table 5-1: SR417 (All Crashes 2008) Model Estimates, Hazard Ratio, and Fit Statistics

Posterior Summaries					**Hazard Ratios**			
			Credible Interval				**Credible Interval**	
Parameter	**Mean**	**Standard Deviation**	**2.5%**	**97.5%**	**Mean**	**Standard Deviation**	**2.5%**	**97.5%**
SD_C2	0.1256	0.0639	0.00312	0.2562	1.1362	0.0729	1.0031	1.2920
AV_D2	-0.0614	0.0257	-0.1167	-0.0153	0.9408	0.0241	0.8899	0.9848

Fit Statistics	
DIC (smaller is better)	647.695
pD (Effective Number of Parameters)	2.149

One limitation however in the current AVI archiving system is that the system does not record the percentage of lane change per segment, this percentage can be calculated by developing an algorithm to compare the unique tag ID for each individual vehicle at the beginning and end of each segment. Moreover, the algorithm can process the original raw AVI data in a way that provides space mean speed by lane and hence a better picture can be comprehended about not only the longitudinal speed variation at the AVI segment but also the variation across the lanes. It is to be noted that by having detailed lane speed data may help to identify other types of crashes such as sideswipe and angle crashes.

It should be noted that using the informative prior slightly enhanced the model fit; the DIC decreased from 652.371 to 647.695 for all crashes model and from 111.278 to 106.097 for rear-end crashes.

Table 5-2: SR417 (Rear-End Crashes 2008) Model Estimates, Hazard Ratio, and Fit Statistics

Posterior Summaries					Hazard Ratios			
			Credible Interval				Credible Interval	
Parameter	**Mean**	**Standard Deviation**	**2.5%**	**97.5%**	**Mean**	**Standard Deviation**	**2.5%**	**97.5%**
SD_C2	0.9151	0.3852	0.1986	1.7065	2.6949	1.1318	1.2197	5.5096
AV_D2	-0.2627	0.1520	-0.6147	-0.0313	0.7776	0.1124	0.5408	0.9692

Fit Statistics	
DIC (smaller is better)	106.097
pD (Effective Number of Parameters)	1.611

5.5.2 Classification Accuracy of the Models (All Crashes vs. Rear-End Crashes)

In order to implement the estimated models in real-time application, sensitivity analyses are conducted. Table 5-3 and Table 5-4 show sensitivity and the specificity for the final models. Sensitivity is the proportion of crashes that are correctly identified as crashes while specificity is the proportion of non-crashes that are correctly identified as non-crashes by the model (Agresti, 2002). The sensitivity and the specificity can be calculated using the odds ratio as explained in chapter 4, given by Equation (4.4). For example, the mean of the two variables (SD_C2, standard deviation of speed of the crash segment at time slice 2 (5-10 minutes prior the crash time)) and (AV_D2, average speed of the downstream segment at time slice 2) of all 4 non-crash cases were calculated within each matched set. The estimated vector of these non-crash means replaced the vector in Equation (4.4) for the j^{th}-matched set. The odds ratio can be estimated by utilizing the beta coefficients from the updated model using 2008 dataset in Equation (4.4) where the vector is the actual observation in the 2009 dataset for all crashes and rear-end crashes.

The sensitivities were found to be 69.44% and 72.22% for all crashes and rear-end crashes, respectively using Bayesian matched case-control model with non-informative priors while it increased to 72.92% and 75.93% using Bayesian updating approach with specified informative priors from year 2007. Both models have reasonable false positive rates, at threshold value of unity, about 42% and 46% were classified as crashes incorrectly for all crashes and rear-end crashes, respectively. Different false positive rates can be obtained by changing the threshold depending on the management strategy. The threshold should be chosen carefully in real-world application; large number of false alarms might affect the drivers' compliance with the system and hence reduce its effectiveness. Nevertheless, Advanced Traffic Management (ATM) objectives of reducing turbulence to improve operation can still be achieved even with high percentage of false alarms. ITS strategies such as variable speed limits could be introduced without the drivers' knowledge of false alarm or not.

Table 5-3: Classification Results (All Crashes)

		All Crashes		
Frequency **Percent** **Row Percent** **Col Percent**		**Predicted**		**Total**
		0(Non-Crash)	**1(Crash)**	
Actual	**0(Non-Crash)**	334 46.39 **Specificity** 57.99 89.54	242 33.61 **False Positive Rate** 42.01 69.74	576 80.00
	1(Crash)	39 5.42 **False Negative Rate** 27.08 10.46	105 14.58 **Sensitivity** **72.92** 30.26	144 20.00
Total		373 51.81	347 48.19	720 100.00

Table 5-4: Classification Results (Rear-End Crashes)

		Rear-End Crashes		
Frequency **Percent** **Row Percent** **Col Percent**		**Predicted**		**Total**
		0(Non-Crash)	**1(Crash)**	
Actual	**0(Non-Crash)**	117 43.33 **Specificity** 54.17 90.00	99 36.67 **False Positive Rate** 45.83 70.71	216 80.00
	1(Crash)	13 4.81 **False Negative Rate** 24.07 10.00	41 15.19 **Sensitivity** **75.93** 29.29	54 20.00
Total		130 48.15	140 51.85	270 100.00

5.6 Conclusions

While traffic flow data collected from ILDs were a good safety measure in real-time proactive safety management, there were no studies that attempted the use of AVIs data in real-time safety risk assessment. AVI data were found to provide reasonably comparable measure to ILDs of crash risk in real-time, the operation-based management of expressways can benefit from the collected data not only for toll collection and travel time estimation but also to provide warnings of increased risk situations. Few studies conducted crash prediction by type using real-time traffic data collected on freeways/expressways. In contrast, explicitly in this study we classify and compare the generic model for all types of crashes with specific crash type (rear-end) model using data collected from tag readers (AVI) systems on expressways.

The chapter presents a Bayesian updating framework to identify real-time traffic conditions prone to crashes using expressway AVI data. Using 3 years of crash data and the corresponding AVI data on SR417 in Orlando, a classical (frequentist) matched case-control model was estimated using data from 2007. With prior knowledge of the likely range of values of the parameters from 2007 at the same expressway corridor, informative priors were specified for the parameters in a semi-parametric Bayesian matched case-control framework to avoid using same data in the updating process. This approach was applied one time on all crashes and another time on rear-end crashes. By contrasting AVI data preceding all crash types and rear-end crashes with non-crashes, it is concluded that the hazard ratio increased for the standard deviation of speed of the crash segment at 5-10 minutes before the crash time for the rear-end crash model by more than twice the hazard ratio for the overall crash model while the hazard ratio decreased for the average speed of the downstream segment at 5-10 min before the crash time. This may indicate that the increase in variation of the speed at any given segment coupled with decrease in average speed in the downstream segment may result in rear-end crash more than any other type of crashes.

The classification accuracy for the rear-end crashes model is more than that achieved by the generic all crashes model, 72.22% of the rear-end crashes may be identified correctly while the generic all crashes model identified only 69.44%. Moreover, the proposed Bayesian updating approach showed better fit in the form of relatively lower DIC values using informative priors, also the accuracy of both models increased to achieve 75.93% and 72.92% for rear-end and all crashes, respectively.

The proposed methodology leads to much more efficient estimation of risk than does ordinary frequentist matched case-control logistic regression. Bayesian updating approach is strongly recommended as a robust technique to reduce uncertainty in the parameters and increase the accuracy of the model fit.

Although the AVI system can provide measures about percentage of lane change per segment by comparing the unique tag ID for each individual vehicle at the beginning and end of the segment as well as providing space mean speed for each lane to estimate the variation in speed across lanes, the AVI algorithm and the archiving system in its current form do not report these information and hence the expressway authorities are encouraged to update their archiving system.

This study suggests that AVI data in the current form can provide an acceptable real-time safety risk assessment for all crash types in general and rear-end crashes in particular, and with minor modifications of how tag readers are structured and how the AVI data are processed and archived, it is possible to enhance the prediction accuracy and extend the proposed methodology to other crash types.

CHAPTER 6. INCORPORTAING ROADWAY GEOMETRY AND REAL-TIME WEATHER DATA IN REAL-TIME RISK ASSESSMENT

6.1 Introduction

In previous studies, weather data were estimated from crash reports for crash cases and from airports weather stations in the vicinity of the freeway section for non-crash cases (Abdel-Aty and Pemmanabonia, 2006; Hassan and Abdel-Aty, 2010). It should be noted that none of these studies had access to actual weather information on the roadway section itself. In this chapter, real-time weather data are gathered by weather stations installed on the roadway solely for the purpose of collecting real-time information about the adverse weather conditions. Moreover, roadway geometrics were considered in few studies (Abdel-Aty and Abdalla, 2004; Abdel-Aty et al., 2007), and their effects were controlled for by a matched case-control framework in other studies (Abdel-Aty et al. 2004, 2005, 2007, 2008; Abdel-Aty and Pande, 2004, 2005; Pande and Abdel-Aty 2006a, 2006b; Hassan and Abdel-Aty, 2010; Ahmed and Abdel-Aty, 2011). These studies were mostly conducted on freeways/expressways that feature normal roadway geometry and hence the traffic flow parameters were found to be the most dominant factors that contribute to crash occurrence. Since the roadway section under study features mountainous terrain of relatively steep grades and sharp horizontal curves' radii, the geometrical characteristics were considered to examine how the interaction between all these factors contributes to crash occurrence. This chapter investigates the identification of freeway locations with high crash potential using traffic data collected from AVI, real-time weather information and geometric features.

According to the Federal Highway Administration (Goodwin, 2002), weather contributed to over 22% of the total crashes in 2001. This means that adverse weather can easily increase the likelihood of crash occurrence. Several studies, in fact, concluded that crashes increase during rainfall by 100% or more (Brodsky and Hakkert, 1988; National Traffic Safety Board, 1980), while others finding more moderate (but still statistically significant) increase (Andreescu and Frost, 1998; Andrey and Olley, 1990).

Automatic Vehicle Identification (AVI) system has been widely used in real-time travel time estimation (Tam and Lam, 2011; Dion and Rakha, 2006). While few studies used traffic data from AVI in real-time traffic safety application (Ahmed and Abdel-Aty, 2011; Ahmed et al. 2012a, 2012b), in this study, AVI data, real-time weather data, and roadway geometry are implemented to assess the safety risk on a freeway section that features mountainous terrain.

6.2 Data Preparation

This study involves four datasets; roadway geometry data, crash data, and the corresponding AVI and weather data. The crash data were obtained from CDOT for a 15-mile segment on I-70 for three years (2007 to 2009). Traffic data consists of space mean speed captured by 20 AVI detectors located on each east and west bounds along I-70. We obtained from CDOT the processed 2-minute space mean speed and the estimated average travel time for each AVI segment. Although the tag readers have the capability of collecting lane by lane data, the processed and archived AVI data included only the combined travel time and space mean speed for all lanes. It is worth mentioning that ATIS was developed and implemented without consideration for safety applications. Weather data recorded by three automated weather stations along I-70 for the same time period were also provided by CDOT. The roadway data were collected from Roadway Characteristics Inventory (RCI) and Single Line Diagrams (SLD).

AVI data corresponding to each crash case were extracted in the following process; the location and time of occurrence for each of the 301 crashes were identified. Since the space mean speeds were archived on 2-minute intervals, the speeds were aggregated to different aggregation level of 2, 4, and 6-minute level to obtain averages and standard deviations and to investigate the best aggregation level that will give better accuracy in the modeling part. Six-min aggregation level was found to provide better fit. Three time slices of the 6-minute prior the crash time were extracted. For example if a crash happened on Sep 16, 2007 (Sunday) at 14:00, at the milepost of 205.42. The corresponding 18-min window for this crash of time intervals (13:42 to 14:00) recorded by AVI segment 34 (Mile marker starts at 200.8 and ends at 205.55). Time slice 1 was discarded in the analysis since it would not provide enough time for successful intervention to reduce crash risk in a proactive safety management strategy. Moreover, the actual crash time might not precisely be known. Golob and Recker (2004) discarded the 2.5 minutes of traffic data immediately preceding each reported crash time to avoid uncertainty of the actual crash time. In general with the proliferation of mobile phones and CCTV cameras on Freeways, crash time is almost usually immediately identified. One-hour speed profiles were also generated (about 30 minutes before and 30 minutes after the crash time) to verify the reported crash time. The modeling procedure required non-crash data, a random selection from the whole remaining AVI dataset where there was no crash within 2-hour before the extraction time was utilized in the study to represent the whole population of different traffic patterns, weather conditions and roadway characteristics.

Similarly, weather data for crash cases and non-crash ceases were extracted. Automated weather stations monitor the weather conditions continuously and the weather parameters are recorded according to a specific change in the reading threshold and hence they do not follow a specific time pattern. The stations report frequent readings as the weather conditions change within short time; if the weather conditions remain the same the station would not update the readings. However, these readings were aggregated over certain time periods to represent the weather conditions. For example; precipitation described by rainfall amount or snowfall liquid equivalent for ten minutes, one hour, three hours, six hours, twelve hours and twenty-four hours and the estimated average hourly visibility which provides an hourly measure of the clear distance in miles that drivers can see. Visibility in general can be described as the maximum distance (in mile) that an object can be clearly perceived against the background sky, visibility impairment can be result of both natural (e.g., fog, mist, haze, snow, rain, windblown dust, etc.) and human induced activities (transportation, agricultural activities, and fuel combustion). The automated weather stations do not directly measure the visibility but rather calculate it from a measurement of light extinction which includes the scattering and absorption of light by particles and gases.

A total number of 301 crashes and 880 non-crashes were finally considered in the analysis in which 70 and 231 crashes and their randomly selected 256 and 624 non-crashes occurred during the dry and the snow seasons, respectively.

6.3 Preliminary Analysis and Results

From the preliminary analysis, it can be found that the environmental conditions have a strong effect on crash occurrence within that section. According to the meteorological data, the study section has two distinct weather seasons; dry season from May through September which experience small amount of rain, and snowy season from October through April. The crash frequencies during the snowy season months were found to be more than double the frequencies during the dry season months. Figure 6-1 shows the 3-year aggregated crash frequency by month and weather for the 15-mile freeway section.

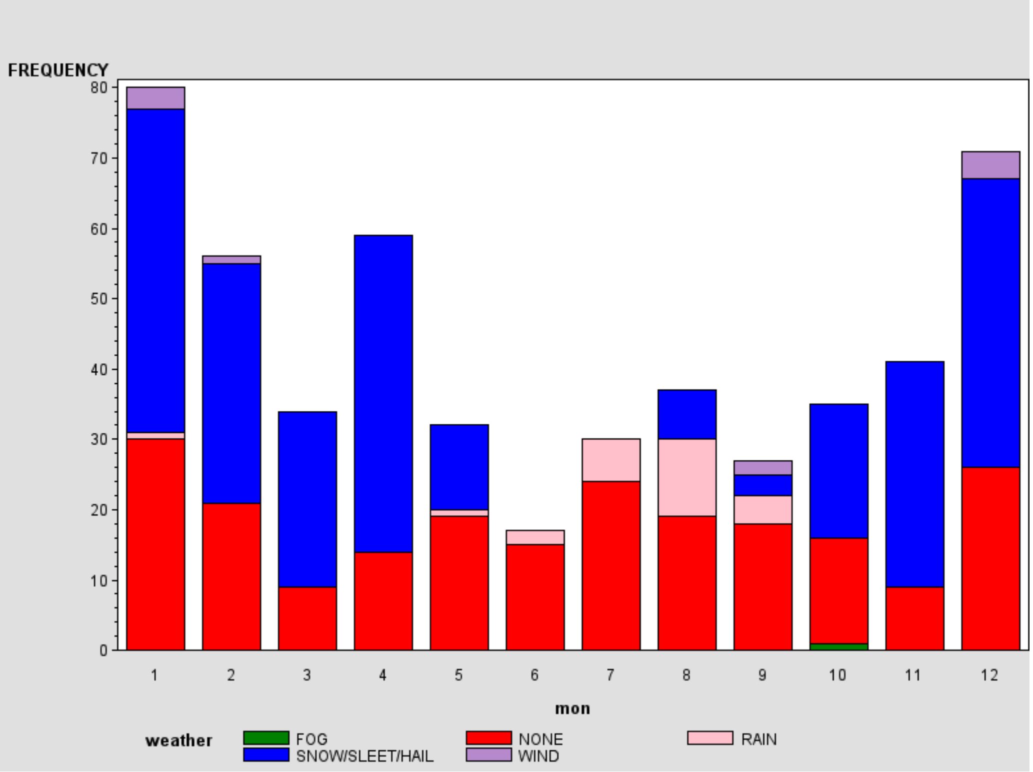


Figure 6-1: Crash Frequency by Month

To compare between the traffic and environmental factors for crash and non-crash cases as well as between snow and dry seasons, a series of statistical tests were conducted. F-test showed that the crash cases and non-crash cases have equal variance and hence t-tests for equal variance were used. The results showed that there is a significant difference between each of the mean of the average speed and the mean of the average 1-hour visibility of crash-cases and non-crash cases. For example, the 6-min average speed 6-12 min prior to the crash cases for both the snowy and the dry seasons was found to be 48.21 mph while it was found to be 55.71 mph prior to the non-crash cases with a resulted t-test p-value of 6.7×10-8. The mean of the estimated visibilities one hour before the crash cases/non-crash cases was found to be significantly higher for non-crash cases than crash-cases, the mean of the estimated visibility for non-crash cases was found to be 1.22 miles while it was found to be 0.95 mile for crash-cases. These results depicts that there is a significant difference between the crash-cases and non-crash cases at the 95% confidence level for the speed and different weather related factors. Similarly, t-tests were used to evaluate weather condition factors in different seasons (dry and snow). The t-test results showed that the dry season had a higher visibility and significantly lower precipitation rate. For visibility, the dry season had a visibility of 1.29 miles while the snow season has 1.09 miles; for ten-minute precipitation, the dry season had a precipitation only as 0.000543 inch while the snow season had 0.057 inch. Average speed for different seasons has also been compared; t-test result shows that in the dry season the average speed is significantly higher than the snow season and with a smaller standard deviation. These observations also suggest that different active traffic management strategies should be implemented for each season.

6.4 Bayesian Logistic Regression

The study utilized a Bayesian logistic regression approach to estimate the probability of crash occurrence in each of the dry and the snow seasons. Bayesian logistic regression has the formulation of a logistic equation and can handle both continuous and categorical explanatory variables. The classical logistic regression treats the parameters of the models as fixed, unknown constants and the data is used solely to best estimate the unknown values of the parameters. In the Bayesian approach, the parameters are treated as random variables, and the data is used to update beliefs about the behavior of the parameters to assess their distributional properties. The interpretation of Bayesian inference is slightly different than the classical statistics; the Bayesian derives updated posterior probability of the parameters and construct credibility intervals that have a natural interpretation in terms of probabilities. Moreover, Bayesian inference can effectively avoid the problem of over fitting that occurs when the number of observations is limited and the number of variables is large.

The Bayesian logistic regression models the relationship between the dichotomy response variable (crash/no-crash) and the explanatory variables of roadway geometry, real-time weather and traffic. Suppose that the response variable y has the outcomes y=1 or y=0 with respective probability p and 1-p. The logistic regression equation can be expressed as:

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta X \qquad (6.1)$$

where β_0 is the intercept, β is the vector of coefficients for the explanatory variables, and X is the vector of the explanatory variables,. The logit function relates the explanatory variables to the probability of an outcome y=1. The expected probability that y=1 for a given value of the vector of explanatory variables X can be theoretically calculated as:

$$p(y=1) = \frac{exp\ (\beta_0 + \beta X)}{1 + exp\ (\beta_0 + \beta X)} = \frac{e^{\beta_0 + \beta X}}{1 + e^{\beta_0 + \beta X}} \qquad (6.2)$$

One advantage of the Bayesian approach over the classical model is the applicability of choosing the parametric family for prior probability distributions. There are three different priors that can be used; 1) informative prior distributions based on the literature, experts' knowledge or explicitly from an earlier data analysis, 2) weak informative priors that do not supply any controversial information but are strong enough to pull the data away from inappropriate inferences, or 3) uniform priors or non-informative priors that basically allow the information from the likelihood to be interpreted probabilistically. In this study, uniform priors following normal distribution with initial values for the estimation of each parameter from the maximum likelihood method was used. Different types of prior distributions using the results from this study as prior could be considered for further research once more data become available to update the estimated models.

As discussed earlier in the preliminary section that Colorado has two distinct weather seasons and hence two models for the snow and dry seasons were considered, these models were estimated by Bayesian inference using the freeware Winbugs (Lunn et al., 2000). For each model, three chains of 10,000 iterations were set up in Winbugs based on the convergence speed and the magnitude of the dataset. The Deviance Information Criterion DIC, a Bayesian generalization of Akaike Information Criterion AIC, is used to measure the model complexity and fit. DIC is a combination of the deviance for the model and a penalty for the complexity of the model. The deviance is defined as $-2\log\ (\text{likelihood})$. The effective number of parameters, pD, is used as a measure of the complexity of the model, $\text{pD} = \text{Dbar} - \text{Dhat}$, where Dbar is the posterior mean of the deviance, and Dhat is a point estimate of the deviance for the posterior mean of the parameters. DIC is given by DIC = Dhat + 2 pD (Spiegehalter et al., 2003). Moreover, receiver operating characteristic (ROC) curve analysis was used to assess the prediction performance.

6.5 Results and Discussion

6.5.1 Model 1 (Dry Season)

The dry season model was estimated using real-time weather, AVI data and roadway geometry for crashes that occurred during May to September for years 2007 through 2009 and the randomly selected non-crashes with their corresponding data. Before drawing inferences from posterior sample, the trace, autocorrelation and density plots were examined visually to ensure that the underlying Markov chains have converged. Following Brooks-Gelman-Rubin (BGR) convergence diagnostics (Brooks and Gelman, 1998), the mixing in the chains was found to be acceptable with no correlation for all included variables in the final model. After ensuring the convergence, first 2,000 samples were discarded as adaptation and burn-in. Table 1 provides the estimates of beta coefficients, credible interval, hazard ratio and fit statistics for the (Dry Season) model; all included roadway alignment factors, i.e. median width, longitudinal grade and horizontal curve were found to be significant. Preliminary analysis on the data indicates that more than 85% of the total crashes occurred on steep grades (grade <-2% or >2%). Steep grades affect the operation and the braking of the vehicles on both upgrade and downgrade, the results indicates that the crash likelihood increases as the grade increases, the effect of various grades are compared to Grade[Flat] (reference condition, flat grade ranges from 0% to ±2%). It can be noted that in order, Grade[Very Steep] (grade (>6% to 8%)and(<-6% to -8%)) is the most hazardous followed by Grade[Steep] (grade (>4% to 6%) and (<-4% to -6%)), and Grade[Moderate] (grade (>2% to 4%) and (<-2% to -4%)). Generally, trends in the results indicate that the steeper the grade, the higher the crash risk. Table 6-1 shows the hazard ratio for the significant variables. Hazard ratio is the exponent of the beta coefficient and it represents an estimate of the expected change in the risk ratio of having crash versus non-crash. The interpretation of the hazard ratio depends upon the measurement scale of the explanatory variable; for interval variables it represents the change in the risk ratio per unit change in the corresponding factor while for categorical variables it represents the change in the risk ratio compared to the base case, e.g. the hazard ratio of 5.63 for the categorical variable Grade[Very Steep] means that the likelihood of a crash at very steep grades is 5.63 times the likelihood at the base case of flat grades Grade[Flat].

A binary variable Grade Index was created to represent the direction of the grade at the crash segment, [1=upgrade] as a reference and [2=Downgrade], the grade index was found to be significant at the 90% credible interval with a positive coefficient which implies that the positive road grades are slightly safer than the negative ones. These results are consistent with the finding from the aggregate models in the literature that the steep grades may increase the likelihood of crash occurrence (Shankar, 1995; Chang and Chen, 2005; Ahmed et al., 2011).

The results imply that the Degree of curvature (β=-0.246, 95%CI(−0.484,-0.024), hazard ratio = 0.78) is significantly associated with crash risk, a unit increase in degree of curvature is associated with 22% decrease in crash likelihood, with all other factors remain constant. High degree of curvature was found to be associated with decrease in crash likelihood in previous studies, it may be explained that the discomfort feeling along sharp curves might make the drivers compensate by driving more cautiously, leading to lower probability of involvement in a crash (29,30,31,32). Median width (β=-0.046, 95%CI(−0.075,-0.019) has a negative coefficient meaning that a wider median is safer since it works as a recovery area for out-of-control vehicles.

The 6-minute average speed of the crash segment during 6-12 minutes prior the crash time as well as the average visibility during the last hour before the crash time were found to be significant during the dry season. Both variables have negative beta coefficients, which mean that the odds of a crash increase as the average speed decreases at the segment of the crash at 6-12 minutes before crash occurrence and the average visibility decreases during one hour prior the crash time. The hazard ratio of 0.926 means that the risk for a crash increases 7.4 percent for each unit decrease in the six minutes average speed, and the hazard ratio of 0.211 means that the risk for a crash increases 79% for each unit mile decrease in the average Visibility measured over one hour before the crash time.

Table 6-1: Parameters and Hazard Ratio Estimates (Dry Season Model)

Variables	**Parameters Estimates**				**Hazard Ratio**			
	Credible interval				Credible interval			
	Mean	S.D.	2.5%	97.5%	Mean	S.D.	2.5%	97.5%
Intercept	2.070	1.37	-0.599	4.830	-	-	-	-
Grade[Flat (0-2)%](reference)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Grade[Moderate >2-4%]	0.510	0.554	-0.565	1.640	1.950	1.210	0.568	5.150
Grade[Steep >4-6%]	1.120	0.485	0.201	2.120	3.470	1.860	1.220	8.330
Grade[Very Steep >6-8%]	1.540	0.604	0.373	2.740	5.630	3.840	1.450	15.600
Grade Index[1=Upgrade](ref.)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Grade Index[2=Downgrade]	0.658	0.354	-0.023	1.350	2.060	0.755	0.977	3.860
Degree of curvature	-0.246	0.116	-0.484	-0.024	0.787	0.091	0.616	0.976
Median width	-0.046	0.014	-0.075	-0.019	0.955	0.014	0.928	0.981
Average Speed	-0.076	0.020	-0.115	-0.037	0.926	0.019	0.891	0.964
Visibility	-1.750	0.636	-3.070	-0.568	0.211	0.141	0.046	0.566
pD: no of effective variables	9.803	-	-	-	-	-	-	-
DIC	297.762	-	-	-	-	-	-	-
ROC	0.783	-	-	-	-	-	-	-
Sensitivity	75.71	-	-	-	-	-	-	-

Summary statistics (Mean, S.D.): Degree of curvature (1.33, 1.49), Median Width (ft) (25.96, 15.11), Average Speed (mph) (56.4, 7.94), and Visibility (mi) (1.29, 0.95).

6.5.2 Model 2 (Snow Season)

Another model was estimated for crash no-crash cases in the snow season to examine whether the same variables have the same effect on crash likelihood as in the dry season. Comparisons between the two models imply very interesting findings. On the one hand, same geometric variables came out to be significant; on the other hand, it is noticeable that all the coefficients increased yielding to the fact that the hazard ratios increase due to the interaction between the snowy, icy, or slushy pavement conditions during snow season, and exacerbated by the steep grades. The hazard ratio for Grade[Very Steep] (grade (>6% to 8%)and(<-6% to -8%)) during snow season increased to 9.67 compared to 5.63 in the dry season which means that the change in risk ratio almost doubled during the snow season. Similar findings were concluded for Degree of Curvature and Median Width. Another interesting observation from the parameter estimate for Grade Index is that the hazard ratio decreased and the variable became insignificant which may indicate that steep grades become hazardous during snow season in both the upgrade and downgrade directions.

While only the 1-hour Visibility was significant in the dry season model, in the snow season model both 1-hour Visibility and the ten-minute Precipitation described by rainfall amount or snowfall liquid equivalent came out to be significant. These results are consistent with the preliminary analysis that the precipitation rates are significantly higher during the snow season than in the dry season, one unit increase in the Precipitation increases the risk of the crash by 165%. Moreover, it can be implied from the results that one unit decrease in the Visibility during the snow season increases the crash likelihood by 88% compared to 79% in the dry season.

Logarithm of the coefficient of variation in speed at the crash segment at time slice 2 (6-12 minutes before the crash) came out to be significant. Log COV Speed has positive beta coefficient, which means that the risk of a crash increases as the variation of the speed increases. The increase in the standard deviation coupled with the decrease in the average speed 6-12 minutes before the crash (since the coefficient of variation of speed includes the standard deviation as the nominator and the average speed as the denominator) may increase the likelihood of crash occurrence.

Table 6-2: Parameters and Hazard Ratio Estimates (Snow Season Model)

Variables	**Parameters Estimates**				**Hazard Ratio**			
	Credible interval				Credible interval			
	Mean	S.D.	2.5%	97.5%	Mean	S.D.	2.5%	97.5%
Intercept	1.596	0.510	0.600	2.541	-	-	-	-
Grade[Flat (0-2)%](reference)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Grade[Moderate >2-4%]	0.820	0.354	0.147	1.533	2.420	0.905	1.158	4.631
Grade[Steep >4-6%]	0.927	0.341	0.279	1.612	2.691	0.952	1.261	4.951
Grade[Very Steep >6-8%]	2.203	0.361	1.533	2.928	9.671	3.730	4.634	18.69
Grade Index[1=Upgrade](ref.)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Grade Index[2=Downgrade]	0.009	0.188	-0.369	0.381	1.031	0.191	0.688	1.456
Degree of curvature	-0.301	0.067	-0.434	-0.175	0.742	0.049	0.648	0.839
Median width	-0.053	0.008	-0.069	-0.038	0.948	0.008	0.933	0.963
Precipitation	0.881	0.418	0.149	1.774	2.652	1.268	1.161	5.892
Visibility	-2.207	0.342	-2.862	-1.533	0.117	0.041	0.057	0.216
Log COV Speed	0.501	0.225	0.056	0.944	1.693	0.388	1.058	2.576
pD: no of effective variables	9.506	-	-	-	-	-	-	-
DIC	802.028	-	-	-	-	-	-	-
Area under ROC Curve	0.84	-	-	-	-	-	-	-
Sensitivity	80.09	-	-	-	-	-	-	-

Summary statistics (Mean, S.D.): Degree of curvature (1.39, 1.52), Median Width (ft) (24.50, 15.45), Visibility (mi) (1.09, 0.47), Precipitation (in) (0.05, 0.29), and Log COV Speed (0.24, 0.38).

In order to implement the estimated model in real-time application, sensitivity analysis is conducted. Tables 6-3 and 6-4 show sensitivity and the specificity for the dry and snow models respectively. Sensitivity is the proportion of crashes that are correctly identified as crashes while specificity is the proportion of non-crashes that are correctly identified as non-crashes by the estimated Bayesian logistic regression models (Agresti, 2002). The sensitivity was found to be 75.71% and 80.09% while the models achieved specificity of 66.41% and 67.79 at cutoff points equal to 0.20 and 0.25 for the dry and the snow seasons, respectively. The cutoff was chosen for each model to reduce the false positive rate; about 33.59% and 32.21% for the dry and snow seasons were classified incorrectly as crashes, respectively.

As mentioned earlier that different classification accuracy can be obtained by changing the threshold depending on the management strategy. The threshold should be chosen carefully for application; large number of false alarms might affect the drivers' compliance to the system and hence reduce the effectiveness of the system. Nevertheless, Advanced Traffic Management (ATM) objectives of reducing turbulence to improve operation can still be achieved even with high percentage of false alarms. False alarm conditions are still non ideal, and reducing the flow turbulence could lead to operation benefits although it might not have lead to a crash. As discussed earlier, ITS strategies such as variable speed limits could be introduced without the drivers' knowledge of false alarm or not.

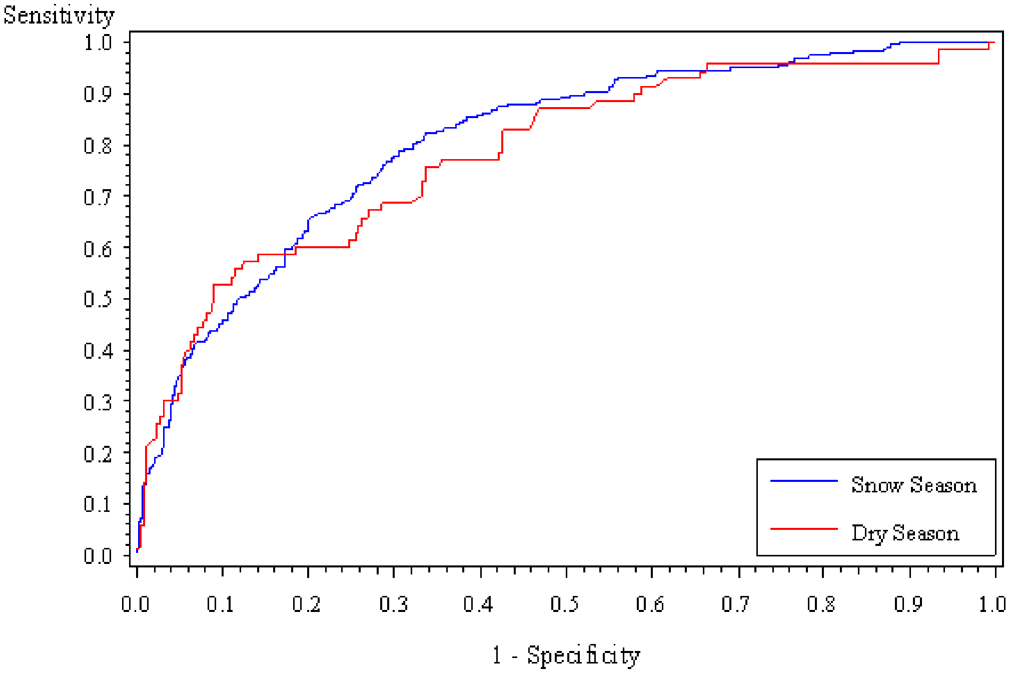


Figure 6-2: Receiver Operating Characteristic (ROC) (Dry and Snow Seasons Models)

The Receiver-Operating Characteristic (ROC) curves were also generated as another way to assess the models' performance. The area under the ROC curve shows how well the model is discriminating between the crash (y=1) and no-crash (y=0) cases in the response variable. This is similar to the misclassification rate, but the ROC curve calculates sensitivity (true positive rate) and 1-specificty (false positive rate) values for many cutoff points. The exact areas under the ROC curves were found to be 0.783 and 0.840 for the dry and the snow seasons, respectively which indicate that the models can provide good discrimination.

Table 6-3: Classification Results (Dry Season Model)

		Dry Season Model		
Frequency **Percent** **Row Percent** **Column Percent**		**Predicted**		**Total**
		0 (Non-Crash)	**1 (Crash)**	
Actual	**0 (Non-Crash)**	170 52.15 **Specificity 66.41** 90.91	86 26.38 **False Positive Rate 33.59** 61.87	256 78.53
	1 (Crash)	17 5.21 **False Negative Rate 24.29** 9.09	53 16.26 **Sensitivity 75.71** 38.13	70 21.47
Total		187 57.36	139 42.64	326 100.00

Table 6-4: Classification Results (Snow Season Model)

		Snow Season Model		
Frequency **Percent** **Row %** **Column%**		**Predicted**		**Total**
		0 (Non-Crash)	**1 (Crash)**	
Actual	**0 (Non-Crash)**	423 49.47 **Specificity 67.79** 90.19	201 23.51 **False Positive Rate 32.21** 52.07	624 72.98
	1 (Crash)	46 5.38 **False Negative Rate 19.91** 9.81	185 21.64 **Sensitivity 80.09** 47.93	231 27.02
Total		469 54.85	386 45.15	855 100.00

6.6 Conclusion

Real-time crash prediction models that depend only on traffic parameters are useful for freeways with normal geometry and at locations that do not encounter severe weather conditions. Most of the previous studies found that the traffic turbulence (e.g., speed variance) defined by the traffic parameters is more dominant to discriminate between crash and non-crash cases and hence the matched case-control design was an adequate technique to account for the small variability in roadway geometry and weather. In this study we illustrate that the same traffic turbulence could affect the driver differently on roadway sections with special geometry and at different weather. Mountainous roadway geometry and adverse weather could exacerbate the effect of traffic turbulence and hence the inclusion of these factors is vital in the context of active traffic management systems.

Although all previous studies used loop detectors data (which provide time mean speed, flow and lane occupancy) we showed in this study that traffic data collected from AVI and real-time weather data were found to provide good measure of crash risk in real-time.

Preliminary analysis on the data and findings discussed in earlier study (Ahmed et al., 2011) indicate that the crash risk during snow season is 82% higher than the crash risk in dry season and hence two models were considered in this study to examine the effect of the interaction between geometric features, weather and traffic data on crash occurrence. While all included geometric factors were significant in the dry and snow seasons, the coefficient estimates indicate that the crash likelihood could be doubled during the snow season because of the interaction between the snowy, icy, or slushy pavement conditions during snow season and the steep grades.

The hazard ratio for the very steep grades (grade (>6% to 8%) and (<-6% to -8%)) during snow season increased to 9.71 compared to 5.63 in the dry season. Same conclusion can be implied for the visibility, reduction of one unit in the visibility was found to increase the crash risk by 88% in the snow season compared to 79% in the dry season. The 10-min. precipitation prior the time of the crash was significant in only the snow season model; one unit increase in the precipitation increases the risk of the crash by 169%. The logarithms of the coefficient of variation in speed at the crash segment during 6-12 minutes prior to the time of the crash is found to be significant in the snow season while the 6-minute average speed at the crash segment 6-12 minutes prior to the crash time was found to be significant in the dry season.

The results from this study suggest that the inclusion of roadway and weather factors in real-time crash prediction models is essential; in particular with roadways that feature challenging roadway characteristics and adverse weather conditions. Also, different active traffic management strategies should be in place during these two distinctive seasons and more resources should be devoted during the snow season.

This study also depicts that traffic management authorities can benefit from the AVI and real-time weather data not only to ease congestion and enhance the operation but also to mitigate increased safety risk.

CHAPTER 7. A DATA FUSION FRAMEWORK FOR REAL-TIME RISK ASSESSMENT ON FREEWAYS

7.1 Introduction

Accurate and reliable estimation of increased risk of crashes is critical to the success of proactive safety management strategies on freeways. In recent years, the advances in electronics have had a tremendous impact on enhancing and improving detection systems, new non-intrusive traffic detection devices are in use more these days because of their easiness of installation and maintenance in addition to their accuracy and affordable cost. Moreover, some freeways have multiple non-intrusive detection systems in place such as the Automatic Vehicle Identification (AVI) and Remote Traffic Microwave Sensor (RTMS). AVI is used mainly for toll collection and for travel time estimation purposes along freeways while RTMS are used mostly for operation and incident management. Research in the field of freeway traffic management has utilized extensively traffic data collected from inductive loop detectors in real-time proactive traffic management (Oh et al., 2001; Abdel-Aty et al., 2004; Abdel-Aty and Pande, 2005; Pande and Abdel-Aty, 2006a, 2006b; Hourdos et al., 2006). Recently, the usefulness of the collected traffic data from AVI has been investigated in real-time safety assessment (Ahmed and Abdel-Aty, 2011; Ahmed et al., 2011, 2012a, 2012b).

Traffic data from AVIs and RTMSs as well as weather data are collected on 15-mile of mountainous Interstate-70 in Colorado to provide roadway users with important information about travel time, congestion, adverse weather conditions and lane closure due to occasional avalanche danger, maintenance on the road and/or road crashes. This information is provided as a part of an Intelligent Transportation System (ITS) and is dynamically disseminated in real time to road users via Dynamic Message Signs (DMS). This system utilizes AVI to estimate the segment travel time by monitoring the successive passage times of vehicles equipped with electronic tags at designated locations. Main traffic flow parameters are collected using RTMS. It is worth mentioning that the AVIs and RTMSs are providing different measures of speeds; AVIs measure space-mean-speed (SMS), which is defined by Gerlough and Huber, 1975 as “the mean of the speeds of the vehicles traveling over a given length of road and weighted according to the time spent traveling that length”, whereas RTMSs measure time-mean-speed (TMS) which is the arithmetic mean of the speed of vehicles passing a point during a given time interval. Hence, TMS only reflects the traffic condition at one specific point. On the other hand, SMS is the average speed of all the vehicles occupying a given stretch of the road over some specified time period (there are several definitions of SMS depending on how it is calculated (Hall, 1996); the definition in this dissertation is the best to describe the AVI’s SMS).

Weather condition is considered one of the most important factors that can contribute to crash occurrences. In previous studies weather data are always estimated from crash reports, in this study real-time weather data are gathered by weather stations located on the roadway section.

Although in previous chapters, it was found that classical statistical models provide interpretable models and acceptable accuracy of crash prediction using AVI and real-time weather data (Ahmed et al. 2011, 2012a); in this study a framework was proposed to augment even more traffic data from multiple sources, weather and geometry data using an advanced machine learning (ML) technique. Machine learning methods are known for their superior performance over the classical statistical ones. In order to enhance the accuracy and increase the reliability of the real-time crash prediction, Stochastic Gradient Boosting (SGB), a recent and promising machine learning technique is attempted to uncover previously hidden patterns preceding a crash relative to non-crash conditions from the large amounts of roadway geometry, weather and AVI and RTMS traffic data.

The following sections illustrate the procedures of preparing the data, modeling technique, interpretation and evaluation, risk assessment framework and the conclusions.

7.2 Data Description and Preparation

There were five sets of data used in this study; roadway geometry data, crash data, and the corresponding AVI, RTMS and weather data. The crash data were obtained from CDOT for a 15-mile segment on I-70 for 13 months (from October 2010 to October 2011). Traffic data consists of space mean speed captured by 12 and 15 AVI detectors located on each east and west bounds, respectively along I-70. Volume, occupancy and time mean speed are collected by 15 RTMSs on each direction. AVI estimates SMS every 2-minute while RTMS provides traffic flow parameters every 30-second. Weather data were recorded by three automated weather stations along the roadway section for the same time period. The roadway data were extracted from Roadway Characteristics Inventory (RCI) and Single Line Diagrams (SLD).

In a previous study (Ahmed and Abdel-Aty, 2011), it was found that crash occurrence was mostly related to the AVI crash segment, one segment in the upstream and another segment in the downstream directions and therefore these AVI segments and their respective RTMS stations were considered in the data extraction process and modeling parts. The crashes have been assigned to the AVI segment and to the closest RTMS station; upstream and downstream AVI segments as well as 3 RTMSs in the upstream and downstream were identified to extract their corresponding traffic data. The upstream, crash, and downstream segments were named U, C and D, respectively while the upstream and downstream RTMSs were named US and DS respectively and assigned numbers in order from the closest to the farthest ones. It is worth mentioning also that most of the RTMSs are located exactly at the same location of the AVIs' tag readers. The arrangement of RTMS and AVI segments and their spacing are illustrated in Figure 7-1.

AVI and RTMS data corresponding to each crash case were extracted in the following process; the location and time of occurrence for each of the 186 crashes were identified. Traffic data were aggregated to 6-minute level to obtain averages, standard deviations, and logarithm of coefficient of variations (standard deviation divided by the average of the traffic parameters) of 2-minute space mean speed obtained from AVIs and 30-second time mean speed, volume, and occupancy raw data obtained from RTMSs. The 6-minute aggregation level was chosen to have consistent time periods between AVIs and RTMSs.

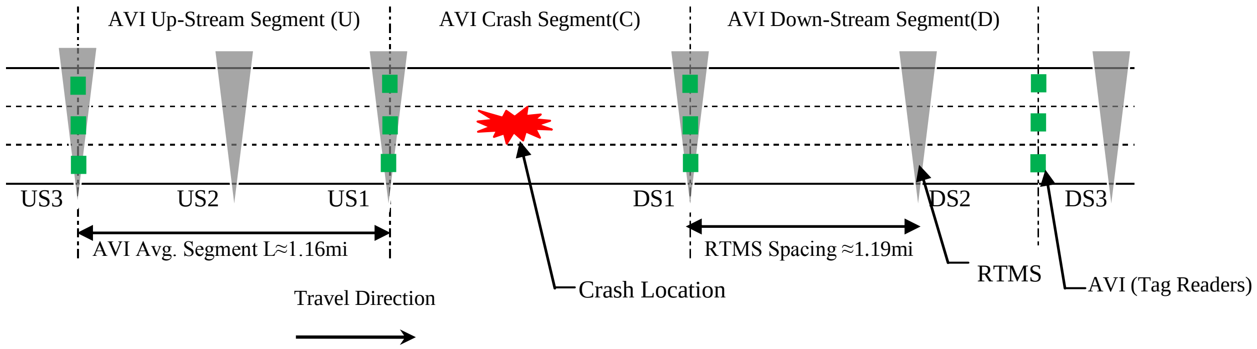


Figure 7-1: Arrangement of RTMS and AVI Segments

Three time slices of the 6-minutes prior to the crash time were extracted. For example if a crash happened on Sep 16, 2010 (Sunday) at 14:00, at the milepost of 210.1 EB. The corresponding 18-min window for this crash of time intervals (13:42 to 14:00) recorded by AVI segment 6 (Mile marker starts at 209.79 and ends at 210.60), upstream AVI segment 5 and downstream AVI segment 7 as well as 3 RTMSs in the upstream and 3 in the downstream were extracted. Time slice 1 was discarded in the analysis since it would not provide enough time for successful intervention to reduce crash risk in a proactive safety management strategy.

Moreover, the actual crash time might not precisely be known. Golob and Recker, 2004 discarded the 2.5 minutes of traffic data immediately preceding each reported crash time to avoid uncertainty of the actual crash time. In general with the proliferation of mobile phones and CCTV cameras on Freeways, crash time is almost usually immediately identified. One-hour speed profiles were also generated (about 30 minutes before and 30 minutes after the crash time) to verify the reported crash time. The modeling procedure required non-crash data, a random selection from the whole remaining AVI and RTMS datasets where there was no crash within 2-hour before the extraction time was utilized in the study to represent the whole population of different traffic patterns, weather conditions and roadway characteristics. A total of 18 (3 parameters x 3 AVI segments x 2 time slices) and 108 (9 parameters x 6 RTMSs x 2 time slices) input variables are prepared from AVI and RTMS raw data respectively.

Similarly, weather data for crash cases and non-crash ceases were extracted. Automated weather stations monitor the weather conditions continuously and the weather parameters are recorded according to a specific change in the reading threshold and hence they do not follow a specific time pattern. The stations report frequent readings as the weather conditions change within short time; if the weather conditions remain the same the station would not update the readings. However, these readings were aggregated over certain time periods to represent the weather conditions. For example; precipitation described by rainfall amount or snowfall liquid equivalent for ten minutes, one hour, three hours, six hours, twelve hours and twenty-four hours and the estimated average hourly visibility which provides an hourly measure of the clear distance in miles that drivers can see. Visibility in general can be described as the maximum distance (in mile) that an object can be clearly perceived against the background sky, visibility impairment can be the result of both natural (e.g., fog, mist, haze, snow, rain, windblown dust, etc.) and human induced activities (transportation, agricultural activities, and fuel combustion). The automated weather stations do not directly measure the visibility but rather calculate it from a measurement of light extinction which includes the scattering and absorption of light by particles and gases.

The basic parameters that define the geometrical characteristics of the roadway section for each crash and non-crash cases were considered in this study, these parameters include longitudinal grade, curve radius, deflection angle, degree of curvature, number of lanes, and width of median.

Multiple Stochastic Gradient Boosting models were calibrated for each dataset separately as well as for fused data from all sources. Each of these data were partitioned into 70% for training, 30% for validation using random sampling, in random sampling every observation in the data set has the same probability of being written to the sample. For example, the 70% of the population that is selected for the training data set, then each observation in the input data set has a 70% chance of being selected. Partitioning provides mutually exclusive datasets; two mutually exclusive datasets share no observations with each other. Partitioning is needed for machine learning (ML) models to have part of the data set for training in order to fit a preliminary model and find the best model weights using this training data set, and since ML techniques have the capacity for overtraining, validation data set will be used to retreat to a simpler fit than to calibrate the model based only on the training dataset. Validation part of the original data set is used for ML models fine-tuning to assess the prediction accuracy of each model. A total number of 186 crashes and 744 non-crashes were finally considered in the analysis.

7.3 Exploratory Comparison between AVI and RTMS Data

Interstate-70 in Colorado is equipped with both Automatic Vehicle Identification (AVI) and Remote Traffic Microwave Sensor (RTMS) Systems as part of the Intelligent Transportation System (ITS). Data from AVI are mainly used for toll collection and travel time estimation. It provides information about the space-mean-speed. RTMS is mostly used as a tool for operation and incident detection. It offers more detailed information on fundamental traffic parameters as time-mean-speed, volume and occupancy of each travel lane on the roadway.

Except that RTMS system keeps record of other traffic parameters that AVI system does not (i.e. volume and occupancy). It is also crucial to recognize that two types of speed are actually collected by the two systems and they differ with each other naturally. As discussed earlier, AVI measures space-mean-speed, which means that it reflects the average speed of all the vehicles occupying the detected road segment over a given time period (basically 2 min interval). RTMS measures time-mean-speed, the arithmetic mean of the speed of vehicles passing a point during specific time slice (normally 30 sec).

Moreover, due to the speed data from AVI are aggregated together without considering for inner or outer lanes, further attention should be paid on the potential difference between AVI speed data and RTMS speed data. For example, the outer lanes are more often travelled by trucks that could result in significantly lower average speed value for outer lane than inner lanes. However, this distinction could not be seen from the AVI speed data.

Therefore it is of great importance as well as interest to look into the data and check on the comparability of these two types of data. If they are comparable, then a useful alternative data source can be used when either one of them is not available.

Data are recorded at each RTMS station and each AVI segment. Two tables have been developed to give a clear view of the two data collection system along the 17-mile roadway section. Table

7-1 shows that RTMS locations are spaced on an average distance of 1.19 mile for both direction with standard deviation of 0.82 mile and 0.77 for eastbound and westbound. AVI segments have similar average length as the RTMS system, namely 1.16 mile and 1.15 mile for east and west bound but a little bit smaller standard deviation. Also, a majority of the starting and end points of AVI segment and RTMS stations are located at the same or close milepost. The spatial distribution of these stations facilitates the comparison between speed data from RTMS stations to those collected from corresponding AVI segments.

Table 7-1: RTMS Station Segment

	Eastbound Segment			Westbound Segment		
	Starting RTMS Station	Ending RTMS Station	Segment Length (mi)	Starting RTMS Station	Ending RTMS Station	Segment Length (mi)
1	205.7	208	2.3	205.7	207.1	1.4
2	208	208.7	0.7	207.1	208.9	1.8
3	208.7	209.79	1.09	208.9	209.79	0.89
4	209.79	210.8	1.01	209.79	210.6	0.81
5	210.8	211.8	1	210.6	211.8	1.2
6	211.8	213.3	1.5	211.8	213.3	1.5
7	213.3	216.7	3.4	213.3	216.7	3.4
8	216.7	217.4	0.7	216.7	217.4	0.7
9	217.4	217.85	0.45	217.4	217.85	0.45
10	217.85	218.1	0.25	217.85	218.1	0.25
11	218.1	218.7	0.6	218.1	218.7	0.6
12	218.7	219.7	1	218.7	219.7	1
13	219.7	221.1	1.4	219.7	221.1	1.4
14	221.1	222.36	1.26	221.1	222.36	1.26
	Average Segment Length		1.19	Average segment length		1.19
	Minimum Segment Length		0.25	Minimum segment length		0.25
	Maximum Segment Length		3.4	Maximum segment length		3.4
	Standard Deviation of Segment Length		0.82	Standard Deviation of Segment Length		0.77

Table 7-2: AVI Segments

	Eastbound			Westbound		
	Starting AVI Station	Ending AVI Station	Segment Length (mi)	Starting AVI Station	Ending AVI Station	Segment Length (mi)
1	205.05	207	1.95	205.7	205	0.7
2	207	208	1.0	207.1	205.7	1.4
3	208	208.7	0.7	208.9	207.1	1.8
4	208.7	208.79	0.09	209.79	208.9	0.89
5	209.79	210.8	1.01	210.6	209.79	0.81
6	210.8	211.8	1.0	211.8	210.6	1.2
7	211.8	213.4	1.6	213.4	211.8	1.6
8	213.4	215.3	1.9	215.3	213.4	1.9
9	215.3	216.7	1.4	216.57	215.3	1.27
10	216.7	217.85	1.15	217.4	216.57	0.83
11	217.85	218.7	0.85	218.1	217.4	0.7
12	221.1	222.4	1.3	218.7	218.1	0.6
13				219.7	218.7	1.0
14				221.1	219.7	1.4
	Average Segment Length		1.16	Average Segment Length		1.15
	Minimum Segment Length		0.09	Minimum Segment Length		0.6
	Maximum Segment Length		1.95	Maximum Segment Length		1.9
	Standard Deviation of Segment Length		0.52	Standard Deviation of Segment Length		0.42

In order to compare between AVI and RTMS data, three scenarios were considered:

1. Normal traffic condition (no crash reported within 2-hour);
2. Crash with property damage only;
3. Crash with injury or fatality.

For each case, an AVI segment is selected and the RTMS stations within this segment are also included. RTMS data are processed according to each lane at each station. Two-hours' records are studied.

Normal condition is defined as the traffic without interruption of crashes. Figure 7-2 represents a typical normal traffic flow condition. Though having variation, the speed curves are mild and without values of sudden drops or rises. The RTMS data give more detailed description about the speed distribution on each lane. From Figure 7-2 below it is shown that at the same station, speeds on inner lanes are higher than that on outer lanes. This can be explained by that outer lanes are designated for truck with lower speed limit. The AVI and RTMS give two different types of speed. Therefore it is not the focus on the direct comparison of the speed profiles. However, from Figure 7-2, it is clear that their patterns are alike.

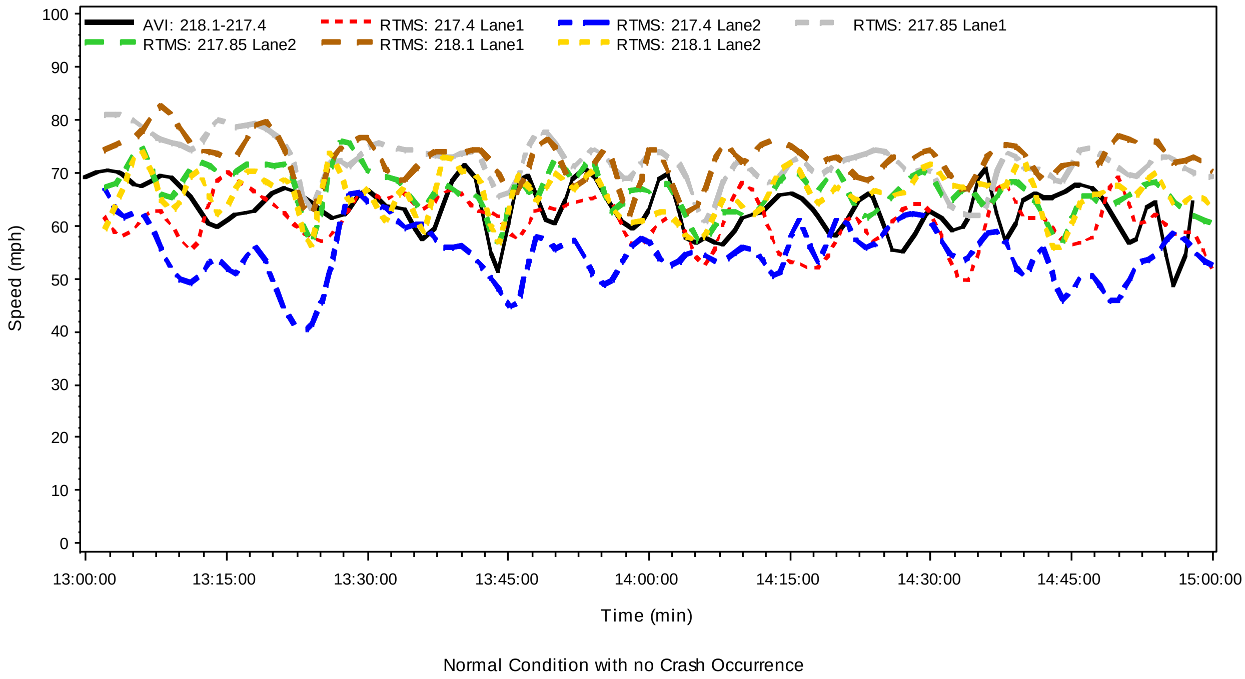
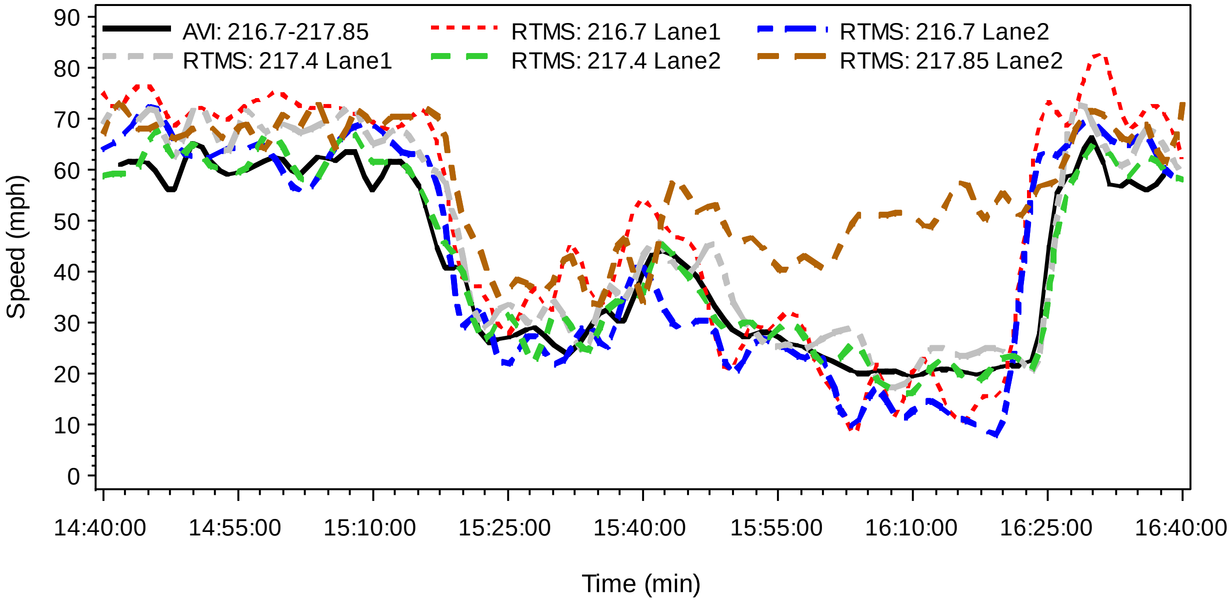


Figure 7-2: Westbound Dry Season Normal Condition

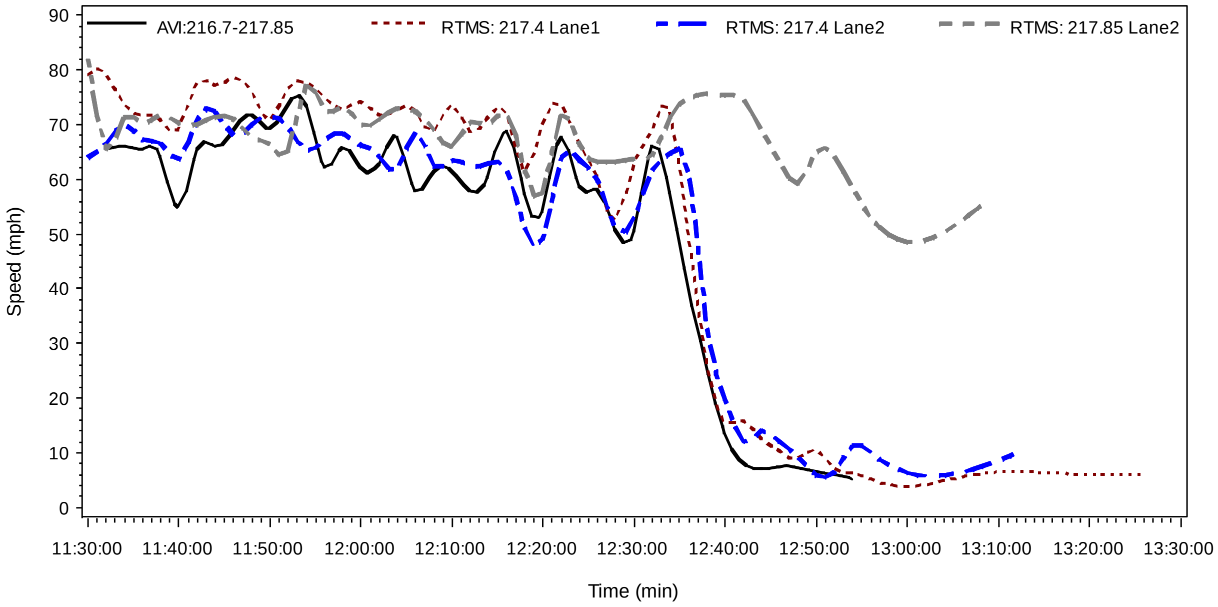
Figure 7-3 shows the occurrence of property-damage-only crash. Speed profile is from one hour before the crash to one hour after the crash. The figure is self-explanatory, when a crash happens on the roadway section, temporary congestion will be generated and vehicles upstream to the crash location will slow down. Once they pass the site, the speed will recover to some extent. So no significant change in speed from RTMS stations downstream after the crash happens shown in the figure has been expected. Figure 7-3 also demonstrates that speed from stations upstream can experience sudden rise, due to the removal of the vehicles involved in crash from the roadway. In this scenario, the AVI and RTMS give very comparable speed profiles.



Crash Location: Milepost 217.7

Figure 7-3: Snow Season Eastbound PDO Crash

When more severe crash occurs, as in Figure 7-4, both AVI and RTMS data show that the speed drops deeply. Different from the case of property damage only crash, when injury or fatality result from traffic crashes, intuitively, it takes longer time for the traffic flow to recover. In this crash happened at 12:30 pm on milepost 217.5, the congestion caused by it lasts more than one hour. Similar with PDO crash, AVI and RTMS still represent consistent pattern of speed.

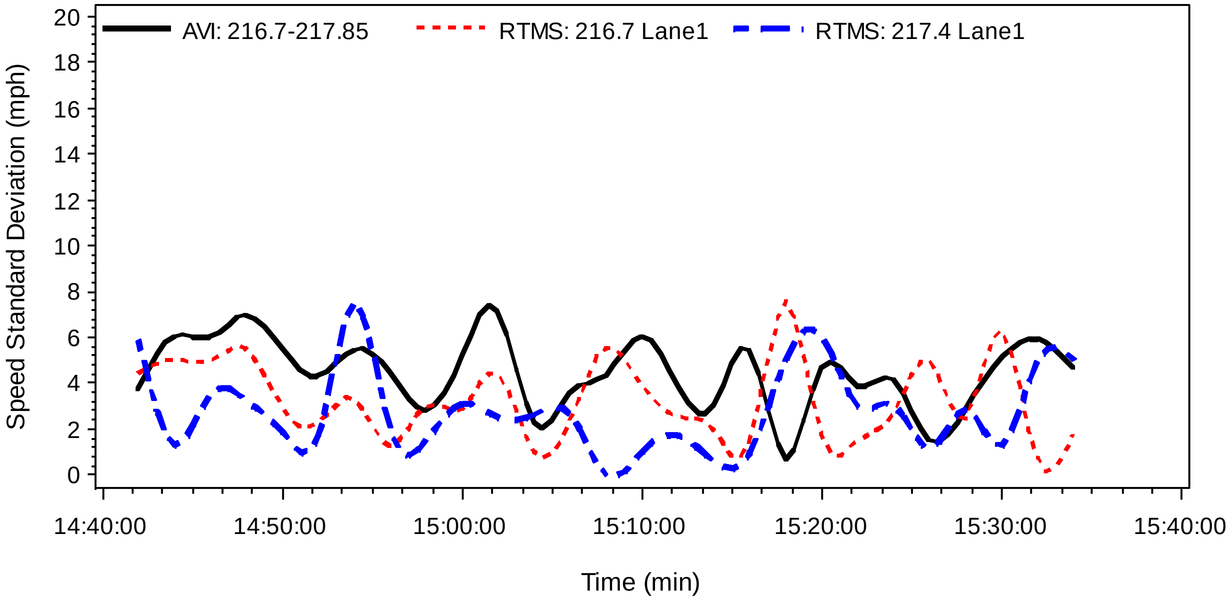


Crash Location: Milepost 217.5

Figure 7-4: Dry Season Eastbound Injury/Fatal Crash

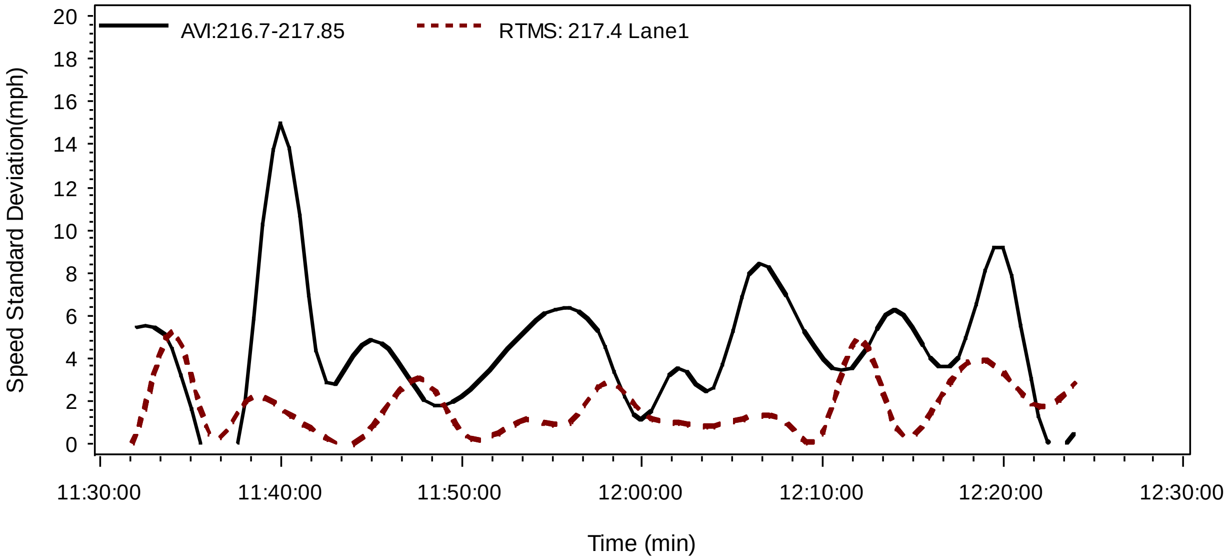
With the comparability of these two types of speed data in mind, variation of the speed from data collected from AVI and RTMS systems was also explored. From Figure 7-3 and 7-4, it can be seen that the existence of significant turbulence in speed prior to the occurrence of traffic crash on road. In order to get better insight about these two types of data, it is to be believed that it is necessary to look into the standard deviation of the speed before the crash occurrence. The 5 minutes' data just prior to the crash are discarded to the possible bias of the reported crash time. The standard deviation of the speed was determined on 2 minutes' interval basis from 1 hour to 5 minutes prior to the reported crash time. And only the inner lanes of RTMS stations upstream to the crash location are studied.

The standard deviation of the speed profiles indicate that AVI system records relatively higher speed variation than RTMS system does. Looking more closely to the 20 minutes period prior to crash, AVI data still provides higher variation.



Crash Location: Milepost 217.7

Figure 7-5: Speed Variation in Crash with Property Damage Only



Crash Location: Milepost 217.5

Figure 7-6: Speed Variation in Crash with Injury/ Fatality

Exploratory analysis and comparison of RTMS and AVI speed data reveals that they are highly comparable in recording the speed trends in normal traffic condition, crash with property damage only and more hazardous conditions involving injury and fatality. The data collected from each system could strengthen one another's credibility when traffic data are missing. RTMS system provides more detailed information in respect of speed. Lane by lane information is provided by RTMS while AVI in its current archiving system provides lane aggregated speed data. On the other hand, AVI system is more sensitive to higher speed variation, which has been attributed as a factor to the occurrence of crash. The examination of two systems suggests that combining them together in the modeling process might help with more accurate crash prediction.

7.4 Stochastic Gradient Boosting

The Stochastic Gradient Boosting (SGB) is a machine learning technique that was introduced by (Friedman, 2001). This technique which is also known under other names such as Multiple Additive Regression Trees (MART), and TreeNet is technically suitable to be used for all data mining problems including regression, logistic regression, multinomial classification and survival models. The general idea of boosting is to create a series of simple learners known as “weak” or basic learners, i.e. a classifier that has a slightly lower error rate than random guessing. Most of the boosting algorithms use binary trees with only two terminal nodes as the basic learner (Hastie et al., 2001). Boosting these simple trees forms a single predictive model. The gradient boosting trees method has been proposed as a recent advancement in data mining that combines the advantages of the non-parametric tree-based methods and the strengths of boosting algorithms. It showed outstanding prediction performance in different fields including; real-time credit card fraud detection and terrorism culpability. The fraud detection application has some similarity to real-time crash prediction; with thousands of credit, debit and online transactions taking place every minute; the probability of a fraud transaction is very small and the variables space is relatively high, the mechanism that is deployed to monitor all transactions in real-time may be adopted in traffic safety applications.

Some of the key features of Stochastic Gradient Boosting are its ability of handling large number of mixed predictors (quantitative and qualitative) without preprocessing of rescaling or transformation which allows real-time traffic and weather data to be directly fed into the SGB algorithms without any time consuming processes. Moreover, by using CART as the basic learner, SGB can automatically handle the missing values which can still yield an accurate prediction in case of missing one of the important variables with no need to consider prior data imputation (Breiman et al, 1983). SGB has the capability of resisting the outliers in predictors and it can perform well with partially inaccurate data, therefore any erroneous traffic data can be handled easily without cleaning. Additional advantage of tree-based models is the robustness of variable selection; tree models have the capability of excluding irrelevant input variables. The main disadvantage however of single tree models is instability and poor predictive performance especially for larger trees which can be mitigated by other techniques that can improve model accuracy such as boosting, bagging, stacking, model averaging and ensemble which merges results from multiple models. Stochastic gradient boosting is uniquely advantageous over other merging techniques because it follows sequential forward stagewise procedure. The process of boosting is an optimization technique to minimize a loss function by adding a new simple learner (tree) at each step that best reduces the loss function, first tree is selected by the algorithm that maximally reduces the loss function. The residuals are the main focus for each following step by performing weighted resampling to boost the accuracy of the model by giving more attention to observations that are more difficult to classify. As the model enlarges, the existing trees are left unchanged; however, fitted value for each observation is to be re-estimated at each new added tree. The sampling weight is adjusted at the end of each iteration for each observation with respect to the accuracy of the model result. Observations with correct classification receive a lower sampling weight while incorrectly classified observations receive a higher weight. In the next iteration, a sample with more misclassified observations would be drawn.

SGB was used for classification in which, traffic, weather, and geometry variables are used as independent variables x to identify the binary crash $\mathrm{y} \in \{-1,1\}$, by using a “training” sample $\{\mathrm{y_i}, \mathrm{x_i}\}_1^{\mathrm{N}}$ of known (y, x) values. The goal of estimating the function that maps the traffic, weather and geometry features to crashes is to be used for prediction of the increased risk for future observations, where only x is known. As explained in Friedman (2001) we need to obtain an approximation $F(x)$ of the function $F^*(x)$ linking x to y, that minimize the expected value of a loss function $\Theta(y, F(x))$ over the joint distribution of all (y, x) values

$$F^*(x) = \arg\ \min_{F(x)} E_{y,x} \Theta(y, F(x)) \tag{7.1}$$

As mentioned earlier, the boosting idea is to build an additive model on a set of basic functions (weak classifier). In case of using a single tree as the individual classifier, the boosted tree model will be a sum of many simple trees:

$$f_T(x) = \sum_{m=1}^{M} T_m(x;\ \gamma_m, R_m) \tag{7.2}$$

where

$$T_m(x;\ \gamma_m, R_m) = \sum_i^{I_m} \gamma_{mi} I(x \in R_{mi}) \tag{7.3}$$

where $R_{mi}, i = 1,2, \dots, I_m$ are disjoint regions that collectively cover the space of all joint values of X. γ_{mi} is a constant that is assigned to each such region. $R_{mi}$is the i^{th} terminal node in tree m with fitted value of γ_{mi}. Ideally, γ_{mi} and $R_{mi}$are fitted by minimizing a loss function;

$$\min_{\{\gamma_m\ R_m\}_1^M} \sum_{j-1}^N \Theta\ (y_j, \sum_{m=1}^M T_m(x_j;\ \gamma_m, R_m)) \tag{7.4}$$

Commonly used loss function for classification is given by;

$$\Theta(y, \hat{F}) = 2\log\ (1 + \exp\ (-2y\hat{F})) \tag{7.5}$$

Where,

$$F(x) = \frac{1}{2}\ \log\ \left[\frac{\Pr\ (Y=1|x)}{\Pr(Y=-1|x)}\right] \tag{7.6}$$

The solution can be approximated by iteratively adding a single tree at each step without adjusting the parameters of the existing trees as mentioned earlier. Therefore, by adding tree k+1, the following equation can be minimized

$$\sum_{j=1}^N \Theta\ (y_j, \sum_{m=1}^K T_m(x_j;\ \gamma_m, R_m) + T_{k+1}(x_j; \gamma_{k+1}, R_{k+1})) \tag{7.7}$$

as a function of γ_{k+1} and R_{k+1}, holding γ_1,…, γ_k and R_1,…, R_k fixed. After M iterations (7.7) will achieve (7.4).

7.5 Results and Discussion

7.5.1 Model Estimation, Interpretation and Diagnostics

This section explains how the calibration, interpretation and evaluation processes were performed.

In this study, Stochastic Gradient Boosting models were fitted in SAS Enterprise Miner 6.1. The SGB was iterated 50 times with different random samples in the validation dataset to stabilize the error rate. The optimization parameters were set at SAS default values; shrinkage (learn rate) =0.1, train proportion (different training observations are taken in each iteration) =60, maximum branch=2 (binary tree), and the maximum depth (number of generation) =2.

In machine learning applications, the data may include easily hundreds of variables; a key question therefore whether or not all these variables actually lead to true information gain? The answer is obviously, no, since there are a lot of redundant variables that may increase the performance of the learning data set but they do not necessarily increase the performance on the actual validation dataset which can be easily controlled for by keeping an eye on the over-fitting. Many data mining techniques such as neural networks, near-neighbor, kernel methods, and support vector machines perform worse when extra irrelevant predictors are added, and therefore variable selection technique should always precede the modeling. On the other hand tree-based models are highly resistant to the inclusion of irrelevant variables; tree-based models perform automatic variable subset selection.

One of the main advantages of tree-based models is their simple interpretability. Single tree model can be graphically illustrated by two-dimensional figure that is easily interpreted. On the other hand, boosted trees are formed of linear combination of many trees (hundreds and in some cases thousands of trees), and therefore forfeit this important feature. The main two components of interpretation are identifying the variables importance and understanding their effect on the classification problem which are provided in all conventional regression models.

Fortunately, unlike other black-box machine learning techniques, SGB can be summarized and interpreted. Relative importance of predictor variables can be conveniently calculated, the variable importance is based on the number of times a variable is selected for splitting rule and weighted by the squared improvement to the model as a result of each split, and averaged over all trees as explained in Friedman and Meulman (2003). Table 7-3 provides the selected variable subsets and their relative importance for each of the calibrated models. The input variables characterized by a relative importance smaller than 25% have been discarded in the SGB models.

Stochastic Gradient Boosting models were estimated for four different datasets; Model-1 was calibrated using all available data collected from AVI, RTMS and weather stations as well as geometrical characteristics for crash/non-crash cases. In order to examine the prediction accuracy that can be achieved depending only on one dataset at a time and to account for any interruption of the data flow from any source, another three models were calibrated; Model-2 based only on RTMS data, Model-3 based only on AVI data, and Model-4 based on real-time weather data.

It may be observed from Model-1 results that the most important variables are traffic data collected from RTMS such as Average occupancies from US2 and US3 sensors during time slice two and three respectively (time slice 2: 6-12 minutes before the crash and time slice 3: 12-18 minutes before the crash), followed by logarithm of the coefficient of variation of speed from AVI crash segment at time slice 2 and average speed from AVI downstream segment at time slice 2, other RTMS and AVI variables were selected but with less relative importance. Weather related variables are relatively important; 1-hour visibility is shown at the top of the list just after some traffic variables. The ten-min precipitation variable was also selected among the important variables. Other site-related variables came out to be important including longitudinal grade, number of lanes, absolute degree of curvature and width of median.

Comparison between models performance is subjective and depends on different criteria; misclassification rate and the area under the Receiver Operating Characteristics (ROC) were used as the main performance criteria in this analysis. The area under the ROC curve shows how well the model is at discriminating between the crash and non-crash cases in the target variable. This is similar to the misclassification rate, but the ROC curve plots sensitivity vs. 1 – specificity values for many cutoff points. The area under the curve seems to be large for the best selected model in red color (model) as shown in Figure 7-7. The exact areas under the ROC curves for all models validation datasets are listed in Table 7-4.

Table 7-3: Variable Importance

Model-1		Model-2		Model-3		Model-4	
Variables	**Variable Importance**	**Variables**	**Variable Importance**	**Variables**	**Variable Importance**	**Variables**	**Variable Importance**
Avg. Occ. Upstream1_Time Slice _2	1.000	Avg. Occ. Upstream 2_Time slice_3	1.000	Log. Coef. of Var. of Speed Crash Segment Time Slice_2	1.000	1-Hour Visibility	1.000
Avg. Occ. Upstream 2_Time slice_3	0.887	Log. Coef. of Var. of Speed Upstream 1_Time Slice_2	0.997	Avg. Speed Downstream Segment Time Slice_2	0.899	10-Minute Precipitation	0.459
Log. Coef. of Var. of Speed Crash Segment Time Slice_2	0.798	Avg. Speed Upstream 2_Time Slice_2	0.804	Avg. Speed Downstream Segment Time Slice_3	0.741	1-Hour Precipitation	0.324
Avg. Speed Downstream Segment Time Slice_2	0.742	S.D. Occ. Upstream 2_Time Slice 2	0.541	Avg. Speed upstream Segment Time Slice_2	0.537		
1-Hour Visibility	0.684	Avg. Speed Downstream 1_Time Slice_2	0.457				
Grade	0.661	Avg. Speed Downstream 2_Time Slice_2	0.391				
S.D. Occ. Upstream 3_Time Slice 2	0.642	Avg. Occ. Upstream1_Time Slice _2	0.374				
No. of Lanes	0.521	Avg. Occ. Upstream2_Time Slice _2	0.348				
Avg. Speed Upstream 1_Time Slice_2	0.519	Log. Coef. of Var. of Volume Downstream 2_Time Slice_2	0.249				
Avg. Speed Downstream Segment, Time Slice_3	0.431						
Abs. Deg. of Curve	0.337						
10-Minute Precipitation	0.335						
Log. Coef. of Var. of Volume Downstream 2_Time Slice 3	0.334						
Log. Coef. of Var. of Speed Upstream Segment_Time Slice 3	0.329						
Med. Width	0.278						

Generally, Model-1 is consistently superior in term of classification accuracy and area under the ROC curve. Model-2 and Model-3 are relatively ranked lower than Model-1 but still providing satisfactory performance. Model-4 is ranked the lowest on these measures. Area under the ROC curves as shown in Figure 7-7 and listed in Table 7-4 was found to be 0.946 for Model-1 validation dataset, 0.762 and 0.721 for Model-2 and Model-3, respectively while Model-4 achieved only ROC of 0.675 all for the validation datasets.

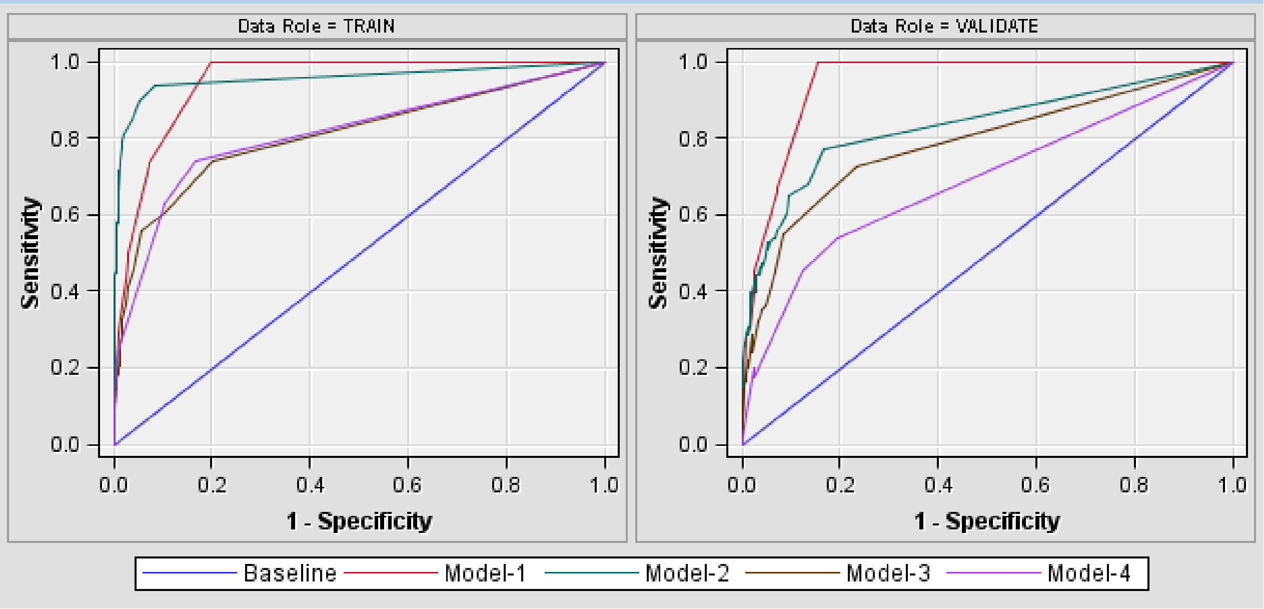


Figure 7-7: Receiver Operating Characteristics Chart

Unlike previous studies that only reported accuracy and misclassification rate at one cutoff value, in this study the accuracy and misclassification rates are graphically illustrated for many cutoff values as shown in Figures 7-8 to 7-11. In terms of accuracy and misclassification rate, also Model-1 outperformed all other individual models in all classification measures. Sensitivity analysis is important for the implementation of the proposed system in real-life application; while the overall classification rate can provide some insight of the model performance, sensitivity which is defined as the proportion of crashes (event cases) that are correctly identified as crashes (known also as true positive rate) is usually the most important measure of accuracy. Other measure that may affect drivers' compliance to the management system and should be kept as minimum as possible is the proportion that is incorrectly classified as crashes (false positive rate). As shown in Figures 7-8 to 7-11 that different false positive rates can be obtained by changing the cutoff value. In order to fairly compare across the four calibrated models, cutoff values have been chosen that achieve the highest possible sensitivity while preserving false positive rates at low values ranging between 5 to 8 percent, specificity (the proportion of correctly identified non-crashes) and overall classification. As illustrated in Figures 7-8 to 7-11 and summarized in Table 7-4 for the chosen cutoff values, Model-1 identified about 89% of crashes correctly while only about 6.5% of non-crash cases were incorrectly identified as crashes; Model-1 also achieved the highest overall accuracy of about 92%. Model-2 and Model-3 ranked the second in term of overall accuracy with Model-2 performed slightly better than Model-3 to the respect of true positive rate and area under ROC curve as mentioned earlier. Model-4 achieved the lowest overall accuracy and true positive rate in the same range of false positive rate defined above.

Table 7-4: Validation: Classification Rates and ROC Index

Model	Model Description	Overall Classification Rate	True Positive Rate	False Positive Rate	True Negative Rate	ROC Index
Model-1	All Data	92.157%	88.889%	6.481%	93.519%	0.946
Model-2	RTMS	87.879%	73.333%	7.154%	92.845%	0.762
Model-3	AVI	87.653%	70.192%	6.393%	93.607%	0.721
Model-4	Weather	84.364%	55.714%	5.854%	94.146%	0.675

Although Model-4 (weather only based model) performed not as good as the other 3 models, inclusion of weather information is essential in risk assessment framework; drivers need to have localized real-time information especially during adverse weather, including pavement conditions, visibility level, lane closure, snow, heavy rain and fog. The weather information would be more relevant if provided at segment level rather than regional level. According to the Federal Highway Administration (Goodwin, 2002), weather contributed to over 22% of the total crashes in 2001. This means that adverse weather can easily increase the likelihood of crash occurrences. Several studies, in fact, concluded that crashes increase during rainfall by 100% or more (Brodsky and Hakkert, 1988; NTSB, 1980), while others founnd more moderate (but still statistically significant) increases (Andreescu and Frost, 1998; Andrey and Olley, 1990). Model-4 may provide an adequate measure of risk in scenarios where weather information is only available and may help toward more weather responsive traffic management.

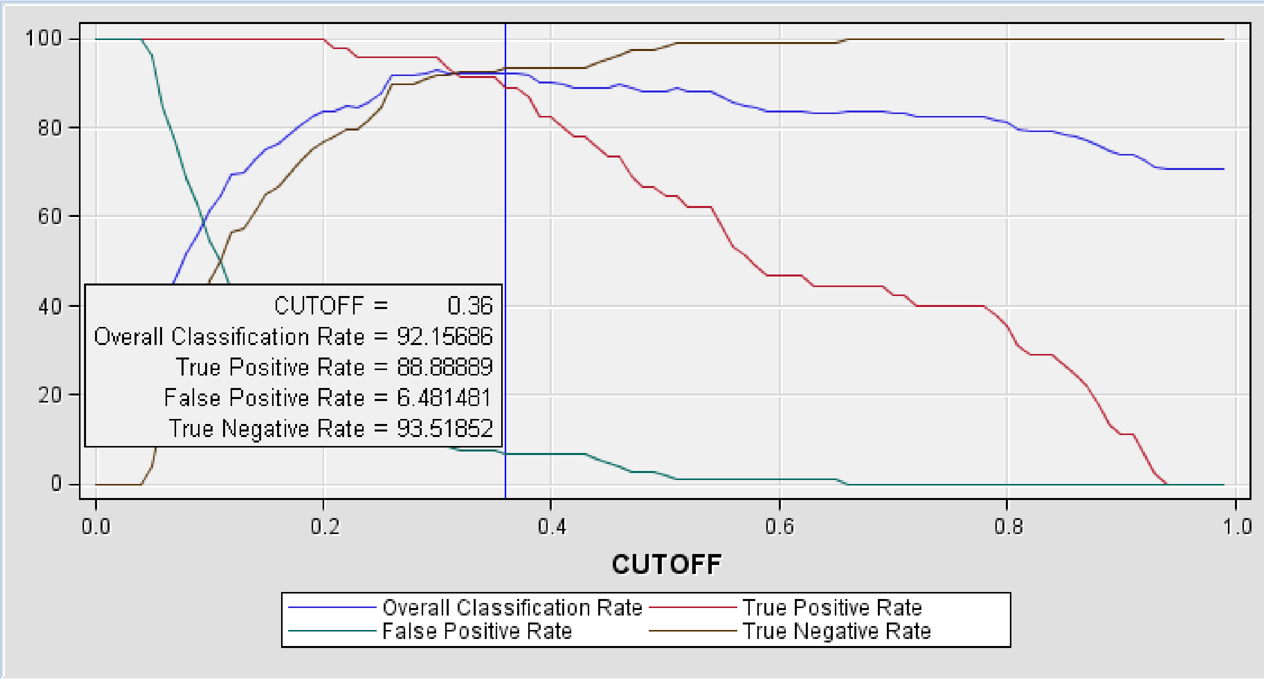


Figure 7-8: Model-1 Classification Rates

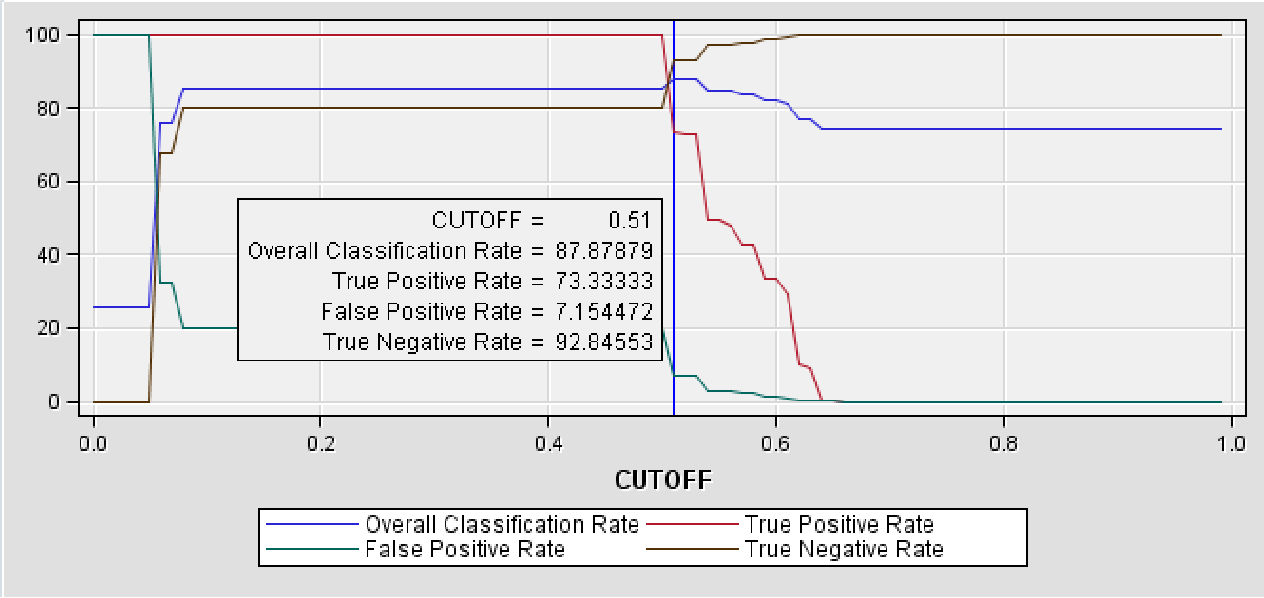


Figure 7-9: Model-2 Classification Rates

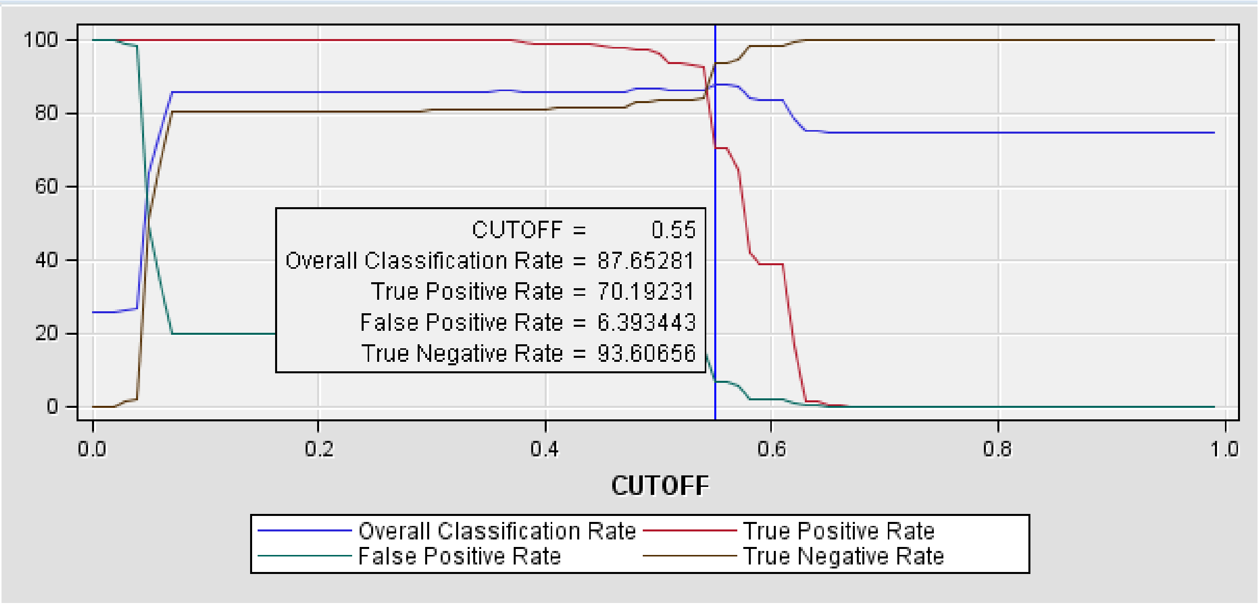


Figure 7-10: Model-3 Classification Rates

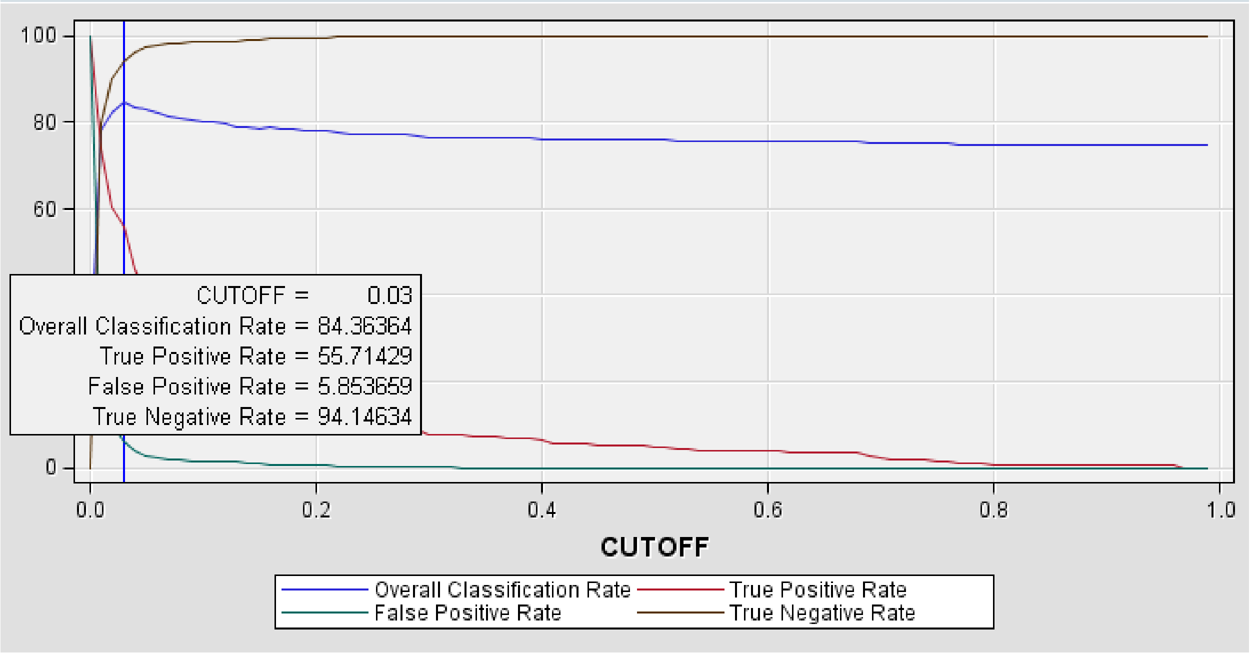


Figure 7-11: Model-4 Classification Rates

7.6 Risk Assessment Framework

The collected data on the study roadway section is one of the greatest assets that should be utilized appropriately to maximize the benefit for the roadway authority as well as for the road users. Buried within this vast amount of data is useful information that could make significant difference in how these roads are managed and operated. Figure 7-12 illustrates a framework to assess the increased real-time risk depending on the availability of on-line data. The idea behind the proposed framework based on the fact that although the traffic detection and meteorological stations became advanced enough to overcome hardware failures and malfunctions, the challenging weather conditions may interrupt the flow of the data in real-time at some point. Therefore, a reliable and robust framework should be in place at all times. Moreover, another issue that was discussed but not explicitly addressed in previous studies is how different the prediction accuracy of traffic data that are collected from different sources at the same location in identifying hot spots on freeway sections in real-time.

There are 4 main models calibrated in the proposed framework; Model-1 based on all available data collected from AVI, RTMS, weather stations and roadway geometry, Model-2 based only on RTMS data, Model-3 based only on AVI data, and Model-4 based on real-time weather data. As shown in the flowchart in Figure 7-12, in case of the availability of all traffic and weather data at the same time, these data would be fused together to provide the most comprehensive data and then Model-1 can be calibrated. If a hazardous traffic condition is detected, this section would be flagged, otherwise, the section would be operated under normal condition. The other 3 models are calibrated for each data separately to examine how each model performs and to substitute the full model in case of absence of other data as mentioned earlier. Based on Model-2, a roadway section can be flagged if unsafe traffic condition was encountered otherwise Model-4 needs to be checked. If a critical visibility or adverse weather encountered from Model-4 then an advisory/warning messages have to be issued to inform drivers about the situation. It should be noted that some specific traffic regimes would not be affected by inclined weather; however, drivers may still need some advisory messages to help them in selecting the safe operating speed. In case that the real-time weather is not available, advisory messages can be issued depending on the forecasted weather. The same logic can be followed by Model-2 using data collected from AVI.

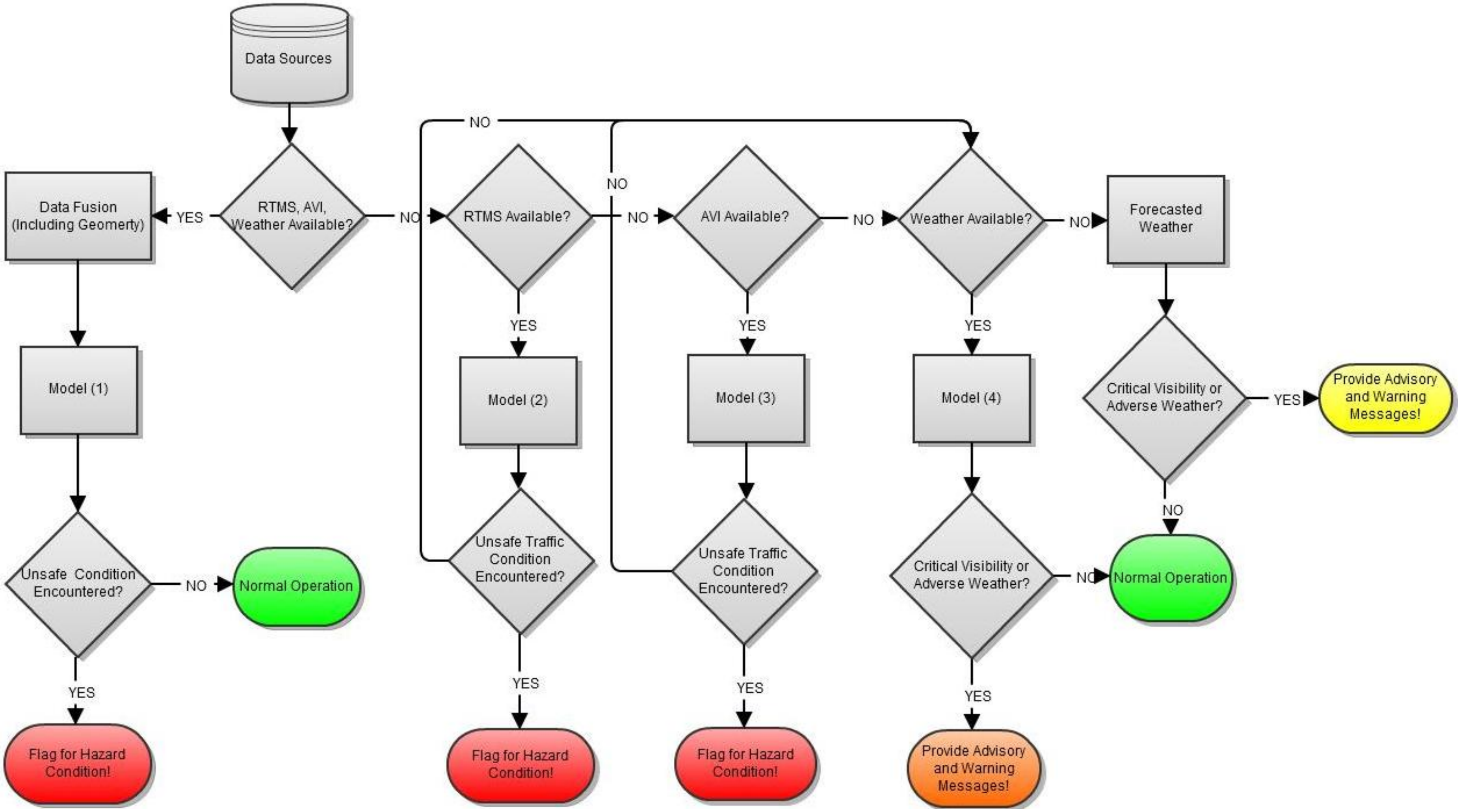

Data Sources
RTMS, AVI, Weather Available?
YES
Data Fusion (Including Geomerty)
Model (1)
Unsafe Condition Encountered?
NO
Normal Operation
YES
Flag for Hazard Condition!
NO
RTMS Available?
YES
Model (2)
Unsafe Traffic Condition Encountered?
YES
Flag for Hazard Condition!
NO
NO
AVI Available?
YES
Model (3)
Unsafe Traffic Condition Encountered?
YES
Flag for Hazard Condition!
NO
NO
Weather Available?
YES
Model (4)
Critical Visibility or Adverse Weather?
YES
Provide Advisory and Warning Messages!
NO
NO
Forecasted Weather
Critical Visibility or Adverse Weather?
YES
Provide Advisory and Warning Messages!
NO
Normal Operation

Figure 7-12: Framework of the Real-Time Risk Assessment

7.7 Conclusion

The recent advances in data collection technologies for traffic and weather on freeway sections provided valuable asset that should be utilized properly to increase safety and mobility and in order to maximize the benefit for highway authorities as well as for road users. These valuable data can be utilized to provide a framework for real-time risk assessment on freeways and expressways.

By fusing data from two different detection systems (AVI and RTMS), real-time weather and geometrical characteristics, the database created in this analysis are by far the most comprehensive database created for a real-time crash prediction study.

In this chapter, a relatively recent machine learning technique known under different names such as Stochastic Gradient Boosting (SGB), Multiple Additive Regression Trees (MART), and TreeNet was used to analyze 186 crashes occurred on 15-mile mountainous freeway section (I-70) in Colorado. The analyses were set up as a binary classification problem in which traffic, geometry, and weather variables are used as independent variable to identify crashes in real-time. The proposed learning machine methodology seems to provide all advantages that are needed in a real-time risk assessment framework. The Stochastic Gradient Boosting inherited all key strengths from tree-based models of their ability of selecting relevant predictors, fitting appropriate functions, accommodating missing values without the need for any prior transformation of predictor variables or elimination of outliers while overcoming the unstable prediction accuracy of single tree models. Boosting is considered unique among other popular aggregation methods; while ensemble, bootstrap or bagging, bagged trees and random forest can improve single tree models performance. Bagged trees and random forest can reduce variance more than single trees, however unlike boosting; they cannot achieve any bias reduction (Prasad et al., 2006).

The proposed methodology has brought considerable advantage over classical statistical approaches. In particular, it has provided outstanding performance. On the other hand, machine learning techniques are being argued against for being black boxes; there are no *P* values to indicate the relative significance of model coefficients and there is no simple model with fewer variables. The proposed methods of interpretation (variable importance) and evaluation (ROC and classification) can be regarded as functional equivalence to many conventional regression techniques, thus addressing the criticisms against machine learning techniques.

Another issue that has been explicitly addressed in this study is how different the prediction accuracy of traffic data that are collected from different sources at the same location in identifying hot spots on freeway sections in real-time; the results showed that crash prediction from AVI is comparably equivalent to RTMS data. Moreover, the accuracy of the main model that is augmenting information from multiple traffic detectors (AVI and RTMS), weather, and geometry performed the best in terms of classification rate and area under the ROC curve. The overall model (Model-1) identified about 89% of crash cases in the validation dataset with only 6.5% false positive.

This study proposed a framework for real-time risk assessment using data from multiple sources that can achieve reliable and robust prediction performance under different scenarios of data availability. The results depict that traffic management authorities as well as road users can benefit from the wealth of collected data from multiple sources not only to alleviate traffic congestion but also to mitigate increased safety risk.

CHAPTER 8. CONCLUSIONS AND RECOMMENDATIONS

This dissertation comprised a multi-level safety assessment for freeways and expressways. The study utilized comprehensive databases from different regions to examine the applicability of such functions on different freeway/expressway types (i.e. geometrical characteristics and environmental conditions). Classical (frequentist), Full Bayesian and Machine Learning statistical approaches were implemented to achieve the objectives discussed in this dissertation.

This chapter discusses key findings, conclusions and future recommendations for freeways/expressways safety analysis on the aggregate and disaggregates levels.

8.1 General

The main objectives of the developed multi-level Safety Performance Functions (SPFs) are 1) to assess constant hazards (site-specific static risks) as well as 2) to identify real-time risks due to turbulent traffic conditions and interactions with other risk factors. To achieve these objectives, SPFs were developed at the aggregate level using historical crash data and the corresponding exposure and risk factors in which the unit of analysis was the crash frequency. Additionally, other SPFs were developed for individual crashes at the disaggregate level to identify crash prone conditions in real-time. Both levels of aggregate and disaggregate analyses were found to be important, the first helped in providing good understanding of different safety problems, ranking the hazardous sites, and developing policies and countermeasures to reduce the number of crashes in total. Also, hazardous sites (hot spots) were identified and hence resources can be allocated more appropriately. In order to assess and enhance the performance of freeways and expressways in real-time, the SPFs based on the disaggregate level can be implemented. This approach may be utilized to provide more proactive traffic management system that will not only enhance the performance of the high speed facilities in particular and the whole traffic network in general but also provides safer mobility for people and goods.

In this dissertation, the most comprehensive data were prepared. There were two main datasets prepared from two different regions; 78-mile on the expressway network in Orlando city, Florida, and a 20-mile mountainous interstate roadway-section west of Denver, Colorado. These datasets comprise of historical crash data, roadway geometrical characteristics, real-time weather and traffic data. The traffic flow parameters were collected from various types of advanced detection systems such as Automatic Vehicle Identification (AVI) and Remote Traffic Microwave Sensors (RTMS).

8.2 Bayesian Hierarchical Approach for Developing SPFs

The safety effects of roadway geometrics on crash occurrence along a freeway section that features mountainous terrain and adverse weather were explored using Poisson models, Bayesian hierarchical models with spatial and random effects were developed to efficiently model the crash frequencies for six years at the roadway section. Furthermore, a Bayesian ranking technique was implemented to rank the hazard levels of the roadway segments. It was found that while the random effect and spatial models outperform the Poisson model, the spatial model may have the problem of redundantly accounting for the geometry dependent effect. Therefore the random effect model was selected for model inference. Estimation of the model coefficients indicates that roadway geometry is significantly associated with crash risk; segments with steep downgrades were found to drastically increase the crash risk. Moreover, this crash risk could be significantly increased during the snow season compared to dry season as a confounding effect between grades and pavement condition. Additionally, sites with higher degree of curvature, wider medians and an increase of the number of lanes appear to be associated with lower crash rate. Based on Bayesian ranking technique; the results confirmed that segments with steep downgrades are more crash prone along the study section. These identified sites should receive more attention from officials and decision makers especially during the snow season. This aggregate level of analysis provided good understanding of the effects of roadway geometrics and weather on crash frequencies on mountainous freeways. Furthermore, the results depict that this step should be considered before proceeding to disaggregate level analysis.

In the future, the Bayesian Hierarchical approach could be extended to utilize informative prior employing real-time traffic and weather data. Instead of using aggregate traffic measure (e.g. ADT and speed limit), and aggregate weather information (e.g. number of rainy days), the mean and the distribution of the archived real-time traffic characteristics of volume, speed and occupancy and real-time weather of visibility, precipitation and temperature could be implemented to provide more certain prior information. Furthermore, with the availability of more crash and risk factors data, the analysis could be expanded to analyze specific crash types (e.g. single-vehicle crashes and multi-vehicle crashes) and different severity levels (e.g. property damage only, injury and fatal crashes). This can shed more light on the different mechanisms for each crash type and identify the different factors that affect different severity levels.

In segment-based SPFs studies, it is believed that crashes are not randomly distributed, but are usually associated with underlying geometrical characteristics, environmental and traffic conditions. In this study a homogeneous segmentation method was adopted, it is worth to investigate different segmentation methods and compare across them to better understand how the segmentation method can affect the analysis results.

8.3 The Viability of Using AVI Data in Real-Time Risk Assessment

Real-time individual crash analysis captured the researchers' interest in the last decade since it has the capability of identifying crashes in real time and hence being more proactive in safety management rather than being reactive. The real-time risk assessment research attempted the use of data from inductive loop detectors; however, no safety analysis has been carried out using traffic data from an increasingly prevalent non-intrusive surveillance system; the tag readers on toll roads known as Automatic Vehicle Identification (AVI). In this dissertation, for the first time, the identification of freeway locations with high crash potential has been examined. Explicitly three main issues were tackled in this study; 1) utilizing matched case-control logistic regression to examine the viability of using AVI data in crash prediction, 2) comparing between the prediction performance of a single generic model for all crashes and a specific model for rear-end crashes using AVI data, 3) applying Bayesian updating approach to generate full probability distributions for the coefficients and to examine the estimation efficiency of the Semi-parametric Bayesian modeling over the frequentist matched-case control logistic regression.

AVI data were found to be promising in providing a good measure of crash risk in real time. The operation-based management of expressways can benefit from the collected AVI traffic data not only to ease congestion and enhance the operation but also to provide warnings of increased risk situations to promote safety on freeways and expressways. By contrasting AVI data (collected on OOCEA expressway network in Orlando) preceding all crashes and rear-end crashes with matched non-crash data, it was found that rear-end crashes can be identified with a 72% accuracy while the generic all crash model achieved accuracy of only 69% using different validation datasets, moreover, using the Bayesian updating approach increased the accuracy of both models by 3.5%.

The current AVI archiving system has some limitations that can be easily addressed, one limitation is that the system does not record the percentage of lane change per segment; this percentage can be calculated by developing an algorithm to compare the unique tag ID for each individual vehicle at the beginning and end of each segment which will add a unique feature to AVI systems over the ILD. Moreover, the algorithm can process the original raw AVI data in a way that provides space mean speed by lane and hence a better picture can be comprehended about not only the longitudinal speed variation at the AVI segment but also the variation across the lanes. It is to be noted that by having detailed lane speed data may help to identify other types of crashes such as sideswipe and angle crashes. Another limitation is that AVI does not provide other traffic parameters such as volume and occupancy. These data can be easily estimated and archived, for example, volumes can be calculated from number of transponders reading and weighted to the total transactions from other payment methods, it could also be provided by lane.

Headways also could be estimated by analyzing time stamps for individual successive vehicles at each tag reader by lane, which can provide a measure of density.

8.4 Incorporating Roadway Geometry and Weather in Real-time Risk Assessment

The effect of the interaction between roadway geometric features, and real-time weather and traffic data on the occurrence of crashes on a mountainous freeway was investigated. The Bayesian logistic regression technique was used to link a total of 301 crash occurrences on I-70 in Colorado with the real-time space mean speed collected from the Automatic Vehicle Identification (AVI) system, real-time weather and roadway geometry data. The results suggest that the inclusion of roadway geometrics and real-time weather with AVI data in the context of active traffic management systems is essential, in particular with roadway sections characterized by mountainous terrain and adverse weather. The modeling results showed that the geometric factors are significant in the dry and the snow seasons and the crash likelihood could double during the snow season because of the interaction between the pavement condition and steep grades. The 6-minute average speed at the crash segment during 6-12 minutes prior to the crash time and the 1-hour visibility before the crash time were found to be significant in the dry season while the logarithms of the coefficient of variation in speed at the crash segment during 6-12 minutes prior to the time of the crash, 1-hour visibility as well as the 10-minute precipitation prior to the time of the crash were found to be significant in the snow season. The results from the two models suggest that different active traffic management strategies should be in place during these two distinctive seasons.

8.5 A Framework for Real-Time Risk Assessment Using Mixed Detection Systems

The increased deployment of non-intrusive detection systems such as automatic vehicle identification (AVI) and remote traffic microwave sensors (RTMS) provided an access to real-time traffic data from multiple sources. The data that are collected from such systems is one of the greatest assets that should be utilized appropriately to maximize the benefit for the roadway authority as well as for the road users. Buried within this vast amount of data is useful information that could make a significant difference in how these roads are managed and operated. Data mining and Machine Learning techniques are known for their capability of extracting the useful hidden information from the massive archived data as well as their superior performance in classification and prediction. Stochastic Gradient Boosting (SGB), a relatively recent and promising machine learning technique was used to calibrate several models utilizing different datasets collected from mixed detection systems as well as real-time meteorological stations (collected on I-70 in Colorado). The results showed that crash prediction from AVI is comparably equivalent to RTMS data, crash prediction model utilizing RTMS data only identified 73% of crash cases with 7% false positive while AVI only model identified 70% with about 6.5% false positive rate. Moreover, the accuracy of the full model that is augmenting information from multiple traffic detectors (AVI and RTMS), weather, and geometry performed the best in terms of classification rate and area under the ROC curve. The full model identified about 89% of crash cases in the validation dataset with only 6.5% false positive.

Based on the results from the machine learning procedure, a framework for real-time risk assessment on freeways was proposed. The proposed framework assesses the increased real-time risk depending on the availability of on-line data. The idea behind the proposed framework based on the fact that although the traffic detection and meteorological stations became advanced enough to overcome hardware failures and malfunctions, the challenging weather conditions may interrupt the flow of the data in real-time at some point. Therefore, a reliable and robust framework should be in place at all times. The proposed framework is considered a good alternative for real-time risk assessment on freeways because of its high estimation accuracy, robustness and reliability.

Overall, the proposed multi-level analyses are useful in providing roadway authorities with detailed information on where countermeasures must be implemented and when resources would be devoted. The study also proves that traffic data collected from different detection systems could be a useful asset that should be utilized appropriately to not only alleviate traffic congestion but also to mitigate increased safety risk in order to maximize the benefit of an existing archived data for freeways/expressways authorities as well as for road users.

The multi-level safety analyses demonstrated in this study are considered as the primary element of a proactive traffic management system. The secondary but vital element would be the traffic control techniques (proactive intervention systems) that will be used to achieve the safer operation conditions. Route diversion, ramp metering, Variable Speed Limit (VSL), and Dynamic Message Signs (DMS) can be used as intervention strategies. Among those strategies, VSL systems are proven to reduce recurrent congestion and speed variation, and maintain higher operating speeds on freeways. Integrating VSL and dynamic safety messages based on the estimated risk level within existing Advanced Traveler Information Systems (ATIS) would be a cost-effective added value to these systems. A good message at the right time is the key to gain drivers' trust and compliance to the system which in return will improve the reliability of the system and increase the revenue on toll roads. Micro-simulation could be used to evaluate different scenarios of route diversion, ramp metering, and VSL. In order to come up with the most appropriate dynamic message(s), based on the findings from the statistical models, tailored sets of messages have to be tested at different traffic and weather conditions. Driving simulator and user preference survey could be used as an effective way to achieve such target. In the near future, with the accelerated development of intelligent vehicle technology, the results from this study could be extended to enable even more advanced proactive traffic management systems utilizing IntelliDrive (vehicle-to-vehicle and vehicle-to-infrastructure communication) that will alleviate congestion and promote safety on roadways.

LIST OF REFERENCES

Abdel-Aty, M., and Abdalla, F., Linking Roadway Geometrics and Real-Time Traffic Characteristics to Model Daytime Freeway Crashes Using Generalized Extreme Equations For Correlated Data. Transportation Research Record. Vol. 1897, pp. 106-115, 2004.

Abdel-Aty, M., and Pande, A., Classification of Real-Time Traffic Speed Patterns to Predict Crashes on Freeways. Presented at the 83rd Annual Meeting of the Transportation Research Board , TRB, Washington D.C., 2004.

Abdel-Aty, M., and Pande, A., Identifying Crash Propensity Using Specific Traffic Speed Conditions. Journal of Safety Research. Vol. 36, pp. 97–108, 2005.

Abdel-Aty, M., and Pande, A., Crash Data Analysis: Collective vs. Individual Crash Level Approach, Journal of Safety Research, Vol. 38 Issue 5, pp. 97-108, 2007.

Abdel-Aty, M., and Pemmanabonia, R., Calibrating a Real-Time Traffic Crash-Prediction Model Using Archived Weather and ITS Traffic Data. IEEE Transactions on Intelligent Transportation Systems, Vol. 7(2), pp.167–174, 2006.

Abdel-Aty, M., Dhindsa, A., Coordinated Use of Variable Speed Limits and Ramp Metering for Improving Safety on Congested Freeways. Compendium of papers CD-ROM, Transportation Research Board 86th Annual Meeting, Washington, D.C, 2007.

Abdel-Aty, M., Gayah, V., Considering Route Diversion as a Real-Time Crash Prevention Strategy on Urban Freeways. Presented at the 87th annual meeting of Transportation Research Board, 2008.

Abdel-Aty, M., Gayah, V., Considering Real-Time Crash Risk Reduction on Freeways Using Coordinated and Uncoordinated Ramp Metering Approaches. Journal of Transportation Engineering, Vol. 136(5), pp. 410–423, 2010.

Abdel-Aty, M., Pande, A., Das, A., and Knibbe, W., Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems. Transportation Research Record. Vol. 2083, pp. 153-161, 2008.

Abdel-Aty, M., Pande A., Lee C., Gayah V., and Santos C., Crash Risk Assessment Using Intelligent Transportation Systems Data and Real-Time Intervention Strategies to Improve Safety on Freeway. Journal of Intelligent Transportation Systems, 11(3):107-120, 2007.

Abdel-Aty, M., Radwan, E., Modeling Traffic Accident Occurrence and Involvement. Accident Analysis and Prevention 32 (5), 633–642, 2000.

Abdel-Aty, M., Uddin N., Pande, A., Abdalla, F., and Hsia L. Predicting Freeway Crashes Based on Loop Detector Data Using Matched Case–Control Logistic Regression. Transportation Research Record. Vol. 1897, pp. 88–95, 2004.

Abdel-Aty, M., Uddin N., and Pande A., Split Models for Predicting Multivehicle Crashes During High-Speed and Low-Speed Operating Conditions on Freeways. Transportation Research Record. Vol. 1908, pp. 51–58, 2005.

Abdel-Aty, M. and Pande A., Identifying Crash Propensity Using Specific Traffic Speed Conditions. Journal of Safety Research. Vol. 36, pp. 97–108, 2005.

Abdelwahab, H., and Abdel-Aty, M., Traffic Safety Analysis for Toll Plazas Using Artificial Neural Networks and Logit Models. Transportation Research Record, Vol. 1784, pp. 115-125, 2002.

Agresti, A., Categorical Data Analysis, 2nd Ed. John Wiley and Sons, Inc., 2002.

Aguero-Valverde, J., Jovanis, P.P., Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. Accident Analysis & Prevention 38, 618–625, 2006.

Aguero-Valverde, J., Jovanis, P.P., Identifying Road Segments with High Risk of Weather-Related Crashes Using Full Bayesian Hierarchical Models. Compendium of papers CD-ROM, Paper No. 07-0590, Transportation Research Board 86th Annual Meeting, Washington, D.C., 2007.

Aguero-Valverde, J., Jovanis, P.P., Bayesian Multivariate Poisson Log-Normal Models for Crash Severity Modeling and Site Ranking. Presented at the 88th Annual Meeting of the Transportation Research Board, 2009.

Aguero-Valverde, J., Jovanis, P.P., Spatial analysis of Fatal and Injury Crashes in Pennsylvania. Accident Analysis & Prevention 38, 618–625, 2006.

Ahmed, M., Abdel-Aty, M., and Yu, R., A Bayesian Updating Approach for Real-Time Safety Evaluation Using AVI Data. Transportation Research Record, (Accepted Feb. 2012a).

Ahmed, M., Abdel-Aty, M., and Yu, R., Assessment of the Interaction between Crash Occurrence, Mountainous Freeway Geometry, Real-Time Weather and AVI Traffic Data. Transportation Research Record, (Accepted Feb. 2012b).

Ahmed M., and Abdel-Aty, M., The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Transaction on Intelligent Transportation Systems. doi: 10.1109/TITS.2011.2171052 (In Press, 2011)

Ahmed M., Huang H., Abdel-Aty, M., and Guevara B., Exploring a Bayesian Hierarchical

Approach for Developing Safety Performance Functions for a Mountainous Freeway. Accident Analysis and Prevention, vol. 43(4), pp. 1581–1589, 2011a.

Ahmed M., Yu R., and Abdel-Aty, M. Safety Application of Automatic Vehicle Identification and Real-Time Weather Data on Freeways. Presented at the 18th ITS World Congress, 2011b.

Al-Ghamdi, A. S., Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity. Accident Analysis and Prevention 34, 729-741, 2002.

Allaby P., Hellinga B., and Bullock M., Variable Speed Limit: Safety and Operational Impacts of a Candidate Control Strategy for and Urban Freeway. Proceedings of the IEEE Intelligent Transportation Systems Conference, TP2.3, pp.897–902, 2006.

American Association of State Highway and Transportation Officials (AASHTO). A Policy on Geometric Design of Highways and Streets, Fifth Edition (Washington, DC, ISBN 1-56051-263-6), 2004.

Amit Y., and Geman D., Neural Computation Vol. 9 (7): pp. 1545–1588, 1997.

Amit Y., and Geman D., Shape Quantization and Recognition with Randomized Trees. Neural Computation Vol. 9 (7): pp. 1545–1588, 1997.

Anastasopoulos, P.C., Tarko A. P., and Mannering, F., Tobit Analysis of Vehicle Accident Rates on Interstate Highways. Accident Analysis and Prevention 40, 768–775, 2008.

Anastasopoulos, P.C. and Mannering, F., A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. Accident Analysis and Prevention 41, 153–159, 2009.

Andreescu P. and Frost D., Weather and Traffic Accidents in Montreal, Canada, Journal Climate Research, vol. 9, no. 3, pp. 225–230, 1998.

Andrey, J., and Olley R., Relationships between Weather and Road Safety, Past and Future Directions, Climatol. Bull., vol. 24, no. 3, pp. 123–137, 1990.

Besag, J., Spatial Interaction And The Statistical Analysis Of Lattice Systems. Journal of the Royal Statistical Society, B 36(2), 192–236, 1974.

Bohenberger, K., and May, A., Advanced Coordinated Traffic Responsive Ramp Metering Strategies. PATH Rep. No. UCB-ITS-PWP- 99-19, Univ. of Calif., Berkeley 1999.

Boughton, C. J., Accidents and Geometric Design Research Report ARR44. Vermont South, Victoria, Australia: Australian Road Research Board, 1975.

Breiman, L., Some Infinity Theory for Predictor Ensembles. Technical Report 579, Statistics Dept. University of California Barkley, 2000.

Breslow, N., and Day, N., Statistical Methods in Cancer Research. Volume I – The Analysis of Case Control Studies. Geneva: IARC Scientific Publications No. 32; 1980.

Breslow, N. and Day, N. Statistical Methods in Cancer Research. Volume I – The Analysis of Case Control Studies. Geneva: IARC Scientific Publications No. 32; 1980.

Brodsky, H., and Hakkert, S., Risk of a Road Accident in Rainy Weather. Accident Analysis and Prevention, vol. 20, no. 2, pp. 161–176, 1988.

Brooks, S. P., and Gelman, A., Alternative Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics. 7, 434-455, 1998.

Carson, J., & Mannering, F., The Effect of Ice Warning Signs on Ice-Accident Frequency and Severity. Accident Analysis and Prevention, 33(1), 99–109, 2001.

Ceder, A., Relationship between Road Accidents and Hourly Traffic Flow–II. Accident Analysis and Prevention 14, 34-44, 1982.

Chan, L.Y. & Chen W.C., Data Mining of Tree-Based Models to Analyze Freeway Accident Frequency. Journal of Safety Research 36, 365-375, 2005.

Chang, G., Kao, Y., An Empirical Investigation of Macroscopic Lane Changing Characteristics on Uncongested Multilane Freeways. Transportation Research Part A 25 (6), 375–389, 1991.

Chang, L., Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network. Safety Science 43 541–557, 2005.

Chang L., Chen W., Data Mining of Tree-Based Models to Analyze Freeway Accident Frequency. Journal of Safety Research 36, pp. 365–375, 2005.

Chin, H.C., Quddus, M.A., Applying the Random Effect Negative Binomial Model to Examine Traffic Accident Occurrence at Signalized Intersections. Accident Analysis and Prevention 35, 253-259, 2003.

Chris L., Abdel-Aty, M., Hsia L., Potential Real-Time Indicators of Sideswipe Crashes on Freeways, Transportation Research Record. Vol. 1953, pp. 41-49, 2006

Congdon, P., Bayesian Statistical Modelling. John Wiley & Sons, Inc., 2007.

Detector Handbook: Third Edition-Volume II, Publication No. FHWA-HRT-06-108, 2006.

Dhindsa, A. Evaluating Ramp Metering and Variable Speed Limits to Reduce Crash Potential on Congested Freeways Using Microsimulation. MS thesis, University of Central Florida, Orlando, 2006.

Dion, F., and Rakha, H., Estimating Dynamic Roadway Travel Times Using Automatic Vehicle Identification Data for Low Sampling Rates. Transportation Research Part B: Methodological, Volume 40, Issue 9, pp 745-766, 2006.

El-Basyouny, K., Sayed, T., Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. Transportation Research Record 1950, 9–16, 2006.

El-Basyouny, K., Sayed, T., Urban Arterial Accident Prediction Models with Spatial Effects. Transportation Research Record, 2385, 27-33, 2009a.

El-Basyouny, K., Sayed, T., Accident Prediction Models with Random Corridor Parameters. Accident Analysis and Prevention 41 (4), 1118–1123, 2009b.

El-Basyouny, K., Sayed, T., Collision Prediction Models Using Multivariate Poisson-Lognormal Regression. Accident Analysis and Prevention 41 (4), 820–828, 2009c.

Federal Highway Administration. Synthesis of safety research related to traffic control and roadway elements. Report TS-82-232. Washington, DC: Federal Highway Administration, Office of Research, Development and Technology; 1982

Feng, C., Synthesis of Studies on Speed and Safety. Presented at the 80th Annual Meeting of Transportation Research Board, Washington, D.C., 2001.

Garber, N., and Ehrhart, A., The Effect of Speed, Flow, and Geometric Characteristics on Crash Frequency for Two-Lane Highways. Transportation Research Record, No. 1717, Transportation Research Board, National Research Council, Washington, D.C., pp. 76-83, 2000.

Gerlough, D. L., and Huber M. J., Traffic Flow Theory: a Monograph. Special Report 165, Transportation Research Board, Washington, D.C., 1975

Gilks, W., Best, N., and Tan, K., Adaptive Rejection Metropolis Sampling with Gibbs Sampling, Applied Statistics, 44, 455–472, 1995.

Gilks, W. and Wild, P., Adaptive Rejection Sampling for Gibbs Sampling, Applied Statistics, 41, 337–348, 1992.

Goldstein, H., Multilevel Statistical Models, third edition. Hodder Arnold, London, 2003.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, H., A User's Guide to MLwiN. London: Institute of Education, 1998.

Golob, T., and Recker W., A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways Transportation Research - Part A, Policy and Practice, 38: 53-80, 2004.

Golob, T., and Recker W., Relationships among Urban Freeway Accidents, Traffic Flow, Weather and Lighting Conditions. California PATH Working Paper UCB-ITS-PWP-2001-19, Institute of Transportation Studies. University of California, Berkeley, 2001.

Golob, T., Recker, W., and Alvarez, V., Freeway Safety as a Function of Traffic Flow. Accident analysis and prevention. Vol. 36, pp. 933-946, 2004.

Golob, T., Recker, W., and Alvarez, V., A Tool to Evaluate the Safety Effects of Changes in Freeway Traffic Flow. Presented at the 82nd annual meeting of Transportation Research Board, Washington, D.C., 2003.

Goodwin, L., Weather Related Crashes on U.S. Highways. Federal Highway Administration December 2002, Falls Church, VA: Mitretek Systems, Inc. Accessed on December 2010: http://ops.fhwa.dot.gov/Weather/best_practices/CrashAnalysis2001.pdf

Hadi, M.A., Aruldhas, J., Lee-Fang Chow, Wattleworth, J.A., Estimating Safety Effects of Cross-Section Design for Various Highway Types Using Negative Binomial Regression. Transport. Res. Rec. 1500, 169–177, 1995.

Haque, M.M., Chin H.C., Huang H. Applying Bayesian Hierarchical Models to Examine Motorcycle Crashes at Signalized Intersections. Accident Analysis & Prevention 42(1), 203-212, 2010.

Harb R., Yan X., Radwan E., and Su X., Crash Avoidance Analysis using Classification Trees and Random Forest, Presented at the 87th Annual Meeting of the Transportation Research Board, Washington, DC., 2008.

Harwood D., Council, F., Hauer, E., Hughes, W., Vogt, A., Prediction of the Expected Safety Performance of Rural Two-Lane Highways. Report No. FHWARD- 99-207, 2000.

Hassan H., and Abdel-Aty, M., Exploring Visibility Related Crashes on Freeways Based on Real-Time Traffic Flow Data., Presented at the 90th annual meeting of Transportation Research Board, 2010.

Hauer, E., Harwood, D.W., Council, F.M., Griffith, M.S., The Empirical Bayes Method for Estimating Safety: A Tutorial. Transportation Research Record 1784, 126–131, 2002.

Heydecker, B.G., Wu, J., Identification of Sites for Accident Remedial Work by Bayesian Statistical Methods: An Example of Uncertain Inference. Advances in Engineering Software 32, 859–869, 2001.

Highway Capacity Manual, Transportation Research Board, Washington, D.C., 2000, Chapter 23, Appendix A.

Ho, T., The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20(8), pp. 832–844, 1998.

Hourdos, J., Garg, V., Michalopoulos, P., and Davis, G., Real-Time Detection of Crash-Prone Conditions at Freeway High-Crash Locations, Transportation Research Record. Vol. 1968, pp. 83-91, 2006.

Hovey P.W., Chowdhury, M., 2005. Development of crash reduction factors. Technical Report. Prepared for the Ohio Department of Transportation, 14801.

http://climate.colostate.edu/climateofcolorado.php Accessed in June 29, 2010.

http://www.coloradodot.info/travel/eisenhower-tunnel/description.html Accessed in June 28, 2010.

http://www.fhwa.dot.gov/interstate/faq.htm Accessed in June 25, 2010.

Huang, H., Abdel-Aty, M., Darwiche, A.L.,. County-Level Crash Risk Analysis in Florida: Bayesian Spatial Modeling. Transportation Research Record, 2148, 27-37, 2010.

Huang, H., Chin, H.C., Haque, M., Empirical Evaluation of Alternative Approaches in Identifying Crash Hotspots: Naive Ranking, Empirical Bayes And Full Bayes. Transportation Research Record 2103, 32-41, 2009.

Huang, H., Chin, H.C., Modeling Road Traffic Crashes With Zero-Inflation And Site-Specific Random Effects. Statistical Methods & Applications, 19(3), 445-462, 2010.

Huang, H., Chin, H.C., Haque, M.M., Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40, 45–54, 2008.

Huang H., Abdel-Aty, M., Multilevel Data and Bayesian Analysis in Traffic Safety. Accident Analysis & Prevention, 42(6), 1556-1565, 2010.

Hughes R., and Council F., On Establishing Relationship(S) Between Freeway Safety and Peak Period Operations: Performance Measurement and Methodological Considerations. Presented at the 78th annual meeting of Transportation Research Board, Washington, D.C., 1999.

Ishak, S., and Alecsandru, C., Analysis of Freeway Pre-Incident, Post-Incident, and Non-Incident Conditions Using Second-Order Spatio-Temporal Traffic Performance Measures. Presented at the 84th annual meeting of Transportation Research Board, Washington, D.C., 2005

Jovanis, P., Chang, H., Modeling the Relationship of Accident to Mile Traveled. Transportation Research Record 1068, pp42-51, 1986.

Karlaftis, M. G., & Golias, I. Effects of Road Geometry and Traffic Volumes on Rural Roadway Accident Rates. Accident Analysis and Prevention, 34(3), pp357–365, 2002.

Kim, D-G, Washington, S., Oh, J., Modeling Crash Types: New Insights into the Effects of Covariates on Crashes at Rural Intersections. Journal of Transportation Engineering 132, pp282-292, 2006.

Kockelman, K., and Ma, J., Freeway Speeds and Speed Variations Preceding Crashes, Within and Across Lanes. Presented at the 83rd annual meeting of Transportation Research Board, Washington, D.C., 2004.

Kuhn, S., Egert, B., Neumann, S., and Steinbeck C., Building Blocks for Automated Elucidation of Metabolites: Machine Learning Methods for NMR Prediction. BMC Bioinformatics, Vol. 9 (400), 2008.

Langford, I.H., Leyland, A.H., Rasbash, J. and Goldstein, H., Multilevel modelling of the geographical distributions of diseases. Applied Statistics, 48 (2), 253-268, 1999.

Lee, C., Abdel-Aty, M., Hsia L., Potential Real-Time Indicators of Sideswipe Crashes on Freeways. Presented at the 85th annual meeting of Transportation Research Board, 2006.

Lee, C., Hellinga, B., and Ozbay, K. Quantifying the Effects of Ramp Metering on Freeway Safety. Accident Analysis and Prevention, No. 28, 279-288, 2006.

Lee, C., Hellinga, B., and Saccomanno, F., Assessing Benefits of Variable Speed Limits. Presented at the 83rd annual meeting of Transportation Research Board, Washington, D.C., 2004.

Lee, C., Saccomanno, F., and Hellinga, B., Analysis of Crash Precursors on Instrumented Freeways. Transportation Research Record. Vol. 1784, pp. 1-8, 2002.

Lee, C., Saccomanno, F., and Hellinga, B., Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic. Transportation Research Record. Vol. 1840, pp. 67-77, 2003.

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

Lee, J., & Mannering, F., Impact of Roadside Features on the Frequency and Severity of Run-Off-Roadway Accidents: An Empirical Analysis. Accident Analysis and Prevention, 34(2), 149–161, 2002.

Li,W., Carriquiry, A., Pawlovich, M.,Welch, T., 2008. The choice of statistical models in road safety countermeasures effectiveness studies in Iowa. Accident Analysis and Prevention 40 (4), 1531–1542.

Lord, D., Park, Y.J., Investigating the Effects of the Fixed and Varying Dispersion Parameters of Poisson-Gamma Models on Empirical Bayes Estimates. Accident Analysis and Prevention 40, 1441–1457, 2008.

Lord, D., Washington, S., Ivan, J.N., Further notes on the application of zero-inflated models in highway safety. Accident Analysis and Prevention 39 (1), 53–55, 2007.

Lord, D., Washington, S., Ivan, J.N., Poisson, Poisson-gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. Accident Analysis and Prevention 37 (1), 35–46, 2005.

Lord D., and Mannering F., The Statistical Analysis of Crash-Frequency data: A Review and Assessment of Methodological Alternatives, Transportation Research Part A 44: 291–305, 2010.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., WinBUGS—A Bayesian Modeling Framework: Concepts, Structure, And Extensibility. Statistics and Computing 10, 325–337, 2000.

Ma, J., Kockelman, K.M., Damien, P., A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40, 964–975, 2008.

Madanat, S., and Liu, P., A Prototype System for Real-Time Incident Likelihood Prediction. IDEA Project Final Report (ITS-2), Transportation Research Board, National Research Council, Washington, D.C., 1995.

Malyshkina, N.V., Mannering, F.L., Tarko, A.P., Markov Switching Negative Binomial Models: an Application to Vehicle Accident Frequencies. Accident Analysis and Prevention 41 (2), 217–226, 2009.

Miaou, S.P., and Song, J.J., Bayesian Ranking of Sites for Engineering Safety Improvement:

Decision Parameter, Treatability Concept, Statistical Criterion and Spatial Dependence. Accident Analysis & Prevention, 37, pp699-720, 2005.

Miaou, S.P., Lord, D., Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes. Transportation Research Record 1840, 31–40, 2003.

Miaou, S.P., The Relationship between Truck Accidents and Geometric Design of Road Section: Poisson versus Negative Binomial Regression. Accident Analysis and Prevention 26, pp471-482, 1994.

Milton, John, and Mannering, Fred, The Relationship among Highway Geometrics, Traffic-Related Elements and Motor-Vehicle Accident Frequencies, Transportation 25: 395–413, 1998.

Miranda-Moreno, L.F., Fu, L., Saccomanno, F.F., Labbe, A., Alternative Risk Models for Ranking Locations for Safety Improvement. Transportation Research Record 1908, 1–8, 2005.

Mitra, S., Washington, S., On the Nature of Over-Dispersion in Motor Vehicle Crash Prediction Models. Accident Analysis and Prevention 39, 459–468, 2007.

National Cooperative Highway Research Program,. Cost and Safety Effectiveness of Highway Design Elements. NCHRP Report 197, Transportation Research Board, Washington DC, 1997.

National Highway Traffic Safety Administration, Analyses of Rear-End Crashes and Near-Crashes in the 100-Car Naturalistic Driving Study to Support Rear-Signaling Countermeasure Development, Report No. DOT HS 810 846, 2007.

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

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

Pande, A., and Abdel-Aty, M., Assessment of Freeway Traffic Parameters Leading to Lane-Change Related Collisions. Accident analysis and prevention. Vol. 38, pp. 936-948, 2006a.

Pande, A., and Abdel-Aty, M., Comprehensive Analysis of the Relationship Between Real-Time Traffic Surveillance Data and Rear-End Crashes On Freeways. TRB Annual Meeting CD-ROM, paper no. 06-0016, 2006b.

Pande, A., Das, A., Abdel-Aty, M., Hassan, H., Real-Time Crash Risk Estimation: Are All Freeways Created Equal? Transportation Research Record, In Press, 2011.

Pande, A. and Abdel-Aty, M. Assessment of Freeway Traffic Parameters Leading to Lane-Change Related Collisions. Accident Analysis and Prevention. Vol. 38, pp. 936-948, 2006.

Park, B., Fitzpatrick, K., Lord, D., Evaluating the Effects of Freeway Design Elements. Compendium of papers CD-ROM, Transportation Research Board 89th Annual Meeting, Washington, D.C, 2010.

Park, B. and Yadlapati, S. Development and Testing of Variable Speed Limit Logics at Work Zones Using Simulation. Presented at the Transportation Research Board 82nd Annual Meeting, Washington, D.C., 2003.

Park, E.S., Lord, D., Multivariate Poisson-lognormal Models for Jointly Modeling Crash Frequency by Severity. Transportation Research Record 2019, 1–6, 2007.

Park, S., and Ritchie, S.G., Exploring the Relationship between Freeway Speed Variance, Lane Changing, and Vehicle Heterogeneity. Presented at the 83rd annual meeting of Transportation Research Board, Washington, D.C., 2004.

Persaud, B.N., Hauer, E., Retting, R., Vallurapalli, R., Mucsi, K., Crash Reductions Following Traffic Signal Removal in Philadelphia. Accident Analysis and Prevention 29, 803–810, 1997.

Persaud, B.N., Retting, R.A., Garder, P.A., Lord, D., Safety Effect of Roundabout Conversions in the United States: Empirical Bayes Observational Before-After Study. Transportation Research Record 1751, 1–8, 2001.

Poch, M., Mannering, F.,. Negative Binomial Analysis of Intersection Accident Frequencies. J. Transport. Eng. 122 (2), 105–113, 1996.

Quddus M.A., Modeling Area-Wide Count Outcomes with Spatial Correlation and Heterogeneity: An Analysis of London Crash Data. Accident Analysis & Prevention 40, pp1486-1497, 2008a.

Quddus M.A., Time Series Count Data Models: An Empirical Application to Traffic Accidents. Accident Analysis & Prevention, 40(5), pp1732-1741, 2008b.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RandomForest: Breiman and Cutler's random forests for classification and regression http://cran.r-project.org/web/packages/randomForest/

Rao, J.N.K., Small Area Estimation. Wiley, New York, 2003.

SAS Institute Inc. SAS/STAT® 9.2 User's Guide, Second Edition. Cary, NC, 2009. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#titlepage.htm Accessed in April, 2011.

Savolainen, P.T., Tarko, A.P., Safety Impacts at Intersections on Curved Segments. Transport. Res. Rec. 1908, 130–140, 2005.

Shankar, V.N, Albin, R.B., Milton, J.C., Mannering F.L., Evaluation of Median Crossover Likelihoods with Clustered Accident Counts: An Empirical Inquiry Using the Random Effect Negative Binomial Model. Transp Res Rec 1635:44–48, 1998.

Shankar, V.N., Mannering, F.L., Barfield, W., Effect of Roadway Geometric and Environmental Factors on Rural Freeway Accident Frequencies. Accident Analysis and Prevention 27, 371-389, 1995.

Shankar, V.N., Milton, J.C. and Mannering, F.L. Modeling Accident Frequencies as Zero-Altered Probability Process: An Empirical Enquiry. Accident Analysis and Prevention, 29(6), 829-837, 1997.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, V.D., Bayesian Measures of Model Complexity and Fit (With Discussion). Journal of the Royal Statistical Society, B 64(4), 583-616, 2003.

Srinivasan, K. K., Injury Severity Analysis with Variable and Correlated Thresholds: Ordered Mixed Logit Formulation. Transportation Research Record 1784, 132-142, 2002.

Tam, M., and Lam, W., Application of Automatic Vehicle Identification Technology for Real- Time Journey Time Estimation. Information Fusion, Volume 12, Issue 1, Pages 11-19, January 2011.

Theodore-Bustle, E., Traffic Crash Statistics Report 2008. Florida Highway Safety and Motor Vehicles. 2008.

Traffic Detector Handbook: Third Edition-Volume II, Publication No. FHWA-HRT-06-108, 2006.

Traffic Safety Facts National Highway Traffic Safety Administration, National Center for Statistics and Analysis. US Department of Transportation, 2008. http://www-nrd.nhtsa.dot.gov/Pubs/811170.pdf, Accessed on July 2010.

Tunaru, R., Hierarchical Bayesian Models for Multiple Count Data. Austrian Journal of Statistics. Vol. 31, pp221-229, 2002

Wong, Y., and Nicholson, A., Driver Behavior at Horizontal Curves: Risk Compensation and the Margin of Safety. Accident Analysis and Prevention, 24, 425–436, 1992.

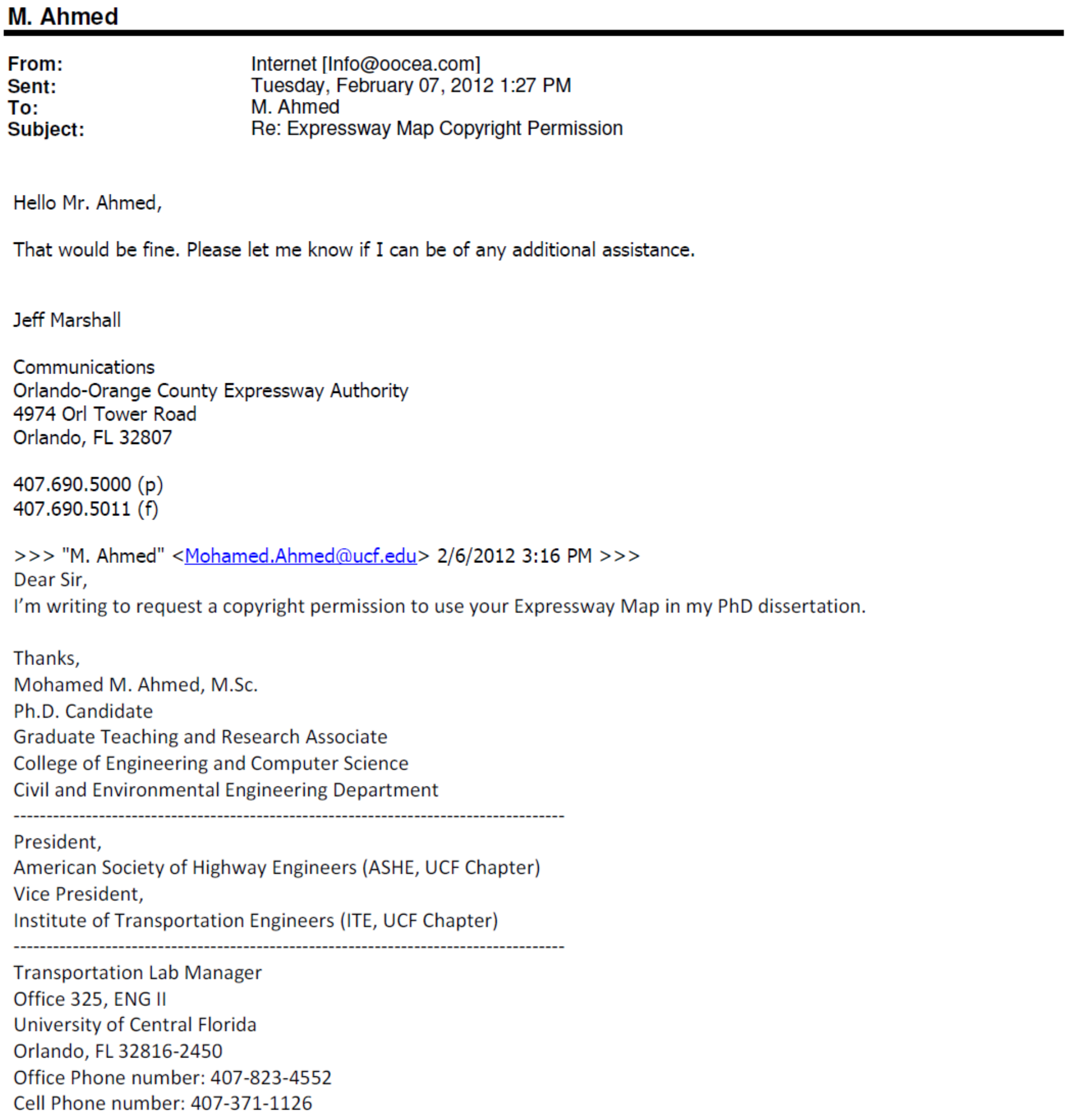
Yan, X., Radwan, E., Mannila, K. K., Analysis of Truck-Involved Rear-End Crashes Using Multinomial Logistic Regression. Advances in Transportation Studies 17, 39-52, 2009.

Ye, X., Pendyala, R.M.,Washington, S.P.,Konduri, K., Oh, J., A Simultaneous Equations Model of Crash Frequency by Collision Type for Rural Intersections. Safety Science 47 (3), 443–452, 2009.

Zhang, C., Ivan, J., N., El-Dessouki, W., M., and Anagnostou, E., N., Relative Risk Analysis for Studying the Impact of Adverse Weather Conditions and Congestion on Traffic Accidents. Presented at the 84th annual meeting of Transportation Research Board, Washington, D.C., 2005.

APPENDIX

Copyright Permission Letter



M. Ahmed

From: Internet [Info@oocea.com]
Sent: Tuesday, February 07, 2012 1:27 PM
To: M. Ahmed
Subject: Re: Expressway Map Copyright Permission

Hello Mr. Ahmed,

That would be fine. Please let me know if I can be of any additional assistance.

Jeff Marshall

Communications
Orlando-Orange County Expressway Authority
4974 Orl Tower Road
Orlando, FL 32807

407.690.5000 (p)
407.690.5011 (f)

>>> "M. Ahmed" <Mohamed.Ahmed@ucf.edu> 2/6/2012 3:16 PM >>>
Dear Sir,
I'm writing to request a copyright permission to use your Expressway Map in my PhD dissertation.

Thanks,
Mohamed M. Ahmed, M.Sc.
Ph.D. Candidate
Graduate Teaching and Research Associate
College of Engineering and Computer Science
Civil and Environmental Engineering Department

President,
American Society of Highway Engineers (ASHE, UCF Chapter)
Vice President,
Institute of Transportation Engineers (ITE, UCF Chapter)

Transportation Lab Manager
Office 325, ENG II
University of Central Florida
Orlando, FL 32816-2450
Office Phone number: 407-823-4552
Cell Phone number: 407-371-1126