CRASH SEVERITY MODELING IN

TRANSPORTATION SYSTEMS

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

AZAD SALIM ABDULHAFEDH

Dr. Timothy Matisziw, Dissertation Supervisor

DECEMBER 2016

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

CRASH SEVERITY MODELING
IN TRANSPORTATION SYSTEMS

presented by Azad Salim Abdulhafedh,

a candidate for the degree of doctor of philosophy of Civil & Environmental Engineering,

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Timothy Matisziw

Dr. Praveen Edara

Dr. Kathleen Trauth

Dr. Ronald McGarvey

ACKNOWLEDGEMENTS

I would like to offer my sincere gratitude to my advisor Dr. Timothy Matisziw, who has supported me throughout my dissertation with his patience and knowledge whilst allowing me the room to work in my own way. I appreciate his help in obtaining Missouri crash data that has been used in this dissertation.

Besides, I would like to thank the rest of my dissertation committee: Dr. Praveen Edara, Dr. Kathleen Trauth, and Dr. Ronald McGarvey, for their time and insightful comments.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES ix

LIST OF TABLES xi

ABSTRACT XIII

CHAPTER 1: INTRODUCTION 1

1.1: FACTORS AFFECTING TRAFFIC CRASHES 3

1.2: Cost of Traffic Crashes 5

1.3: Road Traffic Data Collection Methods 6

1.4: The Contribution of Dissertation to Crash Severity Models 9

CHAPTER 2: BACKGROUND AND OVERVIEW 15

2.1: Overview of Crash Severity Models 15

2.2: Background Literature 17

2.3: The Importance of Collecting Vehicular Crash Data 21

2.4: Sources of Vehicular Crash Data 23

2.4.1: Fatality Analysis Reporting System (FARS) 23

2.4.2: The NASS - GES 24

2.4.3: The NASS - CDS 25

2.4.4: The State Data System (SDS) 26

2.4.5: The Highway Safety Information System (HSIS) 27

2.4.6: Data.gov 27

2.4.7: The U.S. Census Bureau 28

2.4.8: The SHRP2-NDS 28

2.4.9: The Center for Advanced Public Safety (CAPS)................................ 29

2.5: COUNT DATA 29

2.6: COMMON PROBLEMS WITH CRASH DATA 30

2.7: REVIEW OF STATISTICAL APPROACHES OF CRASH MODELING 34

2.7.1: Multiple Linear Regression 34

2.7.2: Poisson Regression 35

2.7.3: Negative Binomial Regression Model (NB)...................................... 36

2.7.4: Poisson-Lognormal Regression Model... 38

2.7.5: Zero Inflated Poisson and Negative Binomial Regression Models 38

2.7.6: Conway-Maxwell Poisson Regression Models 40

2.7.7: Random-Parameter Models 41

2.7.8: Artificial Neural Networks and Fuzzy Logic models......................... 41

2.7.9: Logit and Probit Models 43

2.8: SIGHT DISTANCE ANALYSIS.. 44

2.8.1: Stopping Sight Distance (SSD)... 44

2.8.2: Decision Sight Distance (DSD) 61

2.8.3: Passing Sight Distance (PSD)... 65

2.9: TEMPORAL AUTOCORRELATION 76

2.9.1: Sources of Temporal Autocorrelation... 78

2.9.2: Structure of temporal autocorrelation 78

2.9.3: Detection of Temporal Autocorrelation.. 80

2.9.4: Remedies for Temporal Autocorrelation 87

2.10: SPATIAL AUTOCORRELATION... 87

2.10.1: Weight Matrix of Spatial Autocorrelation 90

2.10.2: Indices of Spatial Autocorrelation 91

2.10.3: Global Indices vs. Local Indices 92

2.10.4: Moran’s I 93

2.10.5: Getis-Ord G statistic 96

2.11: The Multinomial Logistic Regression (MNL) 100

2.11.1: The Advantages of the MNL 101

CHAPTER 3: METHODS 102

3.1: Evaluation of the Temporal Autocorrelation (TA) 102

3.2: Removal of Temporal Autocorrelation 106

3.3: Evaluation of Spatial Autocorrelation 108

3.4: Evaluation of Sight Distance 114

3.4.1: Viewshed Analysis 114

3.4.2: Generating Viewsheds 115

3.4.3: Incorporating Passing Sight Distance in Methodology 120

3.5: Evaluation of Multinomial Logistic Regression (MNL) 121

3.5.1: The Assumptions of MNL 123

3.5.2: The Logit Function and Odd Ratios of MNL 125

3.5.3: Maximum Likelihood Estimation (MLE) 127

3.5.4: The Effects of Independent Variables 128

3.5.5: The Independence of Irrelevant Alternatives (IIA) 129

3.5.6: Multicollinearity 130

3.5.7: The Generalized Hosmer-Lemeshow Statistic 132

3.5.8: The Classification Table of MNL 134

3.5.9: The Pseudo R-squares 134

3.5.10: Estimation of Marginal Effects 137

CHAPTER 4: APPLICATIONS OF MISSOURI CRASH DATA 139

4.1: MISSOURI CRASH DATA 139

4.2: CRASH DATA TERMINOLOGY 139

4.3: MISSOURI GIS ROAD DATA 140

4.4: CRASH SEVERITY CATEGORIES 141

4.5: I-70 CORRIDOR AND BOONE COUNTY ROADS IN MISSOURI 142

4.6: MO ROUTE-5 HIGHWAY FOR ASSESSING THE PSD 145

4.7: THE DIGITAL ELEVATION MODELS (DEMS) 145

4.8: PARTITIONING THE CRASH DATA INTO TRAINING AND TESTING 146

4.9: SELECTION OF INDEPENDENT VARIABLES 147

4.10: APPLICATIONS OF TEMPORAL AUTOCORRELATION (TA) 151

4.10.1: The Significant TA of I-70 (2014) Dataset 152

4.11: APPLICATIONS OF SPATIAL AUTOCORRELATION 152

4.12: APPLICATIONS OF SIGHT DISTANCE 155

4.13: APPLICATIONS OF MULTINOMIAL LOGISTIC REGRESSION 157

4.13.1: Testing the Effects of Independent Variables 158

4.13.2: Testing the IIA Assumption 160

4.13.3: Testing the Generalized Hosmer-Lemeshow Statistic 161

4.13.4: Testing the Multicollinearity 162

4.13.5: The Classification Table 164

4.13.6: The Pseudo R-squares 166

CHAPTER 5: RESULTS 168

5.1: Impacts of Temporal Autocorrelation 168

5.1.1: DW Test Results 168

5.1.2: LM Test Results 169

5.1.3: Differencing Results 170

5.1.4: Cochrane-Orcutt Results 171

5.1.5: The LBQ Test Results 173

5.2: Impacts of Spatial Autocorrelation 176

5.2.1: Global Moran’s I and General Gi* 176

5.2.2: Anselin Local Moran’s I and Local Gi* Statistic 178

5.2.3: Spatial Autocorrelation by a Hybrid Method 189

5.3: Impacts of Sight Distance 192

5.3.1: Sight Distance Along I-70 Corridor 192

5.3.2: Sight Distance Along Boone County Roads 196

5.3.3: Passing Sight Distance along MO Route 5 199

5.4: Impacts of Multinomial Logistic Regression 202

5.4.1: Predicted Odd Ratios for I-70 Corridor 202

5.4.2: Significant Risk Factors for I-70 Corridor 207

5.4.3: Predicted Odd Ratios for Roads in Boone County 209

5.4.4: Significant Risk Factors for Boone County Crashes 212

5.4.5: Marginal Effects for Crashes Along I-70 Corridor 214

5.4.6: Marginal Effects for Boone County Crashes 219

CHAPTER 6: CONCLUSIONS 223

6.1: Conclusion of Results 223

6.2: Potential Impacts of Other Risk Factors 225

6.3: Modeling Crash Points vs. Crash Segments 226

6.4: Effects of Changing DEM’s Resolution 227

6.5: The Amount of Violation of AASHTO Standards 227

REFERENCES 228

VITA 243

LIST OF FIGURES

Figure Page

1.1 Dissertation workflow…………………………………………………………..…14

2.1 AASHTO (2011) criteria for stopping sight distance ……………………………..…46

2.2 AASHTO (2011) SSD criteria on horizontal alignments.……………..……….………55

2.3 SSD parameters used in design of crest vertical curves ….……………………………57

2.4 SSD parameters used in design of sag vertical curves ……………………………...59

2.5 SSD parameters used in design of under passing sag curves ……………………….60

2.6 Recommended AASHTO criteria on DSD ………………………………………….64

2.7 AASHTO (2004) model for PSD calculations ….…………………………………...71

2.8 AASHTO and MUTCD criteria for PSD and marking of NPZs……………………73

2.9 PSD parameters on crest vertical curves ……..………………………………………75

3.1 Observer points in viewshed analysis……...………………………..........……117

3.2 Viewshed parameters ………………………………………………………........118

3.3 Viewshed horizontal and vertical scans...…………………………..……………118

3.4 Classification of road segments ……………………………………………………120

3.5 Passing and no-passing zones ..………………………………………………........121

3.6 Comparison of linear and logistic regression ………………………………........125

4.1 Interstate I-70 and Boone County in Missouri………………………...……..…..143

4.2 Boone County roads and major cities ……………………………………........143

4.3 MO Route 5……………………………………………………………………..…...145

4.4 Speed limit of MO Route 5……………………………………………………..…..157

5.1 Correlogram of I-70 (2013-2015) dataset ……………………………………......175

5.2 Correlogram of Boone County roads (2013-2015) dataset …………………….….…176

5.3 I-70 crash clustering (2013-2015) by Moran’s I ……………………………..…....182

5.4 I-70 crash clustering (2015) by Moran’s I ……………………..………………......182

5.5 I-70 crash clustering (2013-2015) by Gi* …………………..……………………...183

5.6 I-70 crash clustering (2015) by Gi* ……………………….……………………........183

5.7 Boone crash clustering (2013-2015) by Moran’s I ………………….………….........185

5.8 Boone crash clustering (2015) by Moran’s I ……………………………...............186

5.9 Boone crash clustering (2013-2015) by Gi* ……………………………................187

5.10 Boone crash clustering (2015) by Gi* ……………………………………….….…188

5.11 I-70 crash clustering via Hybrid method …….……...……………….…………..…..190

5.12 Boone crash clustering via Hybrid method ………...…………….……………....191

5.13 I-70 segments with potential visibility issues ………...193

5.14 MO I-70 segment 1 with potential visibility issues ……...……………….…..…194

5.15 MO I-70 segment 8 with potential visibility issues ..………………….………195

5.16 MO I-70 segment 3 with potential visibility issues …..…………………….………196

5.17 MO Boone roads with potential visibility issues ….……………..……...…...…...197

5.18 MO Boone roads segment 7 with potential visibility issues …..….……….....….198

5.19 MO Boone roads segment 11 with potential visibility issues……………….........199

5.20 Locations of PZs and NPZs at MO Route 5 ………………..…………….……........201

5.21 End location of segment 8 of MO Route 5 ……………….………………........202

5.22 Start location of segment 27 of MO Route 5 ……………….……….……......202

LIST OF TABLES

Table Page

2.1 Stopping sight distance on level roadways 46

2.2 AASHTO (2011) stopping sight distance on grades 48

2.5 AASHTO recommended decision sight distance 65

2.6 Minimum PSD values for design of two-lane highways 71

2.7 MUTCD warrants for NPZs 72

2.8 Minimum lengths of PZs 73

2.9 PSD design controls for crest vertical curves 76

4.1 Observed crashes along I-70 and roads in Boone County, MO 144

4.2 Observed crashes by severity level 144

4.3 DEMs used in dissertation....................146

4.4 Training and testing datasets....................147

4.5 Analysis group factors 149

4.6 Risk factors and their interpretation 149

4.7 The LR statistic results....................159

4.8 The IIA Assumption results 160

4.9 The Generalized Hosmer-Lemeshow test results....................162

4.10 The VIF results....................163

4.11 I-70 classification table results 164

4.12 Boone roads classification table results 165

4.13 The pseudo R-squares results 166

5.1 DW statistic for I-70 and Boone County roads 169

5.2 LM statistic for I-70 and Boone County roads ……………………………….…....170

5.3 Differencing results for 2014 I-70 data …………………......………………..…...171

5.4 Cochrane-Orcutt results for 2014 I-70 data…………………….……...……….…172

5.5 Overall DW statistic for I-70 and Boone County roads …………………….…...172

5.6 Overall LM statistic for I-70 and Boone County roads ………………….……..…173

5.7 LBQ test results for I-70 ……………………………………………………..173

5.8 LBQ test results for Boone County roads ……………..………..………….....174

5.9 Global Moran's *I* for I-70 and Boone County, MO……………………….......177

5.10 General Gi* for I-70 and Boone County, MO ………………………………………178

5.11 Crash clustering patterns of I-70, MO ………………………………......…...…180

5.12 Crash clustering patterns of Boone County roads, MO ……………….…..…...181

5.13 I-70 crash clustering by Moran's I, Gi*, and Hybrid method ……………….…192

5.14 Boone crash clustering by Moran's I, Gi*, and Hybrid method…….…………..…192

5.15 I-70 segments with potential visibility issues ……………....………..…………193

5.16 Boone roads segments with potential visibility issues ……………………..……...197

5.17 The latitude, longitude, and lengths of NPZs at MO-R5 ……………….……..............199

5.18 Predicted odd ratios for I-70 ………………………………………….....……...204

5.19 Significant risk factors for I-70……………………………………………….....207

5.20 Predicted odd ratios for Boone County crashes ……….……………………..….......210

5.21 Significant risk factors for Boone County crashes …...……………….................213

5.22 Marginal effects for crashes along I-70 ………………...…………….…………215

5.23 Marginal effects for Boone County crashes ………..………………….....……....219

ABSTRACT

Modeling crash severity is an important component of reasoning about the issues that may affect highway safety. A better understanding of the factors underlying crash severity can be used to reduce the degree of crash severity injury, locate road hazardous sites, and adopt suitable countermeasures. In order to provide insights on the mechanism and behavior of the crash severity injury, a variety of statistical approaches have been utilized to model the relationship between crash severity and potential risk factors. Many of the traditional approaches for analyzing crash severity are limited in that they are based on the assumption that all observations are independent of each other. However, given the reality of vehicle movement in networked systems, the assumption of independence of crash incidence is not likely valid. For instance, spatial and temporal autocorrelations are important sources of dependency among observations that may bias estimates if not considered in the modeling process. Moreover, there are other aspects of vehicular travel that may influence crash severity that have not been explored in traditional analysis approaches. One such aspect is the roadway visibility that is available to a driver at a given time that can impact their ability to react to changing traffic conditions, a characteristics known as sight distance. Accounting for characteristics such as sight distance in crash severity modeling involve moving beyond statistical analysis and modeling the complex geospatial relationships between the driver and the surrounding landscape.

To address these limitations of traditional approaches to crash severity modeling, this dissertation first details a framework for detecting temporal and spatial autocorrelation in crash data. An approach for evaluating the sight distance available to drivers along roadways is then proposed. Finally, a crash severity model is developed based upon a multinomial logistic regression approach that incorporates the available sight distance and spatial autocorrelation as potential risk factors, in addition to a wide range of other factors related to road geometry, traffic volume, driver's behavior, environment, and vehicles. To demonstrate the characteristics of the proposed model, an analysis of vehicular crashes (years 2013-2015) along the I-70 corridor in the state of Missouri (MO) and on roadways in Boone County MO is conducted. To assess existing stopping sight distance and decision sight distance on multilane highways, a geographic information system (GIS)-based viewshed analysis is developed to identify the locations that do not conform to AASHTO (2011) criteria regarding stopping and decision sight distances, which could then be used as potential risk factors in crash prediction. Moreover, this method provides a new technique for estimating passing sight distance along two-lane highways, and locating the passing zones and no-passing zones. In order to detect the existence of temporal autocorrelation and whether it's significant in crash data, this dissertation employs the Durbin-Watson (*DW*) test, the Breusch-Godfrey (*LM*) test, and the Ljung-Box Q (*LBQ*) test, and then describes the removal of any significant amount of temporal autocorrelation from crash data using the differencing procedure, and the Cochrane-Orcutt method. To assess whether vehicle crashes are spatially clustered, dispersed, or random, the Moran's *I* and Getis-Ord *Gi** statistics are used as measures of spatial autocorrelation among vehicle incidents. To incorporate spatial autocorrelation in crash severity modeling, the use of the *Gi** statistic as a potential risk factor is also explored. The results provide firm evidence on the importance of accounting for spatial and temporal autocorrelation, and sight distance in modeling traffic crash data.

CHAPTER 1: INTRODUCTION

Vehicular crashes are the world's leading cause of death for individuals between the ages of one and twenty-nine (WHO 2015). Throughout the world, cars, buses, trucks, motorcycles, pedestrians, animals, taxis and other categories of travelers, share the roadways, contributing to economic and social development in many countries. Yet each year, many vehicles are involved in crashes that are responsible for millions of deaths and injuries. Globally, every year, about 1.25 million people are killed in motor vehicle crashes and approximately 50 million more are injured. Following current trends, about two million people could be expected to be killed in motor vehicle crashes each year by 2030 (WHO 2015). Currently, road crashes are ranked as the ninth most serious cause of death in the world, and without new initiatives to improve road safety, fatal crashes will likely rise to the third place by the year 2020 (WHO 2015). In developed countries, road traffic death rates have decreased since the 1960s because of successful interventions such as seat belt safety laws, enforcement of speed limits, warnings about the dangers of mixing alcohol consumption with driving, and safer design and use of roads and vehicles. For example, road traffic fatalities has declined by about 25.0 percent in the United States from 2005 to 2014 and the number of people injured has decreased 13.0 percent from 2005 to 2014 (NCSA 2015). In Canada, the number of road traffic fatalities has declined by about 62.0 percent from 1990 to 2014, and the number of injuries have declined by about 68.0 percent during the same period (Transport Canada 2016). However, traffic fatalities have increased in developing countries between 1990 to 2014 (i.e. 44.0 percent in Malaysia and about 243.0 percent in China) (WHO 2015). Developing countries bear a large share of the burden, accounting for 85.0 percent of annual deaths and 90.0 percent of the disability-adjusted life years. More than one-half of all road traffic deaths globally involve people ages 15 to 44, during their most productive earning years. Moreover, the disability burden for this age group accounts for about 60.0 percent of all disability-adjusted life years. The costs and consequences of these losses are significant. Three-quarters of all poor families who lost a member in a traffic crash reported a decrease in their standard of living, and about 61.0 percent reported having to borrow money to cover expenses following their loss (Beirness and Beasley 2011). The World Bank estimates that road traffic injuries cost 2.0 percent to 3.0 percent of the Gross National Product of developing countries, or twice the total amount of development aid received worldwide by developing countries (World Bank 2015). Crash-related injuries can be prevented or at least minimized by a joint involvement from multiple sectors (i.e. transportation agencies, police, health departments, education institutions) that oversee road safety, vehicles, and the drivers themselves. Effective interventions include design of safer infrastructure and incorporation of road safety features into land-use and transport planning; improvement of vehicle safety features; improvement of post-crash care for victims of road crashes, and improvement of driver behavior, such as setting and enforcing laws relating to key risk factors, and raising public awareness (Mohan 2002).

Modeling of crash data can assist with the development of generalized theories concerning road safety. A range of basic laws have been put forth to help explain the relationship between the occurrence of road crashes and potential risk factors, such as: the universal law of learning, which implies that the crash rate tends to decline as the number of kilometers travelled increases; the law of rare events, which states that rare events, such as environmental hazards, would have more effect on crash rates than regular events; and the law of complexity, which implies that the more complex the traffic situation road users encounter, the higher the probability of crash occurrence (Elvik 2006).

Although transportation agencies often seek to identify the most dangerous road sites, and put great efforts into preventive measures, such as illumination and policy enforcement, the annual number of traffic crashes has not yet significantly decreased. For instance, 35,092 traffic fatalities were recorded in the US during 2015, an increase of 7.2% as compared to the previous year (NCSA 2016). The fatality rate per 100 million vehicle miles traveled (VMT increased 3.7% between 2014-2015. Thirty-five States had more motor vehicle fatalities in 2015 than in 2014. Every month except November saw increases in fatalities from 2014 to 2015, and the highest increases occurred in July and September (NCSA 2016). Given this trend, it is imperative to gain a better understanding of the risk factors that may be associated with traffic crashes.

1.1: Factors Affecting Traffic Crashes

A traffic crash may have many contributing factors, such as those related to driver behavior, road geometry, traffic volumes, vehicle, and environment. The influence of such variables on crash occurrence could significantly vary on a case-by-case basis, but in general, both behavioral factors related to the driver's errors, and non-behavioral factors related to road geometry, traffic flow conditions, vehicle, and environment are thought to significantly affect traffic crashes (Caliendo et al. 2007). Research have revealed that there are generally six major groups of risk factors affecting traffic crash occurrence (Greibe 2003; Delen et al. 2006; Gelman and Hill 2007; Kim et al. 2007):

1. Driver behavior: alcohol and drug use, reckless operation of vehicle, failure to properly use occupant protection devices, the use of cell phones or texting, and fatigue.
2. Vehicle factors: vehicle type, and the engineering and the safety design standards for vehicle performance. For example, the design of windshield glass and the location and durability of gas tanks can increase safety. Passenger protection systems in vehicles (i.e. airbags, safety belts), if used, can eliminate injuries or reduce their severity.
3. Roadway characteristics: road geometries and road side conditions, such as well-designed curves and grades, wide lanes, adequate sight distance, clearly visible striping, flared guardrails, good quality shoulders, roadsides free of obstacles, well-located crash attenuation devices, and well-planned use of traffic signals.
4. Traffic volumes: average annual daily traffic (AADT) or the vehicle miles travelled (VMT). AADT is the average number of vehicles passing a point along a particular road section each day. Thus, AADT represents the vehicle flow over a road section on an average day of the year. VMT refers to the distance travelled by vehicles on roads. It is often used as an indicator of traffic demand and is commonly applied to evaluate mobility patterns and travel trends.
5. Environmental factors: weather conditions, and light conditions.
6. Time factors: the season of the year, the month of the year, weekdays, and the hour of crash occurrence.

1.2: Cost of Traffic Crashes

The highest cost of traffic crashes is in the loss of human lives; however, society also bears the consequences of many costs associated with motor vehicle crashes. Highway crashes currently cost the USA about $1078.0 billion a year, approximately 5.0 percent higher than 2000. Total costs include both economic costs and societal harm (Blincoe et al. 2015). In the year 2010, 3.9 million people were injured and 32,999 killed in 13.6 million motor vehicle crashes in the US (NCSA 2015). The economic costs of these crashes totaled $242.0 billion including lost productivity, medical costs, legal and court costs, emergency service costs, insurance administration costs, congestion costs, property damage, and workplace losses. The $242.0 billion cost of motor vehicle crashes represents the equivalent of nearly $784.0 for each person living in the United States, and 1.6 percent of the $14.96 trillion U.S. Gross Domestic Product for 2010 (Blincoe et al. 2015). When quality of life valuation is considered, the total value of societal harm from motor vehicle crashes in 2010 was $836.0 billion, roughly three and a half times the value measured by economic impacts alone. Lost market and household productivity accounted for $77.0 billion of the total $242.0 billion economic costs, while property damage accounted for $76.0 billion. Medical expenses totaled $23.0 billion. Congestion caused by crashes, including travel delay, excess fuel consumption, greenhouse gases and criteria pollutants accounted for $28.0 billion. Each fatality resulted in an average discounted lifetime cost of $1.4 million. Each critically injured survivor cost an average of $1.0 million (Blincoe 2015). Traffic crashes cost state budgets huge amounts of money every year. For example, traffic crashes cost the state of Missouri in 2013 a total of $981.0 million, and the state of Kansas, a total of $449.0 million (CDC 2016).

1.3: Road Traffic Data Collection Methods

Most studies of traffic related problems begin with the collection of data, vehicle counts in particular. Generally, traffic count collection methods can be classified as one of two categories: intrusive and non-intrusive methods. Intrusive methods typically involve a data recorder and a sensor placing on or in the road (Bar-Gera 2007). The most common intrusive devices are:

- Pneumatic road tubes: rubber tubes placed across the road lanes to detect vehicles from pressure changes that are produced when a vehicle tire passes over the tube. The pulse of air that is created is recorded and processed by a counter located on the side of the road. The main drawback of this technology is that it has limited lane coverage and its efficiency is subject to weather, temperature and traffic conditions.
- Piezoelectric sensors: sensors are placed in a groove along roadway surface of the lane(s) monitored. The principle is to convert mechanical energy into electrical energy. The amplitude and frequency of the signal is directly proportional to the degree of deformation.
- Magnetic loops: this is the most conventional technology used to collect traffic data. The loops are embedded in roadways in a square formation that generates a magnetic field. The information is then transmitted to a counting device placed on the side of the road. This has a generally short life expectancy because it can be damaged by heavy vehicles, but is not affected by bad weather conditions.

Non-intrusive techniques are based on remote observations ranging from human observation to those based on new technologies (Fraser 2007):

- Manual counts: Trained observers gather traffic data such as vehicle occupancy rate, pedestrians and vehicle classifications that cannot be efficiently obtained through automated counts. Equipment needs are rather basic with the observers usually requiring only a tally sheet, mechanical and/or electronic counting devices.
- Passive and active infra-red sensors: the presence, speed and type of vehicles can be detected based on the infrared energy radiating from the detection area. The main drawbacks of this method are the sensor's performance during bad weather, and limited lane coverage.
- Passive magnetic sensors: magnetic sensors can be fixed under or on top of the roadbed. The sensors record the number of vehicles, their type and speed. However, in some operating conditions, the sensors have difficulty differentiating between closely spaced vehicles.
- Microwave radar sensors: these sensors can detect moving vehicles and record vehicle counts, speed and vehicle classification and are not usually compromised by weather conditions.
- Ultrasonic and passive acoustic sensors: these devices emit sound waves to detect vehicles by measuring the time for the signal to return to the device. The ultrasonic sensors can be placed directly over the lane or alongside the road to collect vehicle counts, speed and classification data However, the collection ability of these sensors can be adversely affected by temperature or bad weather.

- Video image detection: video cameras can be used to record vehicle numbers, type and speed by means of different video techniques e.g. trip line and tracking. Video detection systems can be sensitive to weather conditions.

The Floating Car Data (FCD) can be used to collect traffic data by locating the vehicle via mobile phones or GPS over the entire road network. Data such as car location, speed and direction of travel can then be sent anonymously to a central processing center. After being collected and extracted, useful information can be redistributed to the drivers on the road (Robichaud and Gordon 2003).

Two very important traffic measures are average annual daily traffic (AADT) and vehicle miles travelled (VMT). These two traffic variables, usually derived from fixed sensors measurements, play a key role in traffic crash analysis and policy decisions (Sliupas 2006). AADT is the average (calculated over a year) number of vehicles passing a point along a particular counting section each day. Thus, AADT represents the vehicle flow over a road section (e.g. highway segment) on an average day of the year. Methods for calculating AADT are generally based on data from two types of counts: permanent automatic traffic counts and short-period traffic counts. A combination of these two measurements is generally used to obtain an AADT estimate over a larger road network. In the US, the factoring method is a common methodology used to estimate AADT. This method has been adopted by many transportation agencies as a standard protocol corresponding with federal guidelines. The 2013 Traffic Monitoring Guide (TMG 2013) serves as a reference document that provides general guidance on the development of traffic monitoring programs for highway agencies. In particular, the TMG provides guidance on the collection of traffic volume, vehicle classification, and weight information (Ehlert et al. 2006). VMT refers to the distance travelled by vehicles. It is often used as an indicator of traffic demand and for analyzing mobility patterns and travel trends. It plays a key role in various important decision-makings such as air quality compliance, roadway pavement maintenance, and crash analysis. There are four methods commonly used to calculate VMT (Fricker and Kumapley 2002):

- Odometer readings (vehicle-based method) - at regular vehicle inspections, the average distance travelled by the vehicles is determined and then multiplied by the number of road vehicles.
- Traffic counts (road-based method) - for one considered link, the VMT is calculated by multiplying the AADT by the length of the link. VMT for a roadway can then be obtained by summing the VMT of each segment.
- Driver survey - questionnaires sent to households with one or more cars soliciting information such as the number of miles driven by each vehicle during the whole year and unit consumption.
- Fuel consumption - the volume of road traffic is estimated from information about fuel supply and fuel consumption as derived from estimates of miles driven per fuel gallon for typical types of vehicles.

1.4: The Contribution of Dissertation to Crash Severity Models

Modeling of vehicle crash severity is an important component of reasoning about highway safety. Insights resulting from models of crash severity can be used to reduce the degree of crash severity injury, locate road hazardous sites, and adopt suitable countermeasures. To this end, this dissertation explores modeling approaches and considerations that will hopefully improve the quality of models of crash severity. In particular, the following research themes are addressed:

1) Modeling Highway Sight Distance: A GIS-based viewshed analysis is developed to assess existing stopping and decision sight distances along roadways. This method can be used to identify the locations that potentially might not conform to AASHTO (2011) criteria regarding stopping sight distance and decision sight distance. Additionally, the level of sight distance for sections of the roadway could also be used as a potential risk factor in models of crash severity and/or prediction.
2) Locating No-Passing Zones along two-lane highways: The GIS-based viewshed analysis presents a new method for estimating the passing sight distance on two-lane highways, and hence, assisting in the identification of passing zones and no-passing zones along two-lane highways. An application of the methodology to MO Route-5, a two-lane highway, is conducted to assess the effectiveness of this method.
3) Detecting Temporal Autocorrelation in crash data models: Temporal autocorrelation (also called serial correlation) refers to the relationship between successive values (i.e. lags) of the same variable. Although it has long been a major concern in time series models, it is also a very important consideration in crash severity models as well (Washington et al 2010; Lord and Mannering 2010; Savolainen et al. 2011). However, in-depth treatments of temporal autocorrelation in crash models are lacking. To this end, temporal autocorrelation is thoroughly investigated among the

time independent variables in crash data using several test statistics to detect the amount of temporal autocorrelation and its level of significance in crash data. The tests employed are: 1) the Durbin-Watson (*DW*) test; 2) the Breusch-Godfrey (*LM*) test; and 3) the Ljung-Box Q (*LBQ*) test.

4) Removal of Temporal Autocorrelation: When temporal autocorrelation is statistically significant in crash data, it could adversely bias parameter estimates. As such, if present, temporal autocorrelation should be removed prior to use in crash modeling. In this dissertation, two procedures are presented to remove the temporal autocorrelation: 1) differencing; and 2) Cochrane-Orcutt method.

5) Incorporating Spatial Autocorrelation in Crash severity models: Given the spatial nature of vehicle crashes, the potential existence of spatial autocorrelation among crash incidents is a serious concern in crash modeling and if not appropriately accounted for, can bias parameter estimates (Quddus 2004; Washington et al 2010; Lord and Mannering 2010; Savolainen et al. 2011). To determine if the vehicle crashes are spatially clustered, dispersed, or random, two indices of spatial autocorrelation are employed: Moran's *I* and Getis-Ord G_i* statistic. In addition, this dissertation explores integration of the G_i* statistic in crash modeling as a potential risk factor, one whose use has not been reported in prior research.

6) Analysis of Spatial Autocorrelation via a Hybrid Method: In this dissertation, a new method for determining spatial autocorrelation of crashes is presented by combining both Moran's *I* and *Gi** statistic to examine the spatial clustering patterns of crashes.

7) Multinomial Logistic Regression-Testing Outcomes: In this dissertation, a multinomial logistic regression (MNL) approach is applied to model the relationships of crash severity categories (i.e. fatal, disabling injury, minor injury, property-damage-only) with the independent variables. Although there are a few applications of the MNL in the literature regarding crash modeling, this dissertation presents several new outcome results in applying the MNL that have yet to be reported including: 1) the use of odd ratios as regression estimates instead of regression coefficients to interpret the results of prediction; 2) details of testing of the assumption of the independence of irrelevant alternatives that is very important in the MNL applications, using the Hausman specification test; 3) consideration of the generalized Hosmer-Lemeshow test as an important goodness of fit measure to assess whether or not the observed incidents match the predicted incidents; 4) the use of the classification table as a measure of goodness of fit to determine the percent of corrected prediction cases; 5) testing for the multicollinearity among the independent variables as precondition assumption; and 6) the use of the pseudo R squares as potential measures of goodness of fit.

8) Incorporating a Wide Range of Independent Variables in Crash Modeling: Past research has only employed a limited number of independent variables in crash modeling. As such, this dissertation seeks to investigate the use of a wide range of independent variables (i.e. risk factors) that can be obtained from crash data in the modeling process. In particular, 18 risk factors are considered in the analysis, in addition to the spatial auto correlation index *Gi** and the modeled roadway sight distance.

In order to provide insights on the mechanism and behavior of the crash severity injury and to illustrate the developed analysis approaches, three case-studies in the State of Missouri are considered: 1) Interstate I-70; 2) roadways in Boone County, MO, and 3) MO Route-5. The study sites of I-70 corridor and Boone County roads were used to model crash severity along both of them, while MO Route 5 was used to locate passing and no-passing zones along it. Three years of Missouri crash data (2013-2015) are used in the analysis. The response variable (i.e. crash severity) in the MNL specification is modeled with four possible outcome categories: 1) property-damage-only; 2) minor-injury; 3) disabling-injury; and 4) fatal injury.

This dissertation is organized as follows: Chapter 2 provides an overview of crash severity models, a background literature of crash modeling, the importance of collecting crash data, the sources of crash data in the US, common problems with crash data, an overview of crash modeling methods, the AASHTO (2011) sight distance criteria, the concept of temporal autocorrelation, the concept of spatial autocorrelation, and an overview of the multinomial logistic regression. Chapter 3 discusses the theoretical concepts and the general methodologies developed for the temporal autocorrelation, spatial autocorrelation, sight distance, and the multinomial logistic regression. Chapter 4 describes the data used in this analysis, the selection of the independent variables, and the applications of Missouri crash data in regard to temporal autocorrelation, spatial autocorrelation, sight distance, and multinomial logistic regression. Chapter 5 presents the results of the analysis regarding the temporal autocorrelation, spatial autocorrelation, sight distance, and the multinomial logistic regression. Chapter 6 provides conclusions of the research presented in this dissertation. Figure 1.1 shows the dissertation workflow and the methodologies applied at each step of the research.

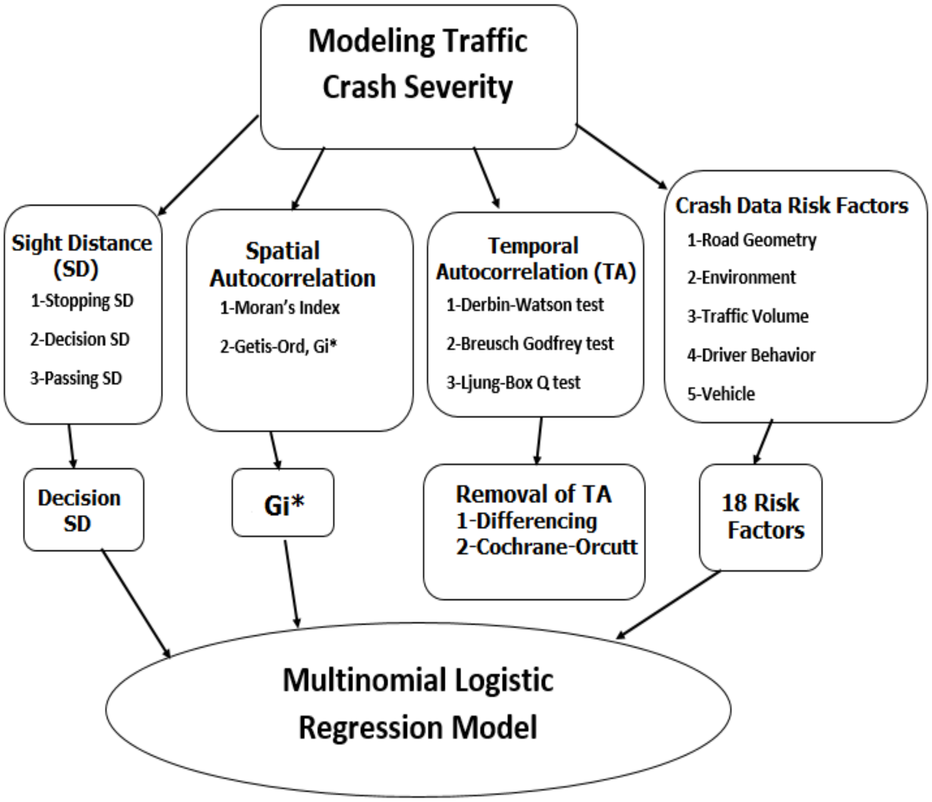

Figure 1.1: Dissertation workflow

CHAPTER 2: BACKGROUND AND OVERVIEW

This chapter presents a general background and an overview of some aspects involved in the crash modeling. First, an overview of crash severity models is presented. Second, a background literature of crash modeling is reviewed. Third, the importance of collecting crash data is presented. Fourth, sources of crash data in the US are detailed and then the structure of those data are discussed. Fifth, common problems with crash data are detailed. Sixth, an overview of crash modeling methods is reviewed. Next, the AASHTO (2011) sight distance as a potential risk factor is detailed. Next, the concept of temporal autocorrelation is discussed. Next, the concept of spatial autocorrelation as potential risk factor is explained in detail. Lastly, the multinomial logistic regression is presented as the modeling process in this dissertation, and its advantages in crash severity modeling.

2.1: Overview of Crash Severity Models

Vehicular crash data can be used to model both the frequency of crash occurrence and the degree of crash severity. Crash frequency refers to the prediction of the number of crashes that would occur on a specific road segment or intersection in a time period (Lord and Mannering 2010). Crash severity methods generally explore the relationship between crash severity injury categories and contributing factors such as driver behavior, vehicle characteristics, roadway geometry, and road-environment conditions. Modeling of crash severity is considered more informative than simply predicting the frequency of crashes (Washington et al. 2010). The data used in modeling crash severity is often attributed with many details relating to the crash occurrence (i.e. such as the number of vehicles involved, age of victims, weather conditions, types of vehicles involved, and crash type) which can be integrated in statistical models (Savolainen et al. 2011). Since the dependent variable (i.e. crash severity) usually has two or more outcome categories (i.e. fatal, injury, property-damage-only), logit and probit models are often used to model the severity of crash data. Discriminant analysis could also be used to model crash severity, but given its rigid assumptions, logit and probit models have been viewed as preferable (Washington et al. 2010; Greene 2012). Binary models consider two response outcomes (i.e. fatal vs. non-fatal or injury vs. property-damage-only), and multinomial models consider three or more response outcomes. Crash severity models can be generally classified as either nominal or ordinal. The nominal models include statistical methods, such as: multinomial logit models; nested logit models, sequential logit models, and mixed logit models. Ordinal models, include: ordered logit models, ordered probit models, and ordered mixed logit models. Based on the existing literature, the multinomial logit models and ordered probit models have been found to be the most prominent types of models used for traffic crash severity analysis. Although there is no consensus on which model is the best, as the selection of the model is often governed by the characteristics of the data, many researchers have opted for nominal models over ordinal models. The rationale for this choice is likely due to the influence that independent variables in ordinal models could exert on the ordered discrete outcome probabilities. That is, in closely related categories (i.e. no injury and possible injury) there may be some shared unobserved effects among adjacent injury categories. Failing to account for such correlation could generate incorrect inferences (Washington et al 2010; Savolainen et al. 2011). Others still prefer ordinal models due to their simplicity and overall performance, especially when less detailed data are available. In addition, the applications of the logistic models are preferred over the applications of probit models due to the simplicity of their calculation (Washington et al. 2010). The most popular and convenient model used in the analysis of crash severity is the multinomial logistic regression that is derived under the assumption of the independence of irrelevant alternatives (IIA). IIA refers to the situation where adding or deleting alternative severity outcome categories does not affect the prediction among the remaining severity outcomes. This means that the odd ratios produced by the logit function for any pair of severity outcomes are determined without reference to the other categories that might be available (McFadden et al. 1976; Hausman 1978; Washington et al. 2010; Savolainen et al. 2011), and therefore it must be checked in the modeling process. If the IIA does not hold, then other models must be considered, such as the nested logit models or the mixed logit models (Washington et al. 2010; Savolainen et al. 2011).

2.2: Background Literature

Modeling crash severity is an important research area in highway safety, given its potential for identifying contributing factors that could then be addressed by transportation policy. In crash modeling research, a wide variety of statistical approaches such as the binary and the multinomial logit models, nested logit models, mixed logit models and ordered probit models have been investigated. For example, Abdel-Aty (2003) apply the ordered probit model to predict crash severity on roadway sections, signalized intersections and toll plazas in Florida. Winston et al. (2006) consider the drivers' decisions to own a vehicle with airbags and/or anti-lock brakes with their probability of being crash-involved and the severity of their crash. They develop a multivariate severity model that relate four binary outcomes: the choice of having airbags, the choice of having anti-lock brakes, the likelihood of being involved in a crash, and the likelihood of such a crash resulting in an injury. Milton et al. (2008) apply a mixed logit model in their research that used the injury outcome of the crash using limited crash data to investigate the proportion of crashes of each severity level on a specific roadway segment over a specified time period. Then, they determine the number of crashes by severity level without the need for detailed crash-specific data. Lee and Abdel-Aty (2008) develop binary severity probit models based on the assumption that drivers with passengers may drive differently than those without passengers. Both the sequential logit or probit models have been applied, which allow the severity categories across the ordered response levels to use separate coefficients for explanatory variables. For instance, Yamamoto et al. (2008) show that sequential models could provide good estimates similar to multinomial logit models when underreporting is a concern. They show that the parameter estimates remain unbiased except for the constant terms. Malyshkina and Mannering (2009) develop a two state Markov switching multinomial logit model to study crash injury severity under the hypothesis that there could exist two states of roadway safety (i.e. safe and unsafe), which may be caused by unobservable risk factors that influence crash severity, assuming that roadway entities can switch between the two states over time. Chang and Wang (2006) use a regression tree approach, which is a data mining technique that does not require a parametric assumption of the relationship between the degree of injury severity sustained and its associated explanatory variables. They show that this approach can provide an efficient technique, but less than the interpretive capabilities of discrete outcome models. Bham et al. (2012)

apply a multinomial logistic regression to model the severity injury of different vehicle collision patterns in urban highways in Arkansas, and recommended the use of the MNL over other models.

Early crash analysis models were generally based on simple multiple linear regression methods assuming normally distributed errors. However, researchers soon discovered that crash occurrence could be better fitted with a Poisson distribution. Hence, a Poisson regression model based upon a generalized linear framework was soon adopted over conventional multiple linear regression techniques. Several such Poisson regression approaches for exploring the relationship between the risk factors and crash frequency have been proposed (Park and Lord 2007; Ma et al. 2008; El-Basyouny and Sayed 2009). However, it has been found that Poisson regression approaches have one important constraint - that the mean must be equal to the variance – which if violated, the standard errors estimated by the maximum likelihood method, will be biased, and the test statistics derived from the model will be incorrect. Recent studies have shown that crash data are usually over-dispersed, when the variance exceeds the mean, therefore, incorrect estimation of the likelihood of crash occurrence could result in applications of the Poisson regression model (Lord and Mannering 2010). In efforts to overcome the problem of over-dispersion, researchers began to employ the Negative Binomial (NB) distribution (also called the Poisson-Gamma) instead of the Poisson distribution, which relaxes the mean equals to variance constraint, and hence can accommodate over-dispersion in crash data counts (Lord and Mannering 2010). NB models have been widely used in crash frequency modeling (Kim et al. 2007; Lord and Bonneson 2007; El-Basyouny and Sayed 2009; Daniels et al. 2010; Malyshkina and Mannering 2010;

Geedipally et al. 2012). However, NB models have some limitations such as the inability to handle under-dispersion of crash counts when the mean of the crash counts is higher than the variance. Although rare, this phenomenon can arise when the sample size is very small, leading to erroneous parameter estimates (Lord 2006; Oh et al. 2006). To address the limitations of NB models, Poisson-lognormal models have been proposed, in which the error term is Poisson-lognormal rather than gamma- distributed to better handle the under-dispersed crash counts (Lord and Miranda 2008; Aguero-Valverde and Jovanis 2008; Daniels et al. 2010). Another widely used type of crash prediction model is the zero-inflated Poisson and zero-inflated negative binomial models, which have been introduced mainly to deal with the over-dispersion problem caused by excessive zeroes (i.e. locations where no crashes can be observed) in traffic data counts. The zero-inflated models have shown great flexibility, although their applicability in crash prediction has been criticized because of the long term mean equals zero in the safe state that could produce some biased estimates (Lord and Mannering 2010; Malyshkina and Mannering 2010). Generalized additive modeling approaches have also been proposed which provide smoothing functions for the explanatory variables. However, these models typically include more parameters than the traditional count models, and therefore their applicability to the crash prediction has been very limited (Xie and Zhang 2008; Li et al. 2009). Random-parameters models have been applied to take the effect of the unobserved heterogeneity from one roadway site to another, however, their application in practice has been very limited (Milton et al. 2008; Anastasopoulos and Mannering 2009; Washington et al. 2010). The finding that road crashes are poorly explained by linear functions of independent variables, has encouraged the exploration of non-linear approximators such as fuzzy logic and neural networks. For example, two fuzzy logic models have been used for predicting the crash frequency on wet pavement and results indicated that the fuzzy logic models exhibited better characteristics over nonlinear regression models (Xiao et al. 1999). In another application, Meng et al. (2009) use a fuzzy logic approach for prediction of urban highway crash occurrence and find the use of fuzzy sets in crash prediction is indeed a viable approach. Neural networks have been applied to highway safety applications as predictive tools, such as in driver behavior analysis, pavement maintenance, vehicle detections, traffic signal control, and vehicle emissions, however, their application to crash analysis has been limited (Abdelwahab and Abdel-Aty 2002; Riviere et al. 2006; Xie and Zhang 2008). For instance, Chang (2005) utilize artificial neural network to analyze the freeway crash frequency in Taiwan, and indicate that an artificial neural network can provide a consistent alternative method for analyzing crash frequency. (Delen et al. 2006) apply a group of artificial neural networks to model the non-linear relationships between the injury severity levels and crash-related factors. Their findings indicate that artificial neural network models can predict crashes more effectively than the traditional statistical methods.

2.3: The Importance of Collecting Vehicular Crash Data

Vehicular crash data are used to respond to requests from the Congress, federal agencies, state and local governments, universities and research organizations, highway safety communities, the media, and private citizens. Accurate data are required to support the development, implementation, and assessment of highway safety programs aimed at reducing crash tolls. An example of the practical importance of collecting and maintaining vehicular crash data is the recent emerging of the crash data retrieval tools, commonly referred to as the vehicle black boxes. Based upon a rule imposed by the National Highway Traffic Safety Administration (NHTSA), most vehicles manufactured and sold in North America after 2012 are equipped with event data recorders (EDRs) that collect, store, and retrieve vehicle crash event data. The EDRs can help law enforcement investigating vehicle crashes to recover crucial crash data parameters from a vehicle that has been involved in a crash, including pre-crash data that will help better understand important factors that led to the crash occurrence (NHTSA Ruling 2010). The anticipated availability of new crash data from vehicle black boxes could lead to important developments in the field of crash frequency and road safety (Lord and Mannering 2010). Another practical example is the use of the Crash Outcome Data Evaluation System (CODES), which is a program managed by NHTSA, to link crash records to injury outcome records collected at the scene by emergency medical services. CODES data has been utilized to improve traffic safety issues in different ways, such as examining whether the increased crash rates for teen drivers have resulted in an increased injury to their passengers, and exploring the seat belt usage in preventing injuries and fatalities. CODES data has also been used to inform and educate traffic safety decision-makers at federal, state, and local levels in many circumstances, for instance, providing federal and state legislators with CODES reports on the importance of seat belt use in preventing injuries and fatalities; delivering data to the state highway administrations to develop long-term, statewide strategic plans for traffic and highway safety; and publishing CODES fact sheets that can help educate the public (NHTSA CODES 2011).

2.4: Sources of Vehicular Crash Data

In the U.S., a variety of efforts to collect, maintain and/or distribute information on vehicular crash data have been utilized. Some of the crash data sources that are publicly available are listed below:

2.4.1: Fatality Analysis Reporting System (FARS)

FARS is an online database of fatal motor vehicle crashes that documents all fatalities that occurred within the 50 States since 1975. FARS qualifying crashes had to involve a motor vehicle traveling on a public traffic way, and must have resulted in the death of a motorist or a non-motorist within 30 days of the crash. FARS is administered by the National Center for Statistics and Analysis (NCSA) within the National Highway Traffic Safety Administration (NHTSA). FARS data are collected from each State's government by trained state employees, who are responsible for gathering, and transmitting their state's data to NCSA in a standard format. After the data file is created, quality checks are performed on the data, and the electronic data are made available online to the public in Statistical Analysis System (SAS) data files as well as Database Files (DBF).The main SAS data files include: the Accident file, which contains information about crash characteristics and environmental conditions at the time of the crash; the Vehicle file, which contains information describing the in-transport motor vehicles and the drivers of in-transport motor vehicle who are involved in the crash; the Person file, which contains information describing all persons involved in the crash including motorists and non-motorists (e.g., pedestrians); the Damage file, which contains information about all areas on the vehicle that were damaged in the crash; the Drimpair file, which contains information about physical impairments of drivers of motor vehicles; the Factor file, which contains information about vehicle circumstances that may have contributed to the crash; the Violatn file, which contains information about violations that were charged to drivers; and the Vindecode file, which contains vehicle descriptors based on the vehicle's VIN. The temporal coverage of FARS data includes some variables such as, the time of the crash, the date, the month, and the year. The spatial coverage of FARS data includes, the latitude and longitude coordinates of each crash location. The FARS data are generally complete, reliable, and publicly available online (NHTSA-FARS 2016). However, one of the FARS data weaknesses is that FARS data cannot be downloaded for multiple years at a time due to the system complexities, and when data is downloaded from FARS website, the user can obtain data by only one variable at a time. In addition, as mentioned above, the FARS data does not provide the injury-severity only crashes, and property- damage only crashes.

2.4.2: The NASS - GES

The National Automotive Sampling System (NASS) - General Estimates System (GES) obtains its data from a representative crash sample selected from more than five million police-reported crashes annually in the US. These crashes include those that result in a fatality or injury and those involving major property damage as well. The data are obtained by NASS - GES data collectors in 60 geographic sites across the United States. These data collectors make visits to approximately 400 police agencies within the 60 sites, where they randomly sample about 50,000 crash per year. NASS-GES data are made available to the public in Statistical Analysis System (SAS) data files as well as Database Files (DBF). The main SAS data files of NASS-GES include similar FARS files mentioned above. The temporal coverage of the NASS-GES data includes variables such as, time of the crash, the date, the month, and the year. The spatial coverage only includes the land use of the crash location without providing the latitude and longitude of the crash location or the x, y coordinates. One weakness in NASS-GES data is that it uses a weighted data element that produces the overall national estimates that may differ from the true state-level values because they are based on a probability sample of crashes among the country, and this cannot give the accurate state-level estimates, which decreases the reliability of the data. Another weakness is that the NASS-GES data are obtained either directly from the police accident report (PAR) or by interpreting the information provided in the PAR through reviewing the crash diagram, or combinations of data elements on the PAR. Because of this interpretation, an important portion of data can be missing in the system (NASS-GES 2016).

2.4.3: The NASS - CDS

The National Automotive Sampling System (NASS) - Crashworthiness Data System (CDS) obtains its data from 24 geographic sites in the US. These data are weighted to represent all police reported motor vehicle crashes occurring in the USA during the year including light vehicles, such as, passenger cars, SUVs, and vans. The NASS-CDS files are available in a Statistical Analysis System (SAS) dataset, and contain similar FARS files. The NASS- CDE system provides temporal coverage of data through variables such as, time of the crash, the date, the month, and the year. There is no spatial coverage within the NASS-CDS data, as it does not provide the latitude and longitude of the crash location nor the x, y coordinates. One weakness of the NASS-CDS data is that the data from these crashes are weighted to produce national estimates, and cannot give the state-level estimates, which decreases the reliability of data (NASS-CDS 2016).

2.4.4: The State Data System (SDS)

The State Data System (SDS) is maintained by NHTSA's National Center for Statistics and Analysis (NCSA), and only thirty-two states are participating in the system, including the state of Missouri. While the (FARS) only has fatal crash data, SDS provides data on injury and property-damage-only crashes as well. In contrast to the data in (NASS-GES), the SDS consists of census data taken directly from police accident reports. The law enforcement agencies within a state are the primary source of information on crashes occurring within a state. All states have requirements for documenting fatal, injury or property damage crashes (with damage above a certain dollar threshold). Each participating state has its own reporting system, for instance, in the state of Missouri, the Missouri Statewide Traffic Accident Records System (STARS) is managed by the Missouri State Highway Patrol (MSHP), and all Missouri law enforcement agencies are required by law to submit a Missouri Uniform Traffic Crash Report to STARS if a traffic crash occurred that involves a death, a personal injury, or a property damage. STARS involves many recording files, such as, the Crash and Personal Severity, which includes fatal, personal injury, and property damage; the Crash Circumstances file, which includes motorcycles crashes by year; Speed Involved Traffic Crash file; Alcohol Involved Traffic Crash file; Young Driver Involved Traffic Crash file; and Mature Driver Involved Traffic Crash file. All files are provided in excel and pdf format, complete, reliable, and available online for the public (MSHP 2016). The temporal coverage of the SDS data includes variables such as, time of the crash, the date, the month, and the year. The spatial coverage only includes the x, y coordinates of the crash locations in only some spots. One weakness of the SDS data is that it does not provide a comprehensive list of risk variables and details that exist in the FARS and NASS-GES systems (NHTSA-SDS 2016).

2.4.5: The Highway Safety Information System (HSIS)

The Highway Safety Information System (HSIS) is a highway data system funded by the U.S. Federal Highway Administration (FHWA), with data voluntarily provided to HSIS by the participating states, which are California, Washington, Minnesota, Illinois, Ohio, Maine, and North Carolina. HSIS began operation in 1987, and the participating states were selected based on their data availability, quantity, and quality of data. HSIS supports the FHWA safety research program, and can be accessed online by researchers, universities, and safety professionals. The HSIS files are available in a (SAS) format, and the main files include four basic files namely; the Accident file, the Vehicle file, the Occupant file, and the Roadway file. The temporal coverage of the HSIS data includes variables such as, time of the crash, date, month, and the year. The spatial coverage only includes the section length, and the milepost of the crash location without providing the latitude and longitude of the crash location nor the x, y coordinates. The HSIS data are generally complete with very few missing data, reliable, and publicly available. One weakness of the HSIS data is that it does not cover all states within the US, and also their main files should be merged in order to get the required information (HSIS 2016).

2.4.6: Data.gov

The Data.gov is a federal open US government online database that includes all states, and local government's metadata describing their open data resources. Data.gov began operation in 2009, and is managed and hosted by the U.S. General Services Administration, Office of Citizen Services and Innovative Technologies, and follows the

Project Open Data schema that includes fields, such as title, description, tags, publisher, etc. for every data set displayed on the website. Different data topics are available, such as Agriculture, Health, Business, Climate, Energy, Finance, and Science. The transportation statistics series consists of analyzed statistical information on motor fuel, vehicle crashes, motor vehicle registrations, driver licenses, highway user taxation, highway mileage, travel, and highway finance. The files are available in CSV format, and can be freely downloaded without registration (Data.gov 2016).

2.4.7: The U.S. Census Bureau

The U.S. Census Bureau is part of the Department of Commerce, and is overseen by the Economics and Statistics Administration. The transportation section within the online database provides data on civil air transportation, water transportation, revenues, passenger and freight traffic volume, trains, highway mileage and finances, highway crash data, characteristics of public transit, and railroads. Data are available in excel format for public use (The U.S. Census Bureau 2016).

2.4.8: The SHRP2-NDS

The Strategic Highway Research Program 2- Naturalistic Driving Study (SHRP2-NDS) is an online database related to the Transportation Research Board (TRB)'s second safety project for an in-vehicle driving behavior field study collected from naturalistic driving data and associated participant, vehicle, and crash-related data. The project was conducted by six site contractors located at geographically distributed data collection sites throughout the United States and more than 3,000 individuals participated in the study. Given that the SHRP 2- NDS is a federally funded study that involves human subjects, the collection of the data and its use in analysis are subject to the approval of institutional review boards. The SHRP 2-NDS database is managed by the Virginia Tech Transportation Institute, and researchers interested in accessing the data must demonstrate that they are qualified researchers seeking the data for research purposes (SHRP2-NDS 2016).

2.4.9: The Center for Advanced Public Safety (CAPS)

The Center for Advanced Public Safety (CAPS) is a research center at the University of Alabama that deals with vehicular crash data, and traffic safety improvements, among other research areas. CAPS has developed a tool for crash data analysis called the Critical Analysis Reporting Environment (CARE), which has many useful analytical functions such as, frequency distributions, cross-tabulations, and statistical significance tests. CARE can compare the performance of one subset of data against another in terms of all potential variables that could demonstrate performance differentials. CARE analysis software is free to download and is required to analyze and visualize the electronic data contained within CAPS datasets. The CAPS online crash datasets are free to download, and contains a variety of crash data files that mainly belongs to the state of Alabama, such as the vehicle crash files, the driver data file, the person data file, and the road data file (CAPS 2016).

2.5: Count Data

When discussing traffic crash modeling, it is important to differentiate between a count, and count data. The term count typically refers to an enumeration of events. Count data, on the other hand, refers to the observations made about events that are enumerated (Hilbe 2014). A common quality of count data is that (0.0) is the most frequently observed value, (1.0) is the next most observed, (2.0) the next, and so on. Use of count data are widespread in many disciplines, including transportation engineering. Examples of count data applications in transportation include the number of driver route changes per day, the number of trip departure changes per week, number of vehicles waiting in a queue, and the number of crashes observed on road segments per some time period, such as a year, or five years. Count data are often described as random events, sporadic (i.e. isolated or scattered), rare, discrete, not continuous, and non-negative integers (Hauer 1992). One frequent pitfall is to model count data as continuous data by applying an ordinary least square regression (Glenberg 1996). This approach is inappropriate because regression models can produce predicted values that are non-integers and can also predict values that are negative, both of which are inconsistent with count data. In addition, many distributions of count data are positively skewed with many observations in the data set having a value of 0.0. The high number of zeros in the data set prevents the transformation of a skewed distribution into a normal one, which is a requirement of normal distribution. An alternative is to use a Poisson distribution or one of its variants. Poisson distributions have a number of advantages over an ordinary normal distribution, including a skew, discrete distribution, and the restriction of predicted values to non-negative numbers (Glenberg 1996).

2.6: Common Problems with Crash Data

Crash data suffer from some problems or issues that have been identified in the literature over the years. These problems are a potential source of error in modeling crash prediction that may cause incorrect estimates and inferences. These issues are summarized below:

- Over- dispersion: over-dispersion occurs when the observed variance exceeds the theoretical variance of the crash counts, which violates the assumption of the most common count-data modeling approach. Over-dispersion in crash data can result from a variety of factors, such as the clustering of data , unaccounted temporal correlation, and model miss-specification (Cameron and Trivedi 1998). When data are over-dispersed, estimation of a crash model can lead to biased parameter estimates, which in turn could lead to incorrect inferences regarding the factors that determine crash-frequencies (Miaou 1994; Cameron and Trivedi 1998; Park and Lord 2007).

- Under-dispersion: under-dispersion occurs when the observed variance of the crash counts is smaller than the assumed (i.e. theoretical) variance, and most likely to occur with small sample sizes. Although rare, however, under-dispersion can lead to incorrect parameter estimates and crash prediction (Cameron and Trivedi 1998; Oh et al. 2006; Park and Lord 2007).
- Small Sample Size: crash data collection process may be expensive, therefore crash data are sometimes characterized by a small number of observations (i.e. small sample size), which can produce low sample-mean. Small samples can cause estimation problems in traditional count prediction models. For example, with small sample sizes, the maximum likelihood estimation of parameters could produce insufficient results (Wood 2002; Lord and Bonneson 2007). Also, Lord (2006) show that the dispersion parameter of the negative binomial model can be incorrectly

estimated when using data characterized by a small sample size and low sample mean.

- Time Interval Variations: crash data are typically collected over some time period, such as one year, three years, and five years. Over the collection period, some explanatory variables and their relationship to the crash incidents may change, a reality that is not usually considered due to the lack of detailed data within the collection period. Ignoring within-period variation in explanatory variables may result in biased estimation of parameters, and incorrect prediction of crashes as a result of unobserved heterogeneity (Glenberg 1996; Lord and Bonneson 2007).
- Temporal and Spatial Autocorrelations: the prediction of crash models can be improved when several years of crash data are utilized, such as a period of three years instead of one year (Mohammadi et al. 2014). However, this means that the same roadway entity will generate multiple observations, which will be correlated over time because many of the unobserved effects associated with a specific roadway entity will remain the same over time. This phenomenon is termed temporal autocorrelation, which can adversely affect the precision of parameter estimates. Similarly, correlation of observations over space can exist given that roadway entities may be in close proximity and may share unobserved effects. This phenomena is termed spatial autocorrelation and if not appropriately addressed, can also lead to incorrect parameter estimates (Gujarati 1992; Lord and Persaud 2000; Wood 2002; Washington et al 2010; Lord and Mannering 2010;

Savolainen et al. 2011).

- Omitted-Variables Bias: modeling crash prediction with few explanatory variables could produce simplified models with omitted-variables bias. Leaving out important explanatory variables can result in biased parameter estimates and incorrect inferences, especially if the omitted variable is correlated with variables included in the model, which is often the case (Arminger et al. 1995; Glenberg 1996; Cameron and Trivedi 1998; Caliendo et al. 2007; El-Basyouny and Sayed 2009; Geedipally et al. 2012).
- Under-Reporting: traffic crash data may suffer from under-reporting effects, especially for minor, and less severe crashes. The unknown parameters in the models are generally estimated assuming random sampling from the population, therefore, if under-reporting is not accounted for, then it could result in biased samples that are likely to produce incorrect parameters in the model-estimation process (Glenberg 1996; Elvik 2006; Caliendo et al. 2007; Anastasopoulos and Mannering 2009).
- Non-Linear Relationships Bias: many crash prediction models assume that explanatory variables influence the dependent variables in linear manner. However, it has been shown that non-linear functions can often better characterize the relationships between crash frequencies and explanatory variables. For example, using traffic flow as a measure of exposure, some have found that the crash prediction per unit of exposure becomes smaller

as traffic flow increases pointing to unobserved heterogeneity and possible other specification problems in the functional form of the model (Myers 1990; Rosenkrantz 1997; Shankar et al. 1997; Lord and Persaud 2000; Wood 2002).

- Endogeneity: endogeneity refers to the cases where explanatory variables are correlated with disturbance terms (i.e. unobserved heterogeneity), which could cause incorrect estimations and inferences. For example, a severity model that considers the presence of an airbag as an explanatory variable in a model of injury-severity outcomes. In this case, the risk of not using the airbag can be difficult to capture in the database, so it is generally captured as a part of the disturbance term, which could overestimate the airbag's effectiveness (Savolainen et al. 2011).

2.7: Review of Statistical Approaches of Crash Modeling

There are different statistical approaches for crash modeling. The next sections present some of the mostly used methods.

2.7.1: Multiple Linear Regression

Crash prediction models have been widely used for decades. Early models were based on the simple multiple linear regression models assuming normally distributed errors. The general form of the linear crash prediction model can be expressed as follows:

$$Y \mid \theta \sim Dist\,(\theta) \text{ with } \theta = f\,(X, \beta, \varepsilon) \qquad (2.1)$$

where,

Y: the dependent variable (i.e. crash frequency),

θ: the crash dataset,

Dist (θ): the model distribution,

X: a vector representing different independent variables (i.e. risk factors),

β : a vector of regression coefficients,

$f(.)$: link function that relates X and Y together,

ε: the disturbance or error terms of the model.

2.7.2: Poisson Regression

Although multiple linear regression models have been widely applied, it has been found that crash occurrence can often be better fitted with a Poisson distribution. Hence, generalized linear modeling variates of the Poisson regression model have been proposed to explore the relationship between the risk factors and crash frequency (Park and Lord 2007; Ma et al. 2008; El-Basyouny and Sayed 2009). Poisson regression has been applied to a wide range of transportation count data, including crash frequency. A Poisson regression model is similar to an ordinary linear regression, with two exceptions. First, it assumes that the errors follow a Poisson (not normal) distribution. Second, rather than modeling the response variable Y as a linear function of the regression coefficients, it models the natural log of the response variable, *ln(Y)*, as a linear function of the coefficients (Lord and Mannering 2010). The Poisson model can be expressed as follows:

$$P\ (n_{\mathrm{i}}) = \frac{\lambda i\ \ EXP(-\ \lambda i\ )}{n!} \qquad (2.2)$$

where,

P (n_{i}): the probability of n crashes occurring on a highway segment i,

n_i: the number of observations per time period (such as a year),

λ_i: the expected crash frequency on road segment i per time period (i.e. the mean of distribution) which can be estimated as follows:

$$\lambda_i = EXP\ (\beta X_i) \qquad (2.3)$$

where

X_i: a vector of the independent variables (i.e. risk factors),

β: a vector of the estimates (coefficients) of the independent variables X_i.

This model is estimable by standard maximum likelihood methods, with the log likelihood (*LL*) function given as:

$$LL\ (\beta) = \sum_1^n [-EXP\ (\beta X_i) + n\ (\beta Xi) - Ln\ (n!)] \qquad (2.4)$$

One assumption of Poisson Models is that the mean and the variance are equal, an assumption that is sometimes violated (Lord and Mannering 2010). This can be dealt with by using a dispersion parameter if the difference is small, or by using a negative binomial regression model if the difference is large (Hilbe 2007).

2.7.3: Negative Binomial Regression Model (NB)

In order to overcoming the problem of over-dispersion, the Negative Binomial (NB) distribution (also called the Poisson-Gamma) has been investigated as an alternative to the Poisson distribution given that it relaxes the condition of mean equals to variance, and hence can take into account over-dispersion in the crash data counts (Lord and Mannering 2010). As a result, NB models have been widely applied in crash frequency modeling (Kim et al. 2007; Lord and Bonneson 2007; El-Basyouny and Sayed 2009;

Daniels et al. 2010; Malyshkina and Mannering 2010; Geedipally et al. 2012).

The NB uses a Gamma probability distribution and can relax the assumption of the mean equals the variance and, hence, the NB can accommodate over-dispersion that may exist in the crash data counts (Hilbe 2014). A primary source of over-dispersion is the clustering of data, and the possible omission of relevant independent variables influencing the Poisson rate across observations (Lord 2006). In order to obtain the NB model, the Poisson regression can be rewritten by adding an error term to its expected number of crashes, and becomes (Lord and Mannering 2010):

$$\lambda i = EXP\ (\beta Xi + \varepsilon_i) \qquad (2.5)$$

where $EXP\ (\varepsilon_i)$ is a gamma-distributed error with mean equals one and variance equals α. The addition of this term allows the variance $VAR\ (n_i)$ to differ from the mean $E\ (n_i)$ as shown in Eq. 2.6:

$$VAR\ (n_i) = E\ (n_i)\ (1 + \alpha E\ (n_i)) \qquad (2.6)$$

This error term is called the over-dispersion parameter, and both α and β can be estimated from the maximum likelihood function. When α is zero, the model becomes Poisson regression, and if α is found to be significantly different from zero, then the NB regression can be used instead of the Poisson regression model to handle the over-dispersion in crash data (Lord 2006). However, the NB model also has some limitations such as its inability to handle the case of under-dispersion of the data count, when the mean of the crash counts is higher than the variance (Lord 2006; Oh et al. 2006).

2.7.4: Poisson-Lognormal Regression Model

To address the limitations of the NB models, the Poisson-lognormal model was introduced, in which the error term is Poisson-lognormal rather than gamma- distributed so as to better handle under-dispersed data counts (Aguero-Valverde and Jovanis 2008; Lord and Miranda 2008; Daniels et al. 2010). The Poisson-lognormal model is similar to the negative binomial model, however, the *EXP* (ε_i) term used in the model is lognormal-rather than gamma-distributed. The Poisson-lognormal model provides more flexibility than the negative binomial model, but it does have some limitations, such as, its complex estimation of parameters due to the fact that the Poisson-lognormal distribution does not have a closed form (Miaou and Lord 2003; Lord and Miranda 2008).

2.7.5: Zero Inflated Poisson and Negative Binomial Regression Models

Another widely used crash frequency modeling approach is the zero-inflated Poisson and zero-inflated negative binomial models, which have been introduced primarily to deal with the over-dispersion problem caused by excessive zeroes (i.e. locations where no crashes can be observed) in traffic data counts. The zero-altered procedure allows modeling the crash frequencies in two states, namely; the zero-crash state, and the non-zero crash state (where crash frequencies follow Poisson or negative binomial distribution), and the probability of a section being in zero or non-zero states can be found by a binary logit or probit model. In crash data, large numbers of zero observations are commonly present largely due to under reporting of minor crashes at these sites, the presence of dangerous crash sites (i.e. non-zero crash sites) in close proximity to the neighboring zero crash sites rendering the zero-crash sites to the safe mode, and given that some of zero crash sites may be free from only certain type of crashes, not all types of crashes (Shankar et al. 1997). Zero-inflated models attempt to account for such excess zeros. A dual state crash system may be assumed, in which one state is the zero crash state that can be regarded as virtually safe during the observation period, while the other state is the non-zero crash state. For example, consider vehicle crash occurring per year on 1-kilometer sections of highway. For straight sections of roadway with wide lanes, low traffic volumes, and no roadside objects, the likelihood of a vehicle crash occurring may be extremely small, but still present because an extreme human error could randomly cause an accident. These sections are considered to be in a zero-crash state that refer to situations where the likelihood of an event occurring is extremely rare in comparison to the non-zero state where crash occurrence is inevitable and follows some count distribution (Lambert 1992). To address the zero-inflated modeling processes, the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) regression models have been developed. The probabilities of the two possible zero- and non-zero states are: p_i for the zero crash state, and (1-p_i) for the non-zero crash state, and the overall probability of crashes is the sum of the probabilities from each state. The probability of crash frequency in the zero state can be modeled as:

$$Pr\ (n_i = 0) = p_i + (1 - p_i)\ R_i(0) \qquad (2.7)$$

where R_i (0) is the probability of zero crashes that occurs in the zero state. The probability of crash frequency in the non-zero state can be modeled as:

$$Pr\ (n_i > 0) = (1 - p_i)\ R_i\ (n_i) \qquad (2.8)$$

where R_i (n_i) is the probability of non-zero crashes in the non-zero state. Maximum likelihood estimates can be used to estimate the parameters of both ZIP and

ZINB regression models and confidence intervals are constructed by likelihood ratio tests. In zero-inflated models, the two state process is assumed to follow a logit (logistic) or probit (normal) probability process (Shankar et al. 1997). Zero-inflated models have shown great flexibility in both states, although their applicability to crash prediction has been criticized because of the long term mean equals to zero in the safe state, and hence, biased estimates may result (Lord and Mannering 2010; Malyshkina and Mannering 2010).

2.7.6: Conway-Maxwell Poisson Regression Models

The Conway–Maxwell Poisson model has been recently investigated with respect to highway safety issues, but it's application in crash frequency modeling has been rather limited (Kadane et al. 2006; Lord and Mannering 2010). Generalized additive models have been explored given that they can provide smoothing functions for the explanatory variables. The Conway–Maxwell Poisson distribution is a generalization of the Poisson distribution that can handle both under-dispersed and over-dispersed crash data. The main advantage of this model is to handle the under-dispersion in crash data that cannot be modeled by the Poisson model or the Negative Binomial model. However, the low sample-mean, and small sample size of the under-dispersed crash data can influence the estimated parameters, and therefore, it has been limited in the application of crash frequency (Kadane et al. 2006; Lord 2006).   However, in practice, the estimation of these models can become very difficult as they require more parameters, a problem that has likely impeded their application to crash frequency prediction (Zhang 2008; Li et al. 2009).

2.7.7: Random-Parameter Models

Random-parameters models have also been investigated to take the effect of the unobserved heterogeneity from one roadway site to another (Shankar and Mannering 2008; Anastasopoulos and Mannering 2009; Washington et al. 2010).

The motivation for random-parameter models is to account for unobserved heterogeneity across observations. Random-parameter models can be derived by assuming that the estimated parameters vary across observations according to some distribution. Estimated parameters can be modeled as (Greene 2008):

$$\beta_n = \beta + \omega_n \qquad (2.9)$$

where

β_n : a vector of estimated parameters of the n observations,

ω_n : a randomly distributed term.

With this equation, the Poisson, and the Negative Binomial parameters become:

$$\lambda i \mid \omega_n = EXP\ (\beta_n X_n) \qquad (2.10)$$

$$\lambda i \mid \omega_n = EXP\ (\beta n\ Xn + \varepsilon_n) \qquad (2.11)$$

2.7.8: Artificial Neural Networks and Fuzzy Logic models

Given that a linear function may not sufficiently explain the relationship between the dependent variables and the associated independent variables in crash modeling, non-linear approximators such as fuzzy logic and neural networks have also been explored. Artificial Neural Networks (ANNs) are a class of computational intelligence tools that can be used for prediction and classification problems. ANNs can model very complex non-linear functions to high accuracy levels using a process of learning that is similar to the learning procedure of the cognitive system in the human brain. The network body is composed of input layers, hidden layers, and output layers. These models can be trained to approximate any nonlinear function to a required degree of accuracy using a learning algorithm (such as back propagation) that would give the desired output, in a supervised learning process. ANNs have some advantages over the statistical models. For instance, regression models need a pre-defined relationship or functional form between the dependent variable (crash frequency) and the independent explanatory variables that can be estimated by some statistical approaches, whereas the ANNs do not require the establishment of these functional forms, and can be easily applied in the analysis. On the other hand the ANNs differ from the statistical models in that they behave as black-boxes and do not provide interpretation for the parameter estimates (Chang 2005; Riviere et al. 2006; Xie and Zhang 2008). Fuzzy logic applications have increasingly been proven to have a significant crash-predicting capability in recent years (Wang et al. 2011). Fuzzy logic system is defined as the nonlinear mapping of an input data set to a scalar output data, and the first step of the process (known as fuzzification) consists of gathering a crisp set of input data that will be converted to a fuzzy set using fuzzy linguistic variables, fuzzy linguistic terms, and membership functions. After that, an inference is made based on a set of fuzzy rules, and then, the resulting fuzzy output is mapped to a crisp output using the membership functions, in the defuzzification step (Meng et al. 2009).

2.7.9: Logit and Probit Models

Logit and Probit models can be applied to study crash severity. Binary models consider two outcomes, and multinomial models consider three or more outcomes. In binomial logit or probit models, the dependent variable, *Y*, can take one of two values 0.0 or 1.0. For example, injury or non-injury, fatal or non-fatal. The binomial logit model denotes π_i = *Pr* (Y_i = 1), allowing the logistic transformation of π in the link function to produce the binomial logit function:

$$Logit\ (\pi_i) = log\ [\frac{\pi_i}{1-\pi_i}] = X_i\ \beta \qquad (2.12)$$

where,

X_i : a vector of explanatory variables (i.e. risk factors),

β : a vector of regression coefficients.

As π approaches zero, *logit* (π) tends toward - ∞; and as π approaches 1.0, *logit* (π) tends toward + ∞ (Mannering and Grosdsky 1995). The binomial probit model is an alternative to the binomial logit model, in which the *probit* (π_i) is the standard cumulative normal distribution function (Θ^{-1}) that can be expressed as:

$$Probit\ (\pi_i) = \Theta^{-1}\ (\pi_i\ ) = Xi\ \beta \qquad (2.13)$$

In practice, the logit model is preferred due to the simplicity of its probability distribution, and density functions (Washington et al. 2010; Greene 2012).

2.8: Sight Distance Analysis

Roadway sight distance is a measure of roadway visibility, which is an important factor in the assessment of road safety. Greater visibility can provide motorists more time to avoid crashes and conflicts, facilitating safe and efficient operation. However, poor visibility can reduce the driver’s ability to react to changing conditions and is a significant factor in roadway crashes and near collisions. A driver’s ability to view ambient roadway conditions is necessary for safe operation of a vehicle. The roadway must have sufficient sight distance that drivers have the time to react to and avoid striking unexpected objects in their path. In addition, certain two-lane, two-way highways should also have adequate passing sight distance to enable drivers to use the opposing traffic lane for passing other vehicles without interfering with oncoming vehicles. Three different types of sight distance are often discussed in the literature: (1) the sight distances needed for stopping, applicable to all highway travels; (2) the sight distances needed for decisions at hazardous complex locations; and (3) the passing sight distance needed on two lane highways.

2.8.1: Stopping Sight Distance (SSD)

Stopping sight distance (SSD) reflects a distance within which a driver can effectively see an object in the roadway and stop their vehicle before colliding with the object (AASHTO 2011). The available sight distance on a roadway should be long enough to enable a vehicle traveling at or near the design speed to stop before reaching a stationary object in its path. Although greater lengths of visible roadway are desirable, the sight distance at every point along a roadway should be at least that needed for a below-average driver or vehicle to stop. Recommended protocols for calculating stopping sight distances account for the basic principles of physics and the relationships between various design’s parameters. Stopping sight distance can be determined as the sum of two distances (AASHTO 2011), namely: 1) Reaction distance (the distance a vehicle travels from the moment a driver sees the object until the driver applies the brakes) and; 2) Braking distance (the distance a vehicle travels from the moment the brakes are applied until the vehicle comes to a complete stop). The following equation shows how SSD is typically computed by combining these two distances (AASHTO 2011):

$$SSD = 0.278VT + 0.039\,V^2/a \qquad (2.14)$$

where:

SSD = stopping sight distance, m;

V = highway design speed, km/h;

T = brake reaction time, seconds;

a = deceleration rate, m/s^2.

AASHTO (2011) recommends a (2.5 seconds) as the driver’s reaction time, and (3.4 m/s^2) as the deceleration rate for stopping sight distance calculations. Figure 2.1 provides an illustration of the factors contributing to the AASHTO recommendations on SSD. Table 2.1 shows the SSD on level terrains. The recommended height of the driver’s eye above the road surface is (1.08 m) and the height of an object above the roadway is (0.6 m).

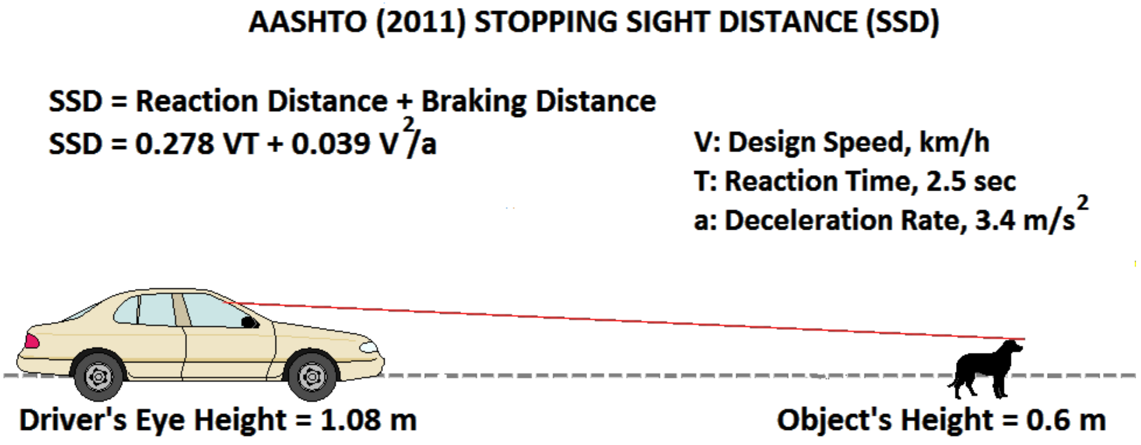
AASHTO (2011) STOPPING SIGHT DISTANCE (SSD)
SSD = Reaction Distance + Braking Distance
SSD = 0.278 VT + 0.039 V²/a
V: Design Speed, km/h
T: Reaction Time, 2.5 sec
a: Deceleration Rate, 3.4 m/s²
Driver's Eye Height = 1.08 m
Object's Height = 0.6 m

Figure 2.1: AASHTO (2011) criteria for stopping sight distance

Table 2.1: Stopping sight distance on level roadways

Design Speed (km/h)	Reaction Distance (m)	Braking Distance (m)	Calculated SSD (m)	Design SSD (m)
20	13.9	4.6	18.5	20
30	20.9	10.3	31.2	35
40	27.8	18.4	46.2	50
50	34.8	28.7	63.5	65
60	41.7	41.3	83.0	85
70	48.7	56.2	104.9	105
80	55.6	73.4	129.0	130
90	62.6	92.9	155.5	160
100	69.5	114.7	184.2	185
110	76.5	138.8	215.3	220
120	83.4	165.2	248.6	250
130	90.4	193.8	284.2	285

2.8.1.1: Driver's Eye Height for SSD

The driver eye height of 1.08 m that is commonly recommended is based on research that suggests average vehicle heights have decreased to 1.30 m (4.25 ft) with a comparable decrease in average eye heights to 1.08 m (3.50 ft). For large trucks, the driver eye height ranges from 1.80 m to 2.40 m (3.50 ft to 7.90 ft). The recommended height for a truck driver for design is 2.33 m (7.60 ft) above the road surface.

2.8.1.2: Object's Height for SSD

An object height of a 0.6 m (2.0 ft) is commonly selected based on studies that have indicated that objects less than 0.60 m in height are less likely to cause crashes. Therefore, an object height of 0.6 m is considered the smallest object that could pose risk to drivers. In addition, an object height of 0.60 m is a good representative of the height of automobile headlights and taillights (AASHTO 2011).

2.8.1.3: Effect of Grades on SSD

For roads having positive grades, braking distance can be calculated by the following equation (AASHTO 2011):

$$d_b = \frac{V^2}{254[\left(\frac{a}{9.81}\right) \pm G]} \qquad (2.15)$$

where,

d_b: Braking distance on grade, m;

V: Design Speed, km/h;

a: Deceleration rate, m/s^2;

G: Grade, rise/run, m/m.

The stopping distances needed on upgrades are shorter than on level roadways; those on downgrades are longer. The AASHTO stopping sight distances for various downgrades and upgrades are shown in Table 2.2. Passenger cars can use grades as steep as 4.0 to 5.0 percent without significant loss in speed below that normally maintained on level roadways. Operation of passenger cars on a 3.0 percent upgrade has only a slight effect on their speeds compared to operations on level terrain. On steeper upgrades, speeds decrease gradually with increases in the grade. On downgrades, passenger car speeds generally are slightly higher than on level terrains. Trucks generally increase speed by up to 5.0 percent on downgrades and decrease speed by 7.0 percent or more on upgrades as compared to their operation on level terrains (AASHTO 2011).

Table 2.2: AASHTO (2011) stopping sight distance on grades

Design Speed (km/h)	Stopping Sight Distance (m)					
	Downgrades			**Upgrades**		
	3%	**6%**	**9%**	**3%**	**6%**	**9%**
20	20	20	20	19	18	18
30	32	35	35	31	30	29
40	50	50	53	45	44	43
50	66	70	74	61	59	58
60	87	92	97	80	77	75
70	110	116	124	100	97	93
80	136	144	154	123	118	114
90	164	174	187	148	141	136
100	194	207	223	174	167	160
110	227	243	262	203	194	186
120	263	281	304	234	223	214
130	302	323	350	267	254	243

2.8.1.4: SSD for Trucks

Trucks are heavier than passenger cars; therefore, they need a longer distance to stop. However, it is believed that adjustment factors for trucks are not necessary since visibility from a truck is typically better given that the driver is seated at a higher elevation above the roadway surface. Thus, this increase in the height of the driver substitutes the need for additional stopping sight distance for trucks (AASHTO 2011).

2.8.1.5: Measuring and Recording Sight Distance

In the field, stopping sight distance is measured along the travel path of vehicles and several methods are typically utilized. The first conventional procedure is called the walking method (Brown and Hummer 2000; Rose et al. 2004) that involves at least two individuals, sighting and a target rods, a measuring wheel, and a chain. The target rod is usually 1.3 m tall representing the vehicle's height, and is usually painted orange on both the top portion and bottom 0.6 m of the rod. The bottom 0.6 m portion of the target rod is the height of object for measuring stopping sight distance. The sighting rod is 1.08 m tall representing the driver's eye height recommended by AASHTO, and is usually painted black. From any point location along the road, the observer should sight from the top of the sighting rod while the assistant moves away in the direction of travel. The assistant stops when the bottom 0.6 m portion of the target rod is no longer visible. The distance from the disappearing point to the observer presents the available stopping sight distance. The analysis procedure consists of comparing the recommended sight distance from AASHTO tables to the measured sight distance in the field. Given that this measurement method requires the observer to be in the travel lane with their back to traffic, measurements along the shoulder are often substituted since they are safer for the personnel conducting the measurement. Similar in scope to the conventional approach, modern technologies have also been utilized to measure sight distance in the field. For instance, the two-vehicle method (Brown and Hummer 2000) employs two vehicles equipped with sensors that measure their spacing, two-way communication device, and a paint sprayer. The vehicles calibrate their spacing to a desired sight distance. As the vehicles traverse a roadway, observers in the trailing vehicle note whether or not portions of the road meet the specified sight distance. Another similar method is the one-vehicle method that also has been used by some transportation agencies (Brown and Hummer 2000; Rose et al. 2004). This method requires one employee in a vehicle equipped with a measuring device, and a paint sprayer. The driver moves slowly through the road, and watches the points at which the view opens up, and marks these points by paint. Another technique that has widely been used is the computer based method, using the global positioning systems (GPS) data (Polus et al. 2000). This method requires two vehicles, the lead vehicle equipped with modern telemetry, and the trailing vehicle equipped with logging laptop computer. The visibility of a target on the lead vehicle, monitored from the trailing vehicle, is recorded to determine if the available sight distance is sufficient. The field-based measurement approaches discussed are advantageous in that a diverse range of roadway conditions can be incorporated. That is, since there are observers on the ground, obstructions to visibility can be accounted for in a more precise manner. However, field measurement techniques are extremely time consuming and may require many years to conduct at a broad regional level. Field measurements can also lack consistency based on the measurement technique and the characteristics of the crew conducting the task. Moreover, field measurements require that individuals work in traffic which presents a significant threat to their safety. As such, a measurement approach that entails a more remote analysis of sight distance and permits a broader, regional perspective would certainly be a valuable tool for providing an initial estimate of sight distance. To address this need, a variety of approaches have been developed to use other data sources to estimate sight distance without using equipped vehicles or deploying individuals to the field. In this sense, Tsai et al. (2010) propose an algorithm to compute roadway geometric data, including roadway length, sight distance, and lane width from images, using emerging vision technology based on 2D, and 3D image reconstruction. Also, Shaker et al. (2011) use stereo high resolution satellite imagery for extracting the highway profiles and constructing 3D highway visualization model using a polynomial-based generic push broom model and rational function model to perform the sensor orientation. Methods that use Global Positioning Systems (GPS) data to estimate sight distance have also been developed. For instance, Ben-Arieh et al. (2004) used a GPS data and B-Spline method to model highway geometric characteristics that utilized B-spline curves and a piecewise polynomial function. Nehate and Rys (2006) used the geometric model developed by Ben-Arieh et al. (2004) to calculate the available sight distance on 3D combined horizontal and vertical alignment. They utilized a piecewise parametric equation in the form of cubic B-splines to represent the highway surface and sight obstructions, and the available sight distance was found analytically by examining the intersection between the sight line and the elements representing the highway surface and sight obstructions. Azimi and Hawkins (2013) proposed a method that uses vector product to derive the visibility of the centerline of the roadway from the spatial coordinates of a set of GPS data of the centerline, and defined the clear zone boundaries on both sides of the roadway to determine the available sight distance at each point of the roadway.

2.8.1.6: Sight Distance Obstructions

On a crest vertical curve, the road surface at some point could limit the driver's stopping sight distance. On horizontal curves, the obstruction that limits the driver's sight distance may be some physical feature outside of the traveled way, such as a longitudinal barrier, a bridge-approach fill slope, a tree, foliage, or the back slope of a cut section. Thus, it is recommended to check all road construction plans for other obstructions to sight distance (AASHTO 2011).

2.8.1.7: SSD on Horizontal Alignments

When a vehicle travels in a circular path, it undergoes a centripetal acceleration that acts toward the center of curvature. This acceleration is sustained by a component of the vehicle's weight related to the roadway super elevation, by the side friction developed between the vehicle's tires and the pavement surface, or by a combination of the two, which is occasionally equals to the centrifugal force (AASHTO 2011). The design of roadway curves should be based on an appropriate relationship between design speed and radius of curvature and on their joint relationships with super elevation (roadway banking) and side friction. When a vehicle travels at constant speed on a curve super elevated so that the friction is zero, the centripetal acceleration is sustained by a component of the vehicle's weight, and no steering force is needed. A vehicle traveling faster or slower than the balance speed develops tire friction as steering effort is applied to prevent movement to the outside or to the inside of the curve. From the basic laws of mechanics, the fundamental equation that governs vehicle operation on a horizontal curve is as follows (AASHTO 2011):

$$\frac{0.01e+f}{1-0.01\,ef} = \frac{v^2}{gR} = \frac{0.0079V^2}{R} = \frac{V^2}{127R} \qquad (2.16)$$

where,

e: rate of roadway super elevation, percent;

f: coefficient of side friction, unitless;

v: vehicle speed, m/s;

V: vehicle speed, km/h;

g: gravitational constant, 9.81 m/s^2;

R: radius of the curve measured to the vehicle’s center of gravity, m.

Values for maximum super elevation rate (e) and maximum side friction coefficient (f) can be determined from the AASHTO Green Book for curve design. Using these values in the curve formula results in determining a minimum curve radius for various design speeds (AASHTO 2011). The coefficient of friction f is the friction force divided by the component of the weight perpendicular to the pavement surface. The value of the product (e f) is always small. As a result, the ($1 – 0.01ef$) term is nearly equal to 1.0 and is normally omitted in highway design. Omission of this term yields the following basic side friction equation, which is widely used in curve design (AASHTO 2011):

$$f = \frac{V^2}{127\,R} - 0.01e \qquad (2.17)$$

The minimum radius is a limiting value of curvature for a given design speed and is determined from the maximum rate of super elevation and the maximum side friction coefficient. Use of sharper curvature for that design speed would call for super elevation beyond the limit considered practical or for operation with tire friction beyond what is considered comfortable by many drivers, or both. The minimum radius of curvature is based on a threshold of driver comfort that is suitable to provide a margin of safety against skidding and vehicle rollover. The minimum radius of curvature, R_{min} can be determined directly from the following equation (AASHTO 2011):

$$R_{min} = \frac{V^2}{127\ (\ 0.01\ e_{max} + f_{max})} \qquad (2.18)$$

If there are sight obstructions (such as walls, cut slopes, buildings, and barriers) on the inside of horizontal curves and their removal to increase sight distance is impractical, a design may need adjustment in the highway alignment. For general use in design of a horizontal curve, the horizontal sight line is a chord of the curve, and the stopping sight distance is measured along the centerline of the inside lane around the curve, as shown in Figure 2.2. The horizontal sight line offset (HSO) can be determined from Equation 2.19. The equation applies only to circular curves longer than the sight distance for the specified design speed (AASHTO 2011).

$$HSO = R[\ 1 - cos(\frac{28.65\ S}{R})] \qquad (2.19)$$

where,

HSO: Horizontal Sightline Offset, m;

S: Stopping Sight Distance, m;

R: Radius of curve, m.

Where adequate stopping sight distance is not available because of a sight obstruction, alternative designs must be used, such as increasing the offset to the obstruction, increasing the radius, or reducing the design speed (AASHTO 2011).

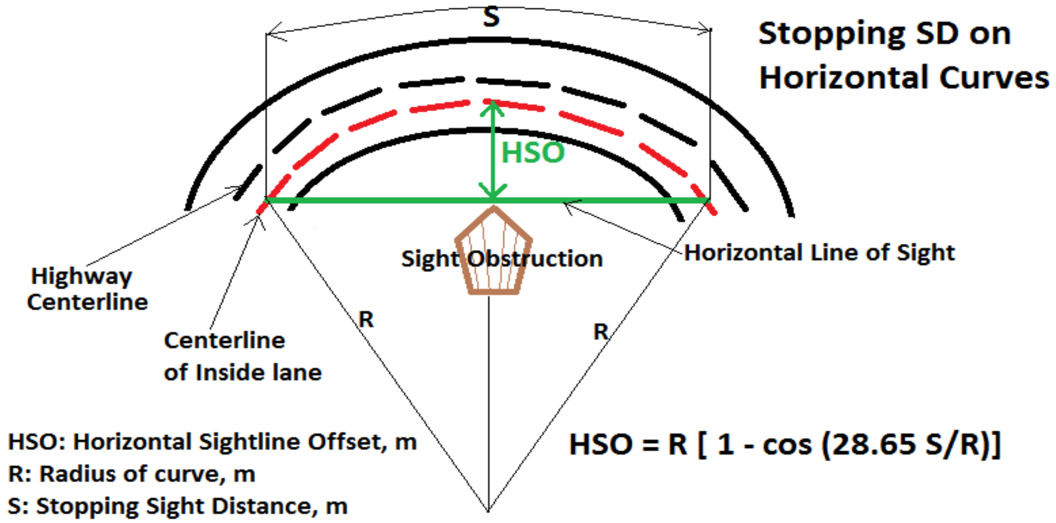
S
Stopping SD on Horizontal Curves
HSO
Sight Obstruction
Horizontal Line of Sight
Highway Centerline
Centerline of Inside lane
R
R
HSO: Horizontal Sightline Offset, m
R: Radius of curve, m
S: Stopping Sight Distance, m
HSO = R [1 - cos (28.65 S/R)]

Figure 2.2: AASHTO (2011) SSD criteria on Horizontal alignments

2.8.1.8: SSD on Crest Vertical Curves

Crest vertical curves should be designed to provide at least the stopping sight distance that is a major design control. Minimum lengths of crest vertical curves based on sight distance criteria generally are satisfactory from the standpoint of safety, comfort, and appearance (AASHTO 2011). The basic equations for length of a crest vertical curve in terms of algebraic difference in grade and sight distance criteria are as follows (AASHTO 2011):

when *S* is less than *L*:

$$L = \frac{AS^2}{100\,(\sqrt{2h_1} + \sqrt{2h_2})^2} \qquad (2.20)$$

when *S* is greater than *L*:

$$L = 2S - \frac{200(\sqrt{h_1} + \sqrt{h_2})^2}{A} \qquad (2.21)$$

where,

L: Length of vertical curve, m;

A: Algebraic difference in grade, percent;

S: Sight distance, m;

h_1: Driver's Eye Height above roadway surface, m;

h_2: Object's Height above roadway surface, m.

When the height of the eye and the height of object are 1.08 and 0.60 m (3.50 ft and 2.0 ft), respectively, as used for stopping sight distance, the equations become:

when S is less than L:

$$L = \frac{AS^2}{658} \qquad (2.22)$$

when S is greater than L:

$$L = 2S - \frac{658}{A} \qquad (2.23)$$

Rate of vertical curvature, *K*, is usually used in the design calculation, which is the length of curve per percent algebraic difference in intersecting grades, (i.e. $K = L/A$). Figure 2.3 shows the AASHTO parameters used in determining the length of a crest vertical curve to provide stopping sight distance. For night driving on highways without lighting, the headlights of the vehicle directly illuminate the length of visible roadway.

Thus, stopping sight distance values exceed road-surface visibility distances afforded by the low-beam headlights regardless of whether the roadway profile is level or curving vertically. Since the headlight, mounting height (typically about 0.60 m) is lower than the driver eye height used for design (1.08 m), the sight distance to an illuminated object is controlled by the height of the vehicle headlights rather than by the direct line of sight. In addition, drivers are aware that visibility at night is less than during the day, regardless of road features, and they may therefore be more attentive and alert (AASHTO 2011).

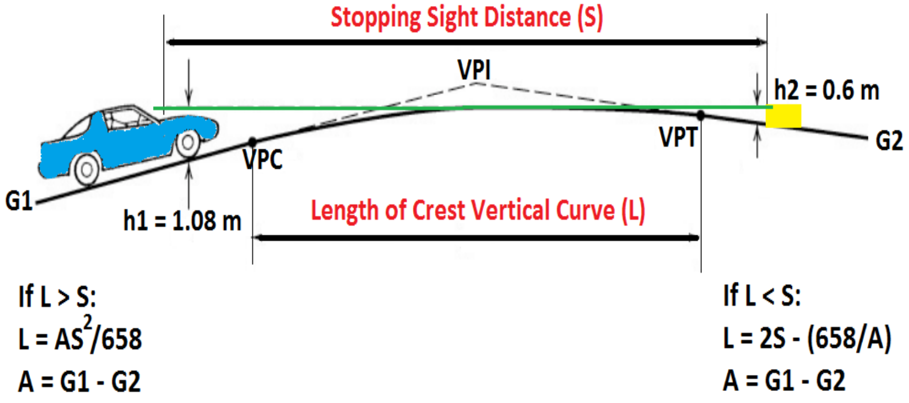

Figure 2.3: SSD parameters used in design of crest vertical curves

2.8.1.9: SSD on Sag Vertical Curves

Design controls for sag vertical curves differ from those for crests, and separate design values are needed. The headlight sight distance is used to determine the length of a sag vertical curve, and the values determined for stopping sight distances are within these limits. As in the case of crest vertical curves, it is convenient to express the design control in terms of the *K* rate for all values of *A*. When a vehicle traverses a sag vertical curve at night, the portion of highway lighted ahead is dependent on the position of the headlights and the direction of the light beam. A headlight height of 0.60 m (2.0 ft) and a 1-degree upward divergence of the light beam from the longitudinal axis of the vehicle are assumed in the design (AASHTO 2011). The following equations are used to determine the length of sag vertical curves based on sight distance criteria (AASHTO 2011):

when *S* is less than *L*:

$$L = \frac{AS^2}{200\ [\ 0.6+S\ (tan\ 1)]} = \frac{AS^2}{120+3.5\ S} \tag{2.24}$$

when *S* is greater than *L*:

$$L = 2S - \frac{200\ [\ 0.6+S\ (tan\ 1)]}{A} = 2S - \frac{120+3.5\ S}{A} \tag{2.25}$$

where,

L: Length of sag vertical curve, m;

A: Algebraic difference in grades, percent;

S: Stopping sight distance (Light beam distance), m.

The light beam distance is approximately the same as the stopping sight distance, and it is appropriate to use stopping sight distances for different design speeds as the value of *S* in the above equations (AASHTO 2011). Figure 2.4 shows the parameters used in the design of a sag vertical curve.

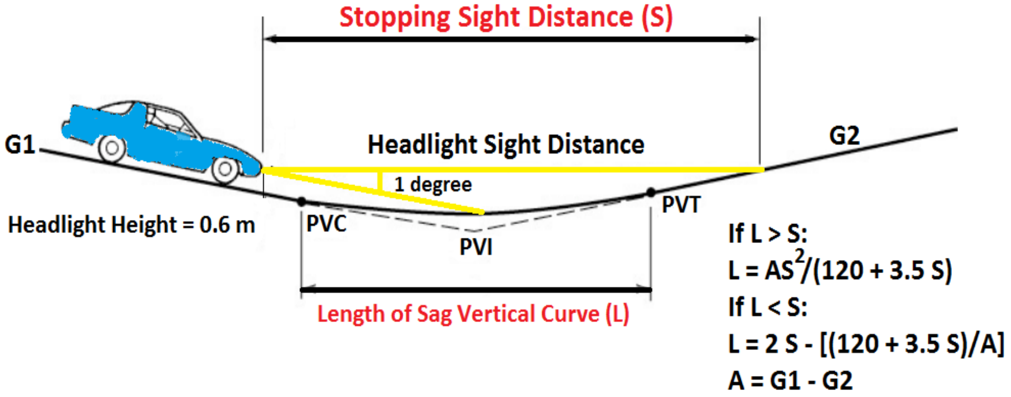

Figure 2.4: SSD parameters used in design of sag vertical curves

2.8.1.10: SSD at Under Crossings

Sag vertical curves under passing a structure should be designed to provide the minimum recommended stopping sight distance for sag curves (AASHTO 2011). The general equations for sag vertical curve length at under crossings are (AASHTO 2011):

when *S* is less than *L*:

$$L = \frac{AS^2}{800\ [\ C - \frac{h_1 - h_2}{2}\ ]} \qquad (2.26)$$

when *S* is greater than *L*:

$$L = 2S - \frac{800\ [\ C - \frac{h_1 - h_2}{2}]}{A} \qquad (2.27)$$

where,

L: Length of Sag Vertical Curve, m;

S: Stopping Sight Distance, m;

C: Vertical Clearance, m;

h_1: height of eye, m;

h_2: height of object, m;

A: Algebraic difference in grades, percent.

AASHTO uses an eye height of 2.4 m (8.0 ft) for a truck driver and an object height of 0.6 m (2.0 ft) for the taillights of a vehicle. Substituting these values, the above equations become (AASHTO 2011):

when S is less than L:

$$L = \frac{AS^2}{800\ (C-1.5)} \qquad (2.28)$$

when S is greater than L:

$$L = 2S - \frac{800\ (C-1.5)}{A} \qquad (2.29)$$

Figure 2.5 shows the AAHSTO parameters used in the design of sag vertical curves under passing a structure.

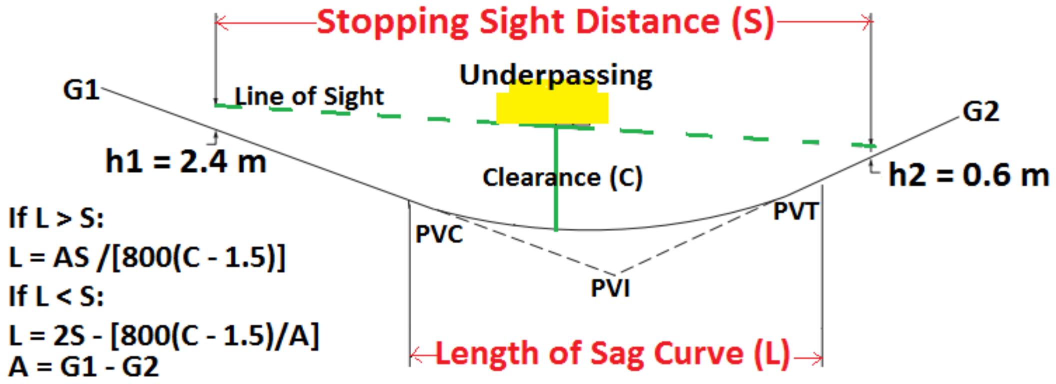

Figure 2.5: SSD parameters used in design of under passing sag curves

2.8.2: Decision Sight Distance (DSD)

While stopping sight distances are usually sufficient to allow average drivers to come to a complete stop under ordinary circumstances, however, greater distances are preferred where drivers must make instantaneous decisions, where information is difficult to perceive, or when unexpected or unusual maneuvers are needed. In these circumstances, decision sight distance provides the greater visibility distance that drivers need. Decision sight distance is defined as the distance required for a driver to detect an unexpected source or hazard in a roadway, recognize the threat potential, select an appropriate speed and path, and complete the required maneuver safely and efficiently (AASHTO 2011). Most traffic situations presented on highways require stopping sight distance at a minimum; however, decision sight distance is also recommended for safer and smoother operations. For example, long traffic queues, problems of driver expectancy, and high traffic volumes require more time and distances to accommodate normal vehicle maneuvers of lane changing, speed changes and path changes.

2.8.2.1: Comparison between SSD and DSD

The distinction between stopping sight distance and decision sight distance must be well understood (AASHTO 2011). Stopping sight distance is applied where only one obstacle must be seen in the roadway and dealt with. Decision sight distance applies when traffic conditions are complex, and driver expectancies are different from normal traffic situation. The difference between stopping in the context of decision sight distance and stopping sight distance is that the vehicle should stop for some complex traffic condition, such as a queue of vehicles or hazardous conditions, rather than an object in the roadway. The values of decision sight distance are greater than the values of stopping sight distance because they provide the driver an additional margin for error and afford sufficient length to maneuver at the same or reduced speed rather than to stop. The added complexity in DSD requires additional perception-reaction time prior to applying the brakes to begin to slow the vehicle to a stop or change the speed or travel path. This allows the driver additional time to detect and recognize the roadway or traffic situation, identify alternative maneuvers, and initiate a response on the highway. AASHTO (2011) suggest that about 3.0 to 9.0 seconds are required for detecting and understanding the unexpected traffic situation with an additional 5.0 to 5.5 seconds required to perform the appropriate maneuver compared to only 2.5 seconds as perception reaction time in stopping sight distance calculations. Similar to the stopping sight distance, AASHTO (2011) recommends assuming the driver's eye height at 1.08 m (3.5 ft), and the object height as 0.60 m (2.0 ft) for decision sight distance calculations.

2.8.2.2: Cases of DSD

Decision sight distance is different for urban versus rural conditions and for stopping versus maneuvering within the traffic stream conditions. Consequently, there are five different cases for decision sight distance (AASHTO 2011) as follows:

- Avoidance Maneuver A: Stop on Rural Road – (t = 3.0 sec),
- Avoidance Maneuver B: Stop on Urban Road – (t = 9.1 sec),
- Avoidance Maneuver C: Speed/Path/Direction Change on Rural Road – (t between 10.2 and 11.2 sec),
- Avoidance Maneuver D: Speed/Path/Direction Change on Suburban Road – (t between 12.1 and 12.9 sec),

- Avoidance Maneuver E: Speed/Path/Direction Change on Urban Road – (t between 14.0 and 14.5 sec).

2.8.2.3: DSD Calculations for Stop Maneuvers A and B

The available decision sight distance for the stop avoidance maneuvers A and B are determined as the sum of two distances (AASHTO 2011), namely: 1) Reaction distance (the distance a vehicle travels from the moment a driver detects a condition or hazard in the roadway until the driver applies the brakes) and; 2) Braking distance (the distance a vehicle travels from the moment the brakes are applied until the vehicle comes to a complete stop). DSD can be computed as a function of these two distances (AASHTO 2011):

$$DSD = 0.278VT + 0.039\, V^2/a \qquad (2.30)$$

where:

DSD = decision sight distance, m;

V = design speed, km/h;

T = Maneuver time, seconds;

a = deceleration rate, m/s^2

AASHTO recommends a (3.0 seconds) as a driver’s reaction time for rural highways, (6.0 seconds) for sub urban highways, and a (9.1 seconds) for urban highways. AASHTO uses (3.4 m/s^2) as the deceleration rate for decision sight distance calculations.

2.8.2.4: DSD Calculations for Maneuvers C D and E

The available decision sight distances for avoidance maneuvers C, D, and E are determined as follows (AASHTO 2011):

$$DSD = 0.278VT \qquad (2.31)$$

where:

DSD = decision sight distance, m;

V = design speed, km/h;

T = Maneuver time, seconds.

AASHTO recommends a (10.2 to 11.2 seconds for maneuver C on rural roads, a 2.1 to 12.9 seconds for maneuver D on suburban roads, and a 14.0 to 14.5 seconds for maneuver E on urban roads) as the driver’s reaction time. Figure 2.6 provides an illustration of the recommended AASHTO criteria on DSD. The recommended height of the driver’s eye above the road surface is (1.08 m) and the height of an object above the roadway is (0.6 m). Table 2.5 shows the AASHTO recommended decision sight distances for various maneuvers. As can be seen in the table, shorter distances are generally needed for rural roads and for locations where a stop is the appropriate maneuver. If it is not practical to provide decision sight distance on some highways, attention should be given to the use of suitable traffic control devices for providing advance warning of the conditions that are likely to be encountered (AASHTO 2011).

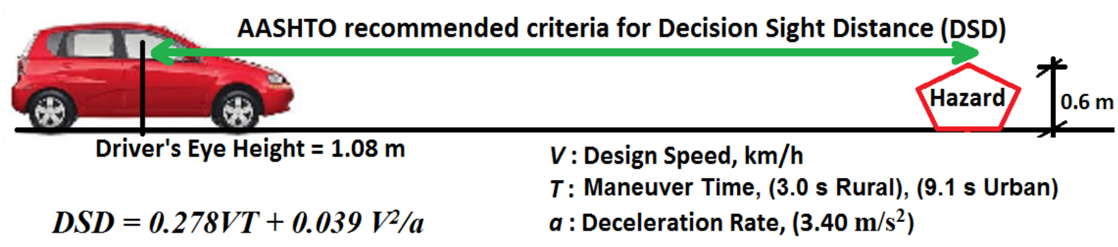
AASHTO recommended criteria for Decision Sight Distance (DSD)
Hazard
0.6 m
Driver's Eye Height = 1.08 m
V : Design Speed, km/h
T : Maneuver Time, (3.0 s Rural), (9.1 s Urban)
a : Deceleration Rate, (3.40 m/s^2)
$DSD = 0.278VT + 0.039\ V^2/a$

Figure 2.6: Recommended AASHTO criteria on DSD

Table 2.5: AASHTO recommended decision sight distance

Design Speed (km/h)	Decision Sight Distance, meters				
	Avoidance Maneuver				
	A	**B**	**C**	**D**	**E**
50	70	155	145	170	195
60	95	195	170	205	235
70	115	235	200	235	275
80	140	280	230	270	315
90	170	325	270	315	360
100	200	370	315	355	400
110	235	420	330	380	430
120	265	470	360	415	470
130	305	525	390	450	510

2.8.3: Passing Sight Distance (PSD)

Passing sight distance (PSD) is the distance that drivers must be able to see along the road ahead to safely and efficiently initiate and complete passing maneuvers of slower vehicles on two-lane, two-way highways using the lane normally reserved for opposing traffic (AASHTO 2011). PSD is a consideration along two-lane roads on which drivers may need to assess whether to initiate, continue, and complete or abort passing maneuvers. In the US, many roads are two-lane, two-way highways on which faster vehicles frequently overtake slower moving vehicles. In order to secure a safe passing maneuver, the passing driver should be able to see a sufficient distance ahead, clear of traffic, to complete the passing maneuver without cutting off the passed vehicle before meeting an opposing vehicle (AASHTO 2011). Therefore, passing sight distance (PSD) is considered an important factor in both the design of two-lane, two-way (TLTW) highways and the marking of passing zones (PZ) and no-passing zones (NPZ) on two-lane, two-way highways. The efficiency of traffic operation of many TLTW highways depends on how often faster drivers are able to pass slower drivers. For example, where faster drivers encounter a slower driver but are unable to pass, vehicle platoons are built up, and cause a decrease in the level of service and inversely affect safety, fuel consumption and emissions. The capacity of a two-lane, two-way road is increased if a large percentage of the roadway's length can be used for passing maneuvers (Haneen and Tomer 2010).

2.8.3.1: PSD on Multilane Highways

There is no need to consider passing sight distance on multilane highways that have two or more traffic lanes in each direction of travel, because passing maneuvers are expected to occur within the limits of the traveled way for each direction of travel. However, multilane roadways should have continuously adequate stopping sight distance, with greater-than-design sight distances preferred (AASHTO 2011).

2.8.3.2: Marking of Passing Zones on Two-Lane Highways

The design of two-lane highway is based on the AASHTO Green book criteria, however, the marking of passing zones (PZs) and No-passing zones (NPZs) is based on the Manual on Uniform Traffic Control Devices for Streets and Highways (MUTCD) criteria. The use of separate PSD criteria for design and marking is justified based on different needs in design and traffic operation (Harwood et al. 2009). Since the current

US highway system operates with relatively low level of crashes related to passing maneuvers and PSD, which indicates that the highway system can be operated safely with passing and no-passing zones marked with the current MUTCD criteria, therefore changing the current MUTCD PSD criteria to equal the AASHTO criteria, or some intermediate value, is not recommended because it would decrease the frequency and length of passing zones on two-lane, two-way highways. This would decrease the traffic level of service and might encourage illegal passes at locations where passing maneuvers are currently legal (Hardwood et al. 2009). As such, the AASHTO Green Book (2011) has adapted the MUTCD PSD values for the design of TLTW highways.

2.8.3.3: Driver’s Eye Height and Object’s Height for PSD

AASHTO Green book uses both the height of the driver’s eye and the object height as 1.08 m (3.5 ft) above the road surface (AASHTO 2011). This object height is based on a vehicle height of 1.33 m (4.35 ft), which represents the 15th percentile of vehicle heights in the current passenger car population, less an allowance of 0.25 m (0.85 ft), which is a near-maximum value for the portion of the vehicle height that needs to be seen for another driver to recognize a vehicle. The choice of an object height equal to the driver eye height makes design of passing sight distance reciprocal (i.e. when the driver of the passing vehicle can see the opposing vehicle, the driver of the opposing vehicle can also see the passing vehicle). Passing sight distances calculated on this basis are also considered adequate for night conditions because headlight beams of an opposing vehicle generally can be seen from a greater distance than a vehicle can be recognized in the daytime (AASHTO 2011).

2.8.3.4: PSD Model Assumptions

While there may be occasions, where multiple passing occurs when two or more vehicles pass a single vehicle or a single vehicle passes two or more vehicles. However, it is not practical to assume such conditions in developing minimum passing sight distance criteria. Instead, PSD is determined for a single vehicle passing a single vehicle (AASHTO 2011). Longer passing sight distances are recommended in the design and these locations can accommodate for an occasional multiple passing. AASHTO (2011) uses two theoretical models for the sight distance needs of passing drivers based on the assumption that a passing driver will abort the passing maneuver and return to his or her normal lane behind the overtaken vehicle if a potentially conflicting vehicle comes into view before reaching a critical position in the passing maneuver beyond which the passing driver is committed to complete the maneuver. The Glennon (1998) model assumes that the critical position occurs where the passing sight distance to complete the maneuver is equal to the sight distance needed to abort the maneuver. The Hassan et al. (1996) model assumes that the critical position occurs where the passing sight distances to complete or abort the maneuver are equal or where the passing and passed vehicles are abreast, whichever occurs first (AASHTO 2011). The following assumptions are made regarding the driver behavior in the passing maneuvers and PSD calculations based on the Glennon (1998) and Hassan et al. (1996) models (AASHTO 2011):

- The speeds of the passing and opposing vehicles are equal to the design speed.
- The overtaken vehicle travels at uniform speed.
- The Speed differential between the passing and overtaken vehicles is 19 km/h (12 mph).

- The passing vehicle has sufficient acceleration capability to reach the specified speed differential relative to the overtaken vehicle by the time it reaches the critical position, which generally occurs about 40 percent of the way through the passing maneuver.
- The lengths of the passing and overtaken vehicles are 5.8 m (19.0 ft).
- The passing driver's perception-reaction time in deciding to abort passing a vehicle is 1.0 sec.
- If a passing maneuver is aborted, the passing vehicle will use a deceleration rate of 3.4 m/s^2 (11.2 ft/s^2), the same deceleration rate used in stopping sight distance criteria.
- For a completed or aborted pass, the space headway between the passing and overtaken vehicles is 1.0 sec.
- The minimum time clearance between the passing and opposed vehicles at the point at which the passing vehicle returns to its normal lane is 1.0 sec.

2.8.3.5: PSD Calculations on Two-Lane Highways

The latest AASHTO Green Book of (2011) does not provide specific formulae for calculating the required PSD, however, previous versions of the Green Book (AASHTO 2001 and 2004) use the minimum passing sight distance for TLTW highways as the sum of the following four distances:

1) d_1 = Distance traversed during perception and reaction time and during the initial acceleration to the point of encroachment on the opposing lane, and is calculated as follows:

$$d_1 = 0.278t_i\,[v - m + (at_i / 2)] \qquad (2.32)$$

where;

t_i = time of initial maneuver, ranges from (3.6 to 4.5) sec,

a = average acceleration, ranges from (2.25 to 2.41) km/h/s,

v = average speed of passing vehicle (km/h),

m = difference in speed of overtaken vehicle and passing vehicle (km/h).

2) d_2 = Distance traveled while the passing vehicle occupies the left lane, and is determined as follows:

$$d_2 = 0.278vt_2 \qquad (2.33)$$

where;

t_2 = time passing vehicle occupies the left lane, ranges from (9.3 to 11.3) sec,

v = average speed of passing vehicle (km/h)

3) d_3 = Distance between the passing vehicle at the end of its maneuver and the opposing vehicle (the clearance length), ranges from (30.0 to 90.0) m.

4) d_4 = Distance traversed by an opposing vehicle for two-thirds of the time the passing vehicle occupies the left lane, or 2/3 of d_2 above, and ranges from (97.0 to 209.0) m. Figure 2.7 shows the AASHTO 2004 model for calculating PSD.

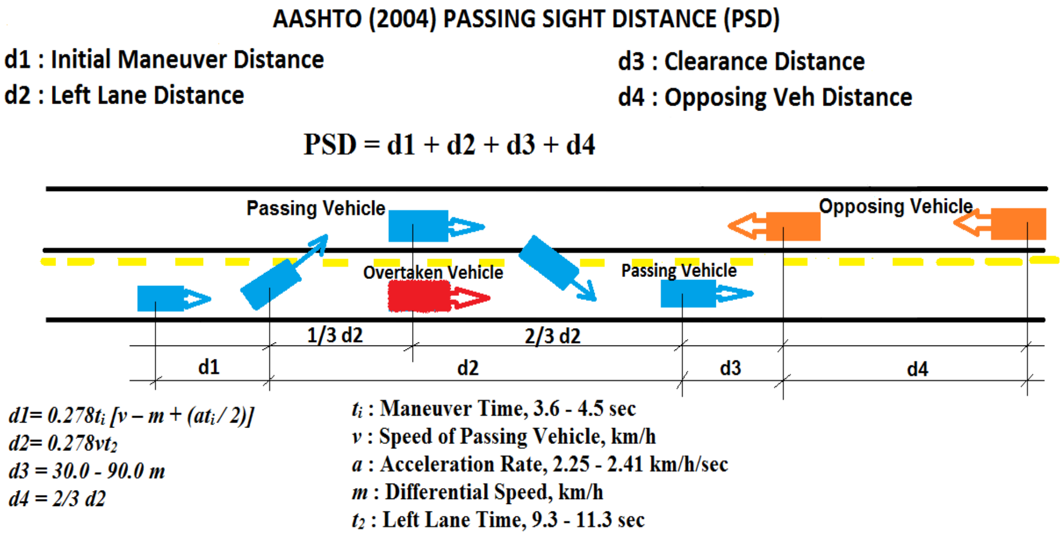
AASHTO (2004) PASSING SIGHT DISTANCE (PSD)
d1 : Initial Maneuver Distance
d2 : Left Lane Distance
d3 : Clearance Distance
d4 : Opposing Veh Distance
PSD = d1 + d2 + d3 + d4
Passing Vehicle
Opposing Vehicle
Overtaken Vehicle
Passing Vehicle
1/3 d2
2/3 d2
d1
d2
d3
d4
d1= 0.278ti [v – m + (ati / 2)]
d2= 0.278vt2
d3 = 30.0 - 90.0 m
d4 = 2/3 d2
ti : Maneuver Time, 3.6 - 4.5 sec
v : Speed of Passing Vehicle, km/h
a : Acceleration Rate, 2.25 - 2.41 km/h/sec
m : Differential Speed, km/h
t2 : Left Lane Time, 9.3 - 11.3 sec

Figure 2.7: AASHTO (2004) model for PSD calculations

Table 2.6 shows the minimum values of PSD required for the design of two-lane highways based on AASHTO 2011 Green Book. These values assume that a passing driver will abort the passing maneuver and return to his or her normal lane behind the overtaken vehicle if a potentially conflicting vehicle comes into view before reaching a critical position in the passing maneuver beyond which the passing driver is committed to complete the maneuver (AASHTO 2011).

Table 2.6: Minimum PSD values for design of two-lane highways

Design Speed (km/h)	Assumed Speeds (km/h)		Minimum Passing Sight Distance (m)
	Overtaken Vehicle	Passing Vehicle	
30	11	30	120
40	21	40	140
50	31	50	160
60	41	60	180
70	51	70	210

80	61	80	245
90	71	90	280
100	81	100	320
110	91	110	355
120	101	120	395
130	111	130	440

Source: AASHTO Green Book, 2011, Table 3-4.

2.8.3.6: Warrants for No-Passing Zones

Each passing zone along a length of roadway with sight distance ahead should be equal to or greater than the minimum passing sight distance should be as long as practical (AASHTO 2011). The criteria for marking passing and no-passing zones on two-lane highways are established by the MUTCD. Passing zones are not marked directly. Rather, the warrants for no-passing zones are set by the MUTCD, and passing zones merely happen where no-passing zones are not warranted (MUTCD 2012). Table 2.7 shows the MUTCD PSD warrants for no-passing zones. These criteria are based on prevailing off-peak 85th-percentile speeds rather than the design speeds.

Table 2.7: MUTCD warrants for NPZs

85th percentile speed Limit (km/h)	Minimum Passing Sight Distance (m)
40	140
50	160
60	180
70	210
80	245
90	280
100	320
110	355
120	395
130	440

2.8.3.7: Minimum Lengths of PZs

The MUTCD uses a minimum passing zone length of 120 m to 240 m (400 ft to 800 ft) depending on the 85th percentile speed limit, (i.e. where two no-passing zones come within 120 m to 240 m of one another, the no-passing barrier stripe should be continued between them). Table 2.8 shows the minimum passing zone Lengths to be Included in marking of PZs and NPZs (MUTCD 2012; AASHTO 2011). Figure 2.8 shows the AASHTO and MUTCD criteria for PSD and marking of NPZs.

Table 2.8: Minimum lengths of PZs

85th Percentile Speed Limit (km/h)	Minimum Passing Zone Length (m)
40	140
50	180
60	210
70	240
80	240
90	240
100	240
110	240
120	240

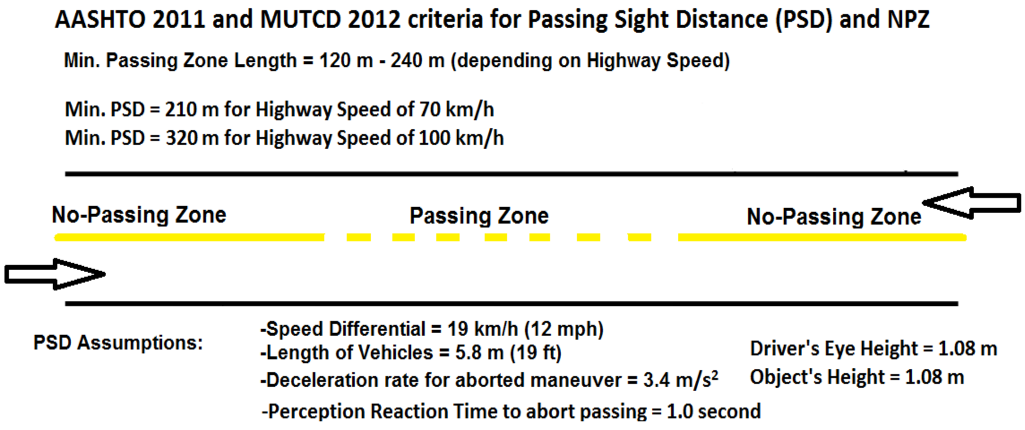
AASHTO 2011 and MUTCD 2012 criteria for Passing Sight Distance (PSD) and NPZ
Min. Passing Zone Length = 120 m - 240 m (depending on Highway Speed)
Min. PSD = 210 m for Highway Speed of 70 km/h
Min. PSD = 320 m for Highway Speed of 100 km/h
No-Passing Zone
Passing Zone
No-Passing Zone
PSD Assumptions:
-Speed Differential = 19 km/h (12 mph)
-Length of Vehicles = 5.8 m (19 ft)
-Deceleration rate for aborted maneuver = 3.4 m/s²
-Perception Reaction Time to abort passing = 1.0 second
Driver's Eye Height = 1.08 m
Object's Height = 1.08 m

Figure 2.8: AASHTO and MUTCD criteria for PSD and marking of NPZs

2.8.3.8: PSD on Horizontal Curves

The minimum passing sight distance for a two-lane road is greater than the minimum stopping sight distance at the same design speed (AASHTO 2011). To stick with those greater sight distances, Equation 18.2 for SSD on curves is directly applicable to passing sight distance but is of limited practical value except on long curves, because it would be difficult to maintain passing sight distance on other than very flat curves. Therefore, design for passing sight distance should be only limited to tangents and very flat curves. Even in level terrain, provision of passing sight distance would need a clear area inside each curve that would extend beyond the normal right-of-way line (AASHTO 2011).

2.8.3.9: PSD on Crest Vertical Curves

Length values of crest vertical curves for passing sight distance differ from those for stopping sight distance because of the different sight distance and object height criteria. Using the 1.08 m (3.50 ft) height of object results in the following formulas (AASHTO 2011):

when S is less than L:

$$L = \frac{AS^2}{864} \qquad (2.34)$$

when S is greater than L:

$$L = 2S - \frac{864}{A} \qquad (2.35)$$

where,

L: Length of vertical curve, m;

A: Algebraic difference in grade, percent;

S: Passing sight distance, m.

The minimum lengths of crest vertical curves are substantially longer than those for stopping sight distances (AASHTO 2011). The extent of difference is evident by the values of K, or length of vertical curve per percent change in A. Figure 2.9 shows the parameters used in determining the length of crest vertical curve based on PSD. Table 2.9 shows the minimum lengths of crest vertical curve as determined by PSD. Generally, it is impractical to design crest vertical curves that provide passing sight distance because of high cost and the difficulty of fitting the resulting long vertical curves to the terrain. Normally, passing sight distance is provided only at locations where combinations of alignment and profile do not need significant grading (AASHTO 2011).

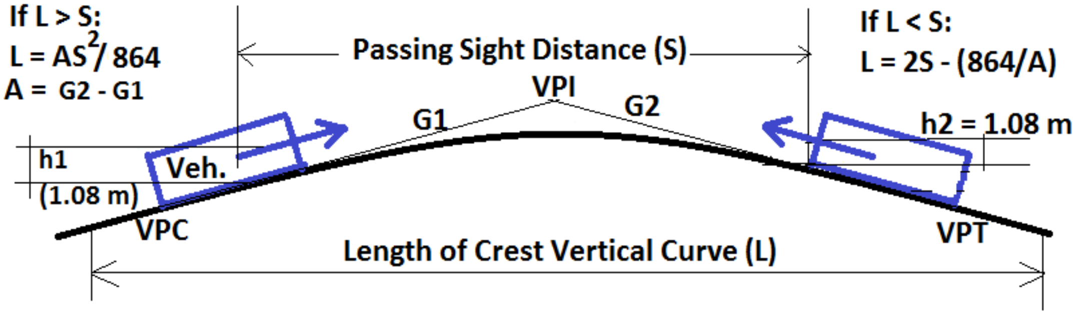

Figure 2.9: PSD parameters on crest vertical curves

Table 2.9: PSD design controls for crest vertical curves

Design Speed (km/h)	Passing Sight Distance (m)	Rate of Vertical Curvature, K
30	120	17
40	140	23
50	160	30
60	180	38
70	210	51
80	245	69
90	280	91
100	320	119
110	355	146
120	395	181
130	440	224

2.9: Temporal Autocorrelation

Temporal autocorrelation (i.e. serial correlation) is a special case of correlation, and refers not to the relationship between two or more variables, but to the relationship between successive values of the same variable. Temporal autocorrelation is closely related to the correlation coefficient between two or more variables, except that in this case we do not deal with variables *X* and *Y*, but with lagged values of the same variable. Most regression methods that are used in crash modeling assume that the error terms are independent from one another, and they are uncorrelated. This assumption is formally expressed (King 1981) as:

$$E\ (\varepsilon_i\ \varepsilon_j) = 0.0 \text{ for all } i \neq j \qquad (2.36)$$

where,

E: the expected value of all pair-wise products of error terms,

$\varepsilon_i\ \varepsilon_j$: error terms of the i and j observations respectively,

which means that the expected value of all pair-wise products of error terms is zero, and when the error terms are uncorrelated, the positive products will cancel those that are negative leaving an expected value of 0.0 (King 1981). If this assumption is violated, the standard errors of the estimates of the regression parameters are significantly underestimated which leads to erroneously inflated coefficients values, and incorrect confidence intervals. The presence of correlated error terms means that these types of inferences cannot be made reliably (Anderson 1984). The violation of this assumption occurs because of some temporal (time) component (i.e. heterogeneity due to time) that can affect the observations drawn across the time, such as time series data, panel data in the form of serial correlation, and any other dataset that might be collected over a period of time. In this context, the error in a first time period influences the error in a subsequent time period (either the previous period, or the next period or beyond) (King 1983). For example, we might expect the disturbance (i.e. error term) in year t to be correlated with the disturbance in year $t-1$ and with the disturbance in year $t+1$, $t+2$, and so on. If there are factors responsible for inflating the observation at some point in time to an extent larger than expected (i.e. a positive error), then it is reasonable to expect that the effects of those same factors linger creating an upward (positive) bias in the error term of a subsequent period. This phenomenon is called positive first-order autocorrelation, which is the most common manner in which the assumption of independence of errors is violated. For instance, if a dataset influenced by quarterly seasonal factors, then a resulting model that ignores the seasonal factors will have correlated error terms with a lag of four periods.

2.9.1: Sources of Temporal Autocorrelation

Temporal autocorrelation can arise from omitted explanatory variables, misspecification of the model form, and misspecification of the error terms. These will be discussed below in more detail.

2.9.1.1: Omitted Explanatory Variables

Omission some important explanatory variables can create temporal autocorrelation that can produce biased parameter estimates and incorrect inferences, especially if the omitted variable is correlated with variables included in the model (King 1981; Cameron and Trivedi 1998; Caliendo et al. 2007; Greene 2008).

2.9.1.2: Misspecification of the Mathematical Form

The model misspecification generates heterogeneity that can create temporal autocorrelation. For example, if a linear form of the model is specified when the true form of the model is non-linear, the resulting errors may reflect some temporal autocorrelation (Gujarati 1992; Miaou et al. 2003; Lord and Bonneson 2007; Hilbe 2014).

2.9.1.3: Misspecification of The Error Term

Successive values of the error term may be related due to some purely random factors, such as changes in weather conditions, economic factors, and other unaccounted for variables, which could have changing effects over successive periods. In such instances, the value of the error term in the model could be misspecified (King 1983).

2.9.2: Structure of temporal autocorrelation

There are different structure types of temporal autocorrelation: 1st order, 2nd order, and so on. The form of temporal autocorrelation that is encountered most often is called the first order serial correlation in the first autoregressive term, which is denoted by AR (1). The AR (1) autocorrelation assumes that the disturbance in time period *t* (current period) depends upon the disturbance in time period *t-1* (previous period) plus some additional amount, which is an error, and can be modeled as (King 1983):

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + \epsilon_t \qquad (2.37)$$

where,

ε_t : the disturbance in time period *t*,

ε_{t-1} : the disturbance in time period *t-1*,

ρ: the autocorrelation coefficient,

ϵ_t : the model error term.

The parameter ρ can take any value between negative one and positive one. If $\rho > 0$, then the disturbances in period *t* are positively correlated with the disturbances in period *t-1*. In this case, positive autocorrelation exists which means that when disturbances in period *t-1* are positive disturbances, then disturbances in period *t* tend to be positive. When disturbances in period *t-1* are negative disturbances, then disturbances in period *t* tend to be negative. Temporal datasets are usually characterized by positive autocorrelation. If $\rho < 0$, then the disturbances in period *t* are negatively correlated with the disturbances in period *t-1*. In this case there is negative autocorrelation. This means that when disturbances in period *t-1* are positive disturbances, then disturbances in period *t* tend to be negative. When disturbances in period *t-1* are negative disturbances, then disturbances in period *t* tend to be positive.

The second order serial correlation is called the second-order autoregressive process or AR (2). The AR (2) autocorrelation assumes that the disturbance in period *t* is related to both the disturbance in period *t-1* and the disturbance in period *t-2*, and can be modeled as (King 1983):

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \epsilon_t \qquad (2.38)$$

where,

ρ_1: the autocorrelation coefficient in time period *t-1*.

$\rho_{1:}$ the autocorrelation coefficient in time period *t-2*.

The disturbance in period *t* depends upon the disturbance in period *t-1*, the disturbance in period *t-2*, and some additional amount, which is an error (ϵ_t). In a similar manner, the temporal autocorrelation can be extended to the ρth order autocorrelation AR (ρ). However, the most often used temporal autocorrelation is the first-order autoregressive process (King 1983).

2.9.3: Detection of Temporal Autocorrelation

There are several ways to detect the existence of the temporal autocorrelation in the dataset, including the residuals scatter plots, the Durbin-Watson test, the Durbin h test, the Breusch-Godfrey test, the Ljung-Box Q test, and correlograms. These will be described in detail below.

2.9.3.1: Scatter Plot of Residuals

The error for the i^{th} observation in the dataset is usually unknown and unobservable. However, the residual for this observation can be used as an estimate of the error, then the residuals can be plotted against the variables that may be related to time. The residual would be measured on the vertical axis. The temporal variables such as, years, months, or days would be measured on the horizontal axis. Next, the residual plot can be examined to determine if the residuals appear to exhibit a pattern of temporal autocorrelation. If the data are independent, then the residuals should be randomly scattered about 0.0. However, if a noticeable pattern emerges (particularly one that is cyclical or seasonal) then temporal autocorrelation is likely an issue. It must be emphasized that this is not a formal test of serial correlation. It would only suggest whether temporal autocorrelation may exist. We should not substitute a residual plot for a formal test (King 1981; Hilbe 2014).

2.9.3.2: The Durbin-Watson (DW) Test

The most often used test for first order temporal autocorrelation is the Durbin-Watson *DW* test (Hilbe 2014). The *DW* test is a measure of the first order autocorrelation and it cannot be used to test for higher order temporal autocorrelation. The *DW* test is constructed to test the null and alternative hypotheses regarding the temporal autocorrelation coefficient (ρ):

$$H_0: \ \rho = 0.0, \quad H_a: \ \rho \neq 0.0 \tag{2.39}$$

The null hypothesis of ρ = 0.0 means that the error term in one period is not correlated with the error term in the previous period, while the alternative hypothesis of $\rho \neq 0.0$ means the error term in one period is either positively or negatively correlated with the error term in the previous period. To test the hypothesis, the *DW* test statistic on a dataset of size n is formulated as (King 1981):

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t{}^2} \tag{2.40}$$

where,

DW: the Durbin-Watson statistic,

e_t: the residual error term in time period t,

e_{t-1}: the residual error term in the previous time period $t - 1$.

The *DW* statistics ranges from 0.0 to 4.0, and it can be shown that:

$$DW = 2\ (1 - \rho^{\wedge}) \tag{2.41}$$

where,

$\rho^{\wedge}$: the residual temporal autocorrelation coefficient.

When $\rho^{\wedge} = 0.0$, (i.e. no autocorrelation), then $DW = 2.0$.

When $\rho^{\wedge}$ tends to 1.0, then $DW = 0.0$.

When $\rho^{\wedge}$ tends to -1.0, then $DW = 4.0$.

The critical values of *DW* for a given level of significance, sample size and number of independent variables can be obtained from published tables that are tabulated as pairs of values: DL (lower limit of *DW*) and DU (upper limit of *DW*). To evaluate *DW* King (1983) suggests:

1) Locate values of DL and DU in Durbin-Watson statistic table.

2) For positive temporal autocorrelation:

a) If *DW* < DL then there is positive autocorrelation.

b) If *DW* > DU then there is no positive autocorrelation.

c) If DL < *DW* < DU then the test is inconclusive.

3) For negative temporal autocorrelation:

a) If *DW* < (4.0 – DU) then there is no negative autocorrelation.

b) If *DW* > (4.0 – DL) then there is negative autocorrelation.

c) If (4.0 – DU) < *DW* < (4.0 – DL) then the test is inconclusive.

A rule of thumb that is sometimes used is to conclude that there is no first order temporal autocorrelation if the *DW* statistic is between 1.5 and 2.5. A *DW* statistic below 1.5 indicates positive first order autocorrelation. A *DW* statistic of greater than 2.5 indicates negative first order autocorrelation (King 1983). Alternatively, a significant *p*-value for the *DW* statistic would suggest to reject the null hypothesis and conclude that there is first order autocorrelation in the residuals, and a non-significant *p*-value would suggest accepting the null hypothesis and concluding that there is no evidence of first order autocorrelation in the residuals.

2.9.3.3: The Durbin *h* Test

When one or more lagged dependent variables are present in the data, the *DW* statistic will be biased towards 2.0, this means that even if temporal autocorrelation is present it may be close to 2.0, and hence it cannot detect it. Durbin suggests a test for temporal autocorrelation when there is a lagged dependent variable in the dataset, and it is based on the h statistics. The Durbin h statistics is defined as:

$$h = \rho^\wedge \sqrt{\frac{T}{1-T\ [\ VAR\ (\beta^\wedge)}} \qquad (2.42)$$

where,

T: the number of observations in the dataset,

$\rho^\wedge$: the temporal autocorrelation coefficient of the residuals,

$VAR\ (\beta^\wedge)$: the variance of the coefficient on the lagged dependent variable.

Durbin has shown that the h statistics is approximately normally distributed with a unit variance, hence the test for first order autocorrelation can be done using the standard normal distribution. If Durbin h statistic is equal to or greater than 1.96, it is likely that temporal autocorrelation exists (King 1981).

2.9.3.4: The Breusch-Godfrey Lagrange Multiplier (LM) Test

The Breusch-Godfrey test is a general test of serial correlation and can be used to test for first order temporal autocorrelation or higher order autocorrelation. This test is a specific type of Lagrange Multiplier test. The *LM* test is particularly useful because it is not only suitable for testing for temporal autocorrelation of any order, but also suitable for models with or without lagged dependent variables (Thomas 1993). The null and alternative hypotheses used with this test for a second order autocorrelation are:

$$H_0: \rho_1 = \rho_2 = 0.0, \quad H_1: \text{ At least one } \rho \text{ is not zero} \qquad (2.43)$$

The *LM* test statistic is given by:

$$LM = (n - i)\ R^2 \qquad (2.44)$$

where,

LM: the Lagrange multiplier test statistic,

n: the number of observations in the dataset,

i: the order of the autocorrelation,

R^2: the unadjusted R^2 statistic (coefficient of determination) of the model.

The *LM* statistic has a chi-square distribution with two degrees of freedom, $\chi 2(2)$ (Studenmund 2001).

2.9.3.5: The Ljung-Box Q (LBQ) Test

The Ljung-Box *Q* test (sometimes called the Portmanteau test) is used to test whether or not observations taken over time are random and independent for any order of temporal autocorrelation. It is based on asymptotic Chi-Square distribution χ^2. In particular, for a given i lag, it tests the following hypotheses:

H_0: the autocorrelations up to i lags are all zero, (2.45)

H_a: the autocorrelations of one or more lags differ from zero (2.46)

The test statistic provided by Box et al. (1994) is:

$$LBQ_i = n\,(n+2)\sum_{j=1}^{i}\frac{{r_j}^2}{n-j} \qquad (2.47)$$

where,

LBQ_i: the Ljung-Box *Q* statistic,

n: the number of observations in the data,

j: the lag being considered,

i: the autocorrelation order,

r: the residual error term in lag j.

2.9.3.6: Correlograms

The correlograms are autocorrelation plots that can show the presence of temporal autocorrelation. The autocorrelation would appear in lag 1.0 and progress for n lags then disappear. In these plots the residual autocorrelation coefficient (ρ^) is plotted against n lags to develop a correlogram. This will give a visual look at a range of autocorrelation coefficients at relevant time lags so that significant values may be seen (Chatfield 1996). In most software packages, two types of autocorrelation functions are presented: the autocorrelation function (ACF), and the partial autocorrelation function (PACF). The ACF is the amount of autocorrelation between a variable and a lag that is not explained by correlations at all lower-order-lags, and the PACF is the difference between the actual correlation at specific lag and the expected correlation due to propagation of correlation at the previous lag. If the PACF displays a sharp cutoff while the ACF decays more slowly we conclude that the data displays an autoregressive model (AR), and the lag at which the PACF cuts off is the indicated number of AR terms. If the ACF of the data displays a sharp cutoff and/or the lag-1 autocorrelation is negative then we have to consider adding a moving average term (MA) to the model, and the lag at which the ACF cuts off is the indicated number of MA terms. In general, the diagnostic patterns of ACF and PACF for an AR (1) term (Warner 1998) are:

ACF: declines in geometric progression from its highest value at lag 1.0.

PACF: cuts off abruptly after lag 1.0.

If the ACF of a specific variable shows a declining geometric progression from the highest value at lag 1.0, and the PACF shows an abrupt cut off after lag 1.0., this would indicate that this variable has not encountered temporal autocorrelation.

2.9.4: Remedies for Temporal Autocorrelation

When temporal autocorrelation is determined to be present in the dataset, then one of the first remedial measures should be to investigate the omission of one or more of the key explanatory variables, especially variables that are related to time. If such a variable does not aid in reducing or eliminating temporal autocorrelation of the error terms, then a differencing procedure should be applied to all temporal independent variables in the dataset to convert them into their differences values, and rerun the regression model by deleting the intercept from the model (Chatfield 1996). If this remedy does not help in eliminating temporal autocorrelation, then certain transformations on all variables can be performed for the AR (1) term. These transformations aim at performing repeated iterative steps to minimize the squared sum of errors in the regression model. Examples of such transformations are: Cochrane-Orcutt procedure; and Hildreth-Lu procedure. More advanced methods can also be used for big datasets such as: the Fourier series analysis; and the spectral analysis (Chatfield 1996; Warner 1998).

2.10: Spatial Autocorrelation

Spatial autocorrelation is the correlation of a variable with itself through space. In most vehicle accident studies, crash incidents are aggregated to a spatial unit of analysis, such as intersections, road segments, zip codes, wards or county levels (Amoros et al. 2003; Noland and Quddus 2004). However, aggregation of individual crash incidents can potentially misrepresent relationships among the original observations that may be important when reasoning about the factors underlying the occurrence of crashes. One such concern is the existence of spatial autocorrelation among crash incidents, which if present, can adversely influence predictive measures, resulting in higher variances of the estimates and consequently, underestimated standard errors (MacNab 2004).

The spatial autocorrelation phenomenon can be best summarized by the Tobler's first law of Geography that everything is usually related to all else but those which are near to each other are more related when compared to those that are further away (Tobler 1970). Accordingly, spatial autocorrelation is a measure of the correlation of an observation with other observations through space. Spatial autocorrelation can be positive or negative among observations. Positive spatial autocorrelation occurs when observations having similar values are closer (i.e. clustered) to one another, and negative spatial autocorrelation occurs when observations having dissimilar values occur near (i.e. clustered) one another (Anselin 1988; Bailey and Gatrell 1995). Most statistical analyses are based on the assumption that the values of observations in each sample are independent of one another. Spatial autocorrelation violates this assumption, because samples taken from nearby locations are related to each other, and hence, they are statistically not independent of one another (Black 1992; Baily and Gatrell 1995). Crash data are usually collected with reference to locations measured as points (with x- and y-coordinates) in space, or to road segments (i.e. mile markers). Two problems may be faced when sample data has a locational dimension: (1) the existence of spatial autocorrelation between the observations, and (2) the variation of the relationship over the space that could be described as spatial heterogeneity (LeSage and Pace 2009) or spatial non-stationarity (Fotheringham et al. 2002). Therefore, the consideration of spatial autocorrelations has been gaining attention in crash modeling in recent years, and researchers have shown that ignoring this factor may lead to a biased estimation of the model parameters (El-Basyouny and Sayed 2006; Mitra and Washington 2007; Aguero-Valverde and Jovanis 2008; Mohammadi, et al. 2014).

Taking the spatial autocorrelation into account in crash modeling can improve model parameter estimation, and the overall model fit (Aguero-Valverde and Jovanis 2008; El-Basyouny and Sayed 2009). Traditional non-spatial modeling approaches to crash analysis may not be able to capture the effect of spatial autocorrelation of the neighborhood locations on traffic crashes, which could result in a violation of the traditional Gauss–Markov assumptions used in traditional regression modelling (Wang and Abdel-Aty 2006). Hence, spatial autocorrelation must be incorporated in the crash analysis modeling to properly account for the effect of spatial correlation and any unobserved heterogeneity that may exist in the crash data. Black (1992) examins the differences between the network autocorrelation and spatial autocorrelation. In his study, he demonstrates that the network autocorrelation could influence the values associated with a network link given its relationship to another link in the network. To account for these relationships, spatial autocorrelation was only modeled between neighboring (adjacent) network links. Levine et al. (1995) examines the effect of spatial autocorrelation on traffic crashes by geo-coding them to the nearest intersection or ramp, and then calculating different spatial statistics such as, mean, standard deviation, and standard deviational ellipse. In another study Black and Thomas (1998) explores spatial autocorrelation of road segments by using the Moran's Index. They conclude that there was a significant level of positive spatial autocorrelation in the data. When investigating spatial autocorrelation among traffic crashes, Miaou et al. (2003) estimate a series of crash frequency models aggregated at the county level for the state of Texas. Wang and Abdel-Aty (2006) analyze rear-end crashes at signalized intersections to model the spatial correlation between intersections. In their study, three different correlation structures are considered: independent correlation, exchangeable correlation, and autoregressive correlation, where the correlation decreases as the gap between intersections increases. The models proved that high spatial correlations exist between intersections for rear-end crashes. Guoa et al. (2010) propose spatial models for intersections, using the distance adjacency method for the spatial correlation determination. Further, Wang and Kockelman (2013) apply a multivariate spatial modeling method for pedestrian and bicyclist collision at the census tracts levels. Chiou et al. (2014) utilize spatial multinomial generalized Poisson models to explore the spatial autocorrelation, and find that spatial correlation sharply decreases at distances exceeding 7 km, and shorter road segments with high crash frequency tend to have high spatial dependency. Aguero-Valverde (2013) develop a multivariate spatial modeling approach for excess crash frequency and severity in cantons (counties) for Costa Rica, and report that the multivariate spatial model performed better than univariate spatial models. They also report that the effects of spatial smoothing due to multivariate spatial random effects were evident in the estimation of no-injury collisions.

2.10.1: Weight Matrix of Spatial Autocorrelation

To assess spatial autocorrelation, a distance measure must be specified in order to define what is meant by two observations being close together. These distances are usually presented in the form of a weights matrix, which defines the relationships between locations at which the observations occur (Cliff and Ord 1981). If data are collected at *n* locations, then the weight matrix will be *n x n* with zeroes on the diagonal. The weight matrix is often row-standardized, (i.e. all the weights in a row sum to one), and can be constructed given a variety of assumptions (Bailey and Gatrell 1995), such as:

- A constant distance that represents the weight for any two different locations.
- A fixed weight for all observations within a specified distance.
- *k* nearest neighbors that represents a fixed weight, and all others are zero.
- Weight could be proportional to the inverse distance, or inverse distance squared.

2.10.2: Indices of Spatial Autocorrelation

There are a number of indices or statistics that attempt to measure spatial autocorrelation for count data, such as the Moran's *I*, the Geary's *C*, and the Getis- Ord *G* statistic (Fischer and Wang 2011). These indices can be computed as *Global* or *Local* measures depending on the scope of the analysis. Global spatial autocorrelation identifies and measures the spatial pattern of the entire study area. Local spatial autocorrelation identifies spatial variation across the study area considering the relationship between individual features. Anselin (1995) outline a general class of local indicators of spatial autocorrelation termed the Local Indicator of Spatial Autocorrelation (*LISA*) statistic that satisfies two conditions, first; the *LISA* for each point or section in the space gives an indication of significant spatial clustering (grouping) of similar or dissimilar values around that point or section, and second; the sum of *LISAs* for all points or sections in a given study area is proportional to a corresponding global indicator of spatial autocorrelation for that area, which implies that the *LISA* statistic decomposes global results into their local parts. For example, a significant global index at a given spatial point or section may hide large spatial patches of no autocorrelation, and *LISA* can detect this and show us the location of these insignificant patches in space. Conversely, an insignificant global result may hide patches of strong autocorrelation, and *LISA* can detect this again. Therefore, the *LISA* concept in measuring the local spatial autocorrelation is very useful by uncovering hidden, local patterns in data that the global statistics average over (Anselin 1995).

2.10.3: Global Indices vs. Local Indices

There are two types of spatial autocorrelation indices that can be used: Global and Local indices depending on the scope of the analysis. Global spatial autocorrelation identifies the spatial pattern of the entire study area (i.e. whether the overall area is clustered, dispersed, or random). Local spatial autocorrelation identifies spatial variation across the study area considering the relationship between individual features, resulting in specific areas of clustering (Anselin 1995; Fischer and Wang 2011). Global implies that all elements in the weight matrix are included in the calculation of spatial autocorrelation providing a single measurement of spatial autocorrelation for an entire data. Local indices calculate spatial autocorrelation for individual units within the study area. Indices of spatial autocorrelation are based on the general index of matrix association (i.e. the Gamma Γ index). The Global Gamma index consists of the sum of the cross products of the elements a_{ij} and b_{ij} in two matrices of similarity, using spatial similarity in one matrix (i.e. spatial weight matrix) and value similarity in the other matrix, such that (Anselin 1995):

$$\Gamma = \sum_i \sum_j a_{ij}\, b_{ij} \qquad (2.48)$$

Using different value similarity would result in different indices. For example, setting $a_{ij} = x_i\, x_j$ would result in Moran's *I* statistic, and setting $a_{ij} = (x_i - x_j)^2$ would result in Geary' *C* index (Anselin 1995). The local Gamma index of a location i is defined as (Anselin 1995):

$$\Gamma_i = \sum_j a_{ij} b_{ij} \qquad (2.49)$$

Again using different value similarity would result in different indices. The Global Gamma index equals the sum of local Gamma indices within the study area (Anselin 1995).

2.10.4: Moran's *I*

Moran's *I* statistic is one of the oldest indices of spatial autocorrelation and can be used to test for global and local spatial autocorrelation among continuous data. For any continuous variable, x_i, a mean $\overline{x}$, can be calculated and the deviation of any observation from that mean can be calculated based on the cross products of the deviations from the mean. The statistic then compares the value of the variable at any one location with the values at all other locations (Griffith 1987; Goodchild 1987; Anselin 1992). For n observations on a variable x at locations i, j, Global Moran's *I* can be calculated as follows (Anselin 1995):

$$I = \frac{\boldsymbol{n}}{\boldsymbol{S_0}} \frac{\sum_i^n \sum_j^n \boldsymbol{w_{ij}}\,(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{x_j} - \overline{\boldsymbol{x}})}{\sum_i^n (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2} \qquad (2.50)$$

where,

$\bar{x}$: the mean of the variable x;

x_i : the value of variable x at location i;

x_j : the value of variable x at location j;

w_{ij}: the elements of the weight matrix;

n : number of observations;

S_0 : is the sum of the elements of the weight matrix: $S_0 = \sum_{\boldsymbol{i}}^{\boldsymbol{n}} \sum_{\boldsymbol{j}}^{\boldsymbol{n}} \boldsymbol{w_{ij}}$

The local Moran's *I* for location i can be calculated as follows:

$$I_i = \frac{x_i - \bar{x}}{S_i^2} \sum_{j=1}^{n} w_{ij} \left(x_j - \bar{x}\right) \qquad (2.51)$$

$$S_i^2 = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n-1} - \bar{x}^2 \qquad (2.52)$$

Values for this index typically, range from -1.0 to +1.0, where a value of -1.0 indicates negative spatial autocorrelation, and a value of +1.0 indicates positive spatial autocorrelation. When nearby points have similar values, their cross product is high. Conversely, when nearby points have dissimilar values, their cross-product is low. The expectation of Moran's *I* statistic is:

$$E\ (I) = \left(\frac{\boldsymbol{-1}}{\boldsymbol{n-1}}\right) \qquad (2.53)$$

with a Moran's *I* value larger than *E (I),* indicates positive spatial autocorrelation, and a Moran's *I* less than *E (I),* indicates negative spatial autocorrelation. In Moran's initial formulation, the weight variable, w_{ij}, was a contiguity matrix. If zone j is adjacent to zone $_{i}$, the product receives a weight of 1.0. Otherwise, the product receives a weight of 0.0. Cliff and Ord (1975) generalized these definitions to include any type of weight, and in more current use, w_{ij}, is a distance-based weight which is the inverse distance between locations i and j ($1/d_{ij}$). The z-score of Moran's I can be computed in Eq. (2.54):

$$Z_i = \frac{I - E\ (I)}{\sqrt{V\ (I)}} \qquad (2.54)$$

where $E\ (I)$ is the expected value of I, and $V\ (I)$ is the variance of I, as shown in Eq. (2.55):

$$V\ (I) = E\ (I^2) - E^2\ (I) \qquad (2.55)$$

These z-scores express the difference between the observed and expected value of I in standard deviation units. The distribution of the z-scores is assumed to be approximately normal with a mean of 0.0 and a variance of 1.0 (Cliff and Ord 1981). A statistically significant positive z-score indicates that the distribution of the observations are spatially autocorrelated, whereas a negative z-score indicates that the observations tend to be more dissimilar. A z-score close to zero indicates that observations are randomly and independently distributed in space. By assuming a z-score is from a standard normal distribution, their associated p-value can be obtained, and can be used to determine the significance of the index at each location (Cliff and Ord 1981). To determine if the z- score is statistically significant, it should be compared to the range of values for a particular confidence level. For example, at a significance level of 95%, a z-

score would have to be less than –1.96 or greater than + 1.96 to be statistically significant. The null hypothesis H_0 is that there is no local spatial autocorrelation among the observations. The null hypothesis can be rejected, if the p-value shows that the z-score is significant, and the next step is to inspect the value of the Moran's *I* index. If the value is greater than 0.0, then observations are spatially autocorrelated, and the pattern is clustered, and if the value is less than 0.0, then the pattern is more dispersed (Griffith 1987; Baily and Gatrell 1995). In case of non-normality, the randomization distribution should be used to test the null hypothesis of no local autocorrelation H_0 (Anselin 1995) .

2.10.5: Getis-Ord *G* statistic

The Getis-Ord *G* statistic is a widely used index of spatial autocorrelation but for values that fall within a specified distance of each other (Getis and Ord 1992; Ord and Getis 1995). The *G* statistic is calculated with respect to a specified threshold distance (defined by the user) rather than to an inverse distance, as with the Moran’s *I*. The General (Global) *G* statistic computes a single statistic for the entire study area, while the G_i statistic is an indicator for local spatial autocorrelation for each data point. The Global *G* statistic can be calculated as follows (Fischer and Wang 2011):

$$G = \frac{\sum_i \sum_{j \neq i} w_{ij} x_i x_j}{\sum_i \sum_{j \neq i} x_i x_j} \qquad (2.56)$$

where,

x_i : the value of variable x at location i;

x_j : the value of variable x at location j;

w_{ij}: the elements of the weight matrix.

There are two types of local G_i statistics, although almost the two types produce identical results (Getis and Ord 1996; Berglund and Karlstrom 1999). The first one, G_i, does not include the autocorrelation of a zone with itself, whereas the G_i* includes the interaction of a zone with itself (i.e. the G_i statistic does not include the value of X_i itself, but only the neighborhood values, but G_i* includes X_i as well as the neighborhood values), and both can be computed by the formulae (Fischer and Wang 2011):

$$G_i\ (d) = \frac{\sum_{j\neq i}^{n} w_{ij}\, x_j}{\sum_{j\neq i}^{n} x_j} \qquad (2.57)$$

$$G_i^*\ (d) = \frac{\sum_{j=1}^{n} w_{ij}\, x_j}{\sum_{j=1}^{n} x_j} \qquad (2.58)$$

where d is the neighborhood (threshold) distance, and w_{ij} is the weight matrix that has only 1.0 or 0.0 values, 1.0 if j is within d distance of i, and 0.0 if its beyond that distance. These formulae indicate that the cross-product of the value of X at location i and at another location j is weighted by a distance weight, w_{ij} which is defined by either a 1.0 if the two locations are equal to or closer than a threshold distance, d, or a 0.0 otherwise. The G statistic can vary between 0.0 and 1.0. The statistical significance of the local autocorrelation between each point and its neighbors is assessed by the z-score test and the p-value.

ArcGIS uses the following formulae to calculate the local Getis-Ord *Gi** (ESRI 2016 a):

$$Gi^*= \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{x}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{n\sum_{j=1}^{n} w^2{}_{ij} - \left(\sum_{j=1}^{n} w_{ij}\right)^2}{n-1}}} \qquad (2.59)$$

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \qquad (2.60)$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x^2{}_j}{n} - \bar{x}^2} \qquad (2.61)$$

where,

x_i : the value of variable x at location i;

x_j : the value of variable x at location j;

w_{ij} : the elements of the weight matrix;

n : number of observations.

The expected *G* value for a threshold distance, d, is defined as (Lee and Wong 2005):

$$E\,[G\,(d)] = \frac{W}{n\,(n-1)} \qquad (2.62)$$

where *W* is the sum of weights for all pairs of locations ($W = \sum_{i}^{n}\sum_{j}^{n} w_{ij}$), and n is the number of observations.

Assuming normal distribution, the variance of *G (d)* is defined as (Lee and Wong 2005):

$$Var[G\ (d)] = E\ (G^2) - E^2\ (G) \qquad (2.63)$$

The standard error of *G (d)* is the square root of the variance of *G*. Therefore, a *z*-test can be computed by:

$$S.\ E.\ [G\ (d)] = \sqrt{Var\ [\ G\ (d)]} \qquad (2.64)$$

$$Z\ [G\ (d)] = \frac{G\ (d) - E[G\ (d)]}{S.E.[G\ (d)]} \qquad (2.65)$$

where a positive *z*-value indicates spatial clustering of high values, while a negative *z*-value indicates spatial clustering of low values. Sometimes, the *G* statistic may not follow a normal standard error, and the distribution of the statistic may not be normally distributed, such as the case of a skewed variable with some points having very high values while the majority of other points having low values. In this case, a permutation type simulation should be used (Anselin 1995; Mobley et al. 2008), with a randomization distribution to test the null hypothesis of no local autocorrelation (H_0). This will maintain the distribution of the variable *z* but will estimate the value of *G* under random assignment of this variable, and the user can take the usual 95% or 99% confidence intervals based on the simulation used.

The type of concentration of incidents (i.e. clustering) and its statistical significance is evaluated based on a confidence level and on the output *z*-scores and the correspondent *p*-values. These will determine whether a data point is classified as having a significant high spatial autocorrelation (denoted by High-High, HH), significant low spatial autocorrelation (denoted by Low-Low, LL), a significant dispersed outlier (a high data value surrounded by low data values or vice versa, denoted by High-Low, HL or Low-High, LH), or insignificant random crash.

2.11: The Multinomial Logistic Regression (MNL)

The multinomial logistic regression (MNL) can model the relationships between a polytomous (multinomial) dependent variable (with more than two outcomes) and a set of independent variables. It is an extension of the binary logistic regression, which analyzes dichotomous (binary) dependent variables with only two outcomes. The multinomial logit model may be used to handle a dependent variable that is a categorical, unordered variable (i.e. cannot be ordered in any logical way). Ordered logistic regression is used in cases where the dependent variable is ordered in a certain way. The MNL works by choosing one group as the base (reference) category for the other groups. Then MNL contrasts all the outcomes of the dependent variable with this common reference category, which serves as the contrast point for all analyses, and the effects of the analysis are always in reference to the contrast category (Greene 2012).

The MNL applies the assumption of the independence of irrelevant alternatives (IIA), which means that adding or deleting alternative severity outcome categories does not affect the prediction among the remaining severity outcomes. In other words, the odd ratios produced by the logit function for any pair of severity outcomes are determined without reference to the other categories that might be available (McFadden et al. 1976; Hausman 1978), and therefore it must be checked in the modeling process.

2.11.1: The Advantages of the MNL

The MNL has many advantages in modeling crash severity, such as (Kleinbaum and Klein 2010; Baltagi 2011; Greene 2012):

- It produces sound estimates as it applies transformation of the multinomial dependent variable to a continuous variable ranging from negative infinity to positive infinity. It is usually difficult to model a variable which has restricted range, such as probability. This transformation attempts to overcome this problem. It changes probability ranging between 0.0 and 1.0 to log odds ranging from negative infinity to positive infinity.
- Among all of the many choices of transformation, the log of odds in MNL is one of the easiest to understand and interpret.
- The results of MNL can be interpreted by both the regression coefficient estimates and/or the odd ratios (the exponentiated coefficients) as well.
- The estimates are asymptotically consistent with the requirements of the nonlinear regression process.
- MNL can be used to improve the fitted model by comparing the full model that include all predictors to a chosen restricted models by excluding the non-significant predictors, then picks up the best fit.

CHAPTER 3: METHODS

This dissertation describes a framework for modeling the severity of vehicular crashes given the location (i.e. the latitude and longitude) and other characteristics of crash incidence along roadways that can be obtained from crash data. For this type of analysis, one must first test for temporal autocorrelation among the independent variables that are related to the time, and correct for any significant temporal autocorrelation that might exist before using the data. Next, the spatial autocorrelation of the vehicular crashes must be tested to identify the clustering patterns of the incidents. If the vehicular crashes are found to be significantly clustered due to spatial autocorrelation, then an appropriate spatial autocorrelation index must be incorporated in the modeling process as a risk factor. Next, the available sight distance of the roadway must be determined according to AASHTO (2011) criteria, and roadway segments with sight distances that may potentially not conform to the AASHTO (2011) standards can be incorporated in the modeling process as risk factors. Lastly, one must apply an appropriate statistical process that can identify the risk factors contributing to different crash severity categories.

3.1: Evaluation of the Temporal Autocorrelation (TA)

Temporal autocorrelation (also called serial correlation) refers to the relationship between successive values (i.e. lags) of the same variable. Although it is a major concern in time series models, however, it is very important to be checked in crash modeling as well (Washington et al. 2010; Lord and Mannering 2010; Savolainen et al. 2011). The results of crash modeling can be improved when several years of crash data are utilized in the analysis, such as a period of three years instead of one year (Mohammadi et al. 2014).

However, this means that the same roadway will generate multiple observations over time, which could be correlated due to some temporal (time) component and could adversely affect the precision of parameter estimates (Washington et al. 2010; Lord and Mannering 2010; Savolainen et al. 2011). There are several methods in the literature that can be used to detect the existence of the temporal autocorrelation in the dataset, such as: 1) the residuals scatter plots; 2) the Durbin-Watson (*DW*) test; 3) the Durbin *h* test; 4) the Breusch-Godfrey (*LM*) test; 5) the Ljung-Box Q (*LBQ*) test; and 6) correlograms. The residuals scatter plots and correlograms are not formal tests, and they would only suggest whether temporal autocorrelation may exist within crash data (Hilbe 2014). The Durbin *h* test can only be used when there is a lagged dependent variable in the dataset (King 1981; Hilbe 2014). This dissertation uses the Durbin-Watson (*DW*), Breusch-Godfrey (*LM*), and the *LBQ* tests to detect the temporal autocorrelation among the temporal independent variables in the crash data (i.e. hour, weekday, month). Although the applications of these tests can be found in time series models, they have not been addressed in modeling crash severity (Lord and Mannering 2010; Savolainen et al. 2011). As such, this dissertation investigates the applicability of these tests. These tests can be applied at different levels of temporal aggregation (i.e. over one year, over two years, three years, etc.) to help identify any hidden effects of the temporal autocorrelation that might exist within a timeframe. In this dissertation, the JMP12 software package is used to compute the *DW* statistics, the associated residual temporal autocorrelation coefficients, and their significance at the 95% confidence level (i.e. *p*-values). JMP requires that the input format of the crash data be in either excel spreadsheet (i.e. *.xlsx) or in text (i.e. delimited or *.csv) and then the output is produced as excel spreadsheet or delimited text.

The Eviews 9 software is used to compute the *LM* statistics, and their significance at the 95% confidence level (i.e. *p*-values). The software requires that the input format of the crash data be in either excel spreadsheet (i.e. *.xlsx) or in text (i.e. delimited or *.csv) and then the output is produced as excel spreadsheet or delimited text. The Stata 14 software is used to compute the Box-Ljung *Q* statistic (*LBQ*) at each lag separately with the autocorrelation function (*ACF*) and the partial autocorrelation function (*PACF*) at each lag as well, and their significance at the 95% confidence level (i.e. *p*-values). The software requires that the input format of the crash data be in either excel spreadsheet (i.e. *.xlsx) or in text (i.e. delimited or *.csv) and then the output is produced as excel spreadsheet or delimited text.

The *DW* test is a measure of the first order autocorrelation and it cannot be used to test for higher order temporal autocorrelation. The *DW* test statistic on a dataset of size n can be computed as (King 1981):

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} {e_t}^2} \qquad (3.1)$$

where,

DW: the Durbin-Watson statistic,

e_t: the residual error term in time period t,

e_{t-1}: the residual error term in the previous time period $t - 1$.

The Breusch-Godfrey test is a general test of temporal autocorrelation and can be used to test for first order temporal autocorrelation or higher order autocorrelation, and is based on asymptotic Chi-Square distribution χ^2. This test is a specific type of Lagrange

Multiplier (*LM*) test. The *LM* test statistic can be computed as (Thomas 1993; Studenmund 2001):

$$LM = (n - i)\, R^2 \qquad (3.2)$$

where,

LM: the Lagrange multiplier test statistic,

n: the number of observations in the dataset,

i: the order of the autocorrelation,

R^2: the unadjusted R^2 statistic (coefficient of determination) of the model.

The Ljung-Box *Q* (*LBQ*) test (sometimes called the Portmanteau test) is used to test for any order of temporal autocorrelation, and is based on asymptotic Chi-Square distribution χ^2. The test statistic provided by Box et al. (1994) is:

$$LBQ = n\,(n+2) \sum_{j=1}^{i} \frac{{r_j}^2}{n-j} \qquad (3.3)$$

where,

LBQ: the Ljung-Box *Q* statistic,

n: the number of observations in the data,

j: the lag being considered,

i: the autocorrelation order,

r: the residual error term in lag j.

The minimum recommended number of lags (*m*) that should be considered for the *LM* and *LBQ* tests is roughly taken as the natural logarithm of the number of observations (*n*) within the dataset (i.e. $m = ln\ (n)$) (Tsay 2010), and larger values are recommended to detect the existence of temporal autocorrelation.

3.2: Removal of Temporal Autocorrelation

If the temporal autocorrelation is found to be significant in crash data, then it must be removed before using the data in the modeling process (Washington et al. 2010; Lord and Mannering 2010; Savolainen et al. 2011). In order to remove any significant temporal autocorrelation that may be existed in a dataset, one of the first remedial measures should be to investigate the omission of one or more of the explanatory variables, especially variables that are related to time. In this research, since the three temporal variables in the datasets (month, weekday, hour) have potential influence on the dependent variable (i.e. crash severity), therefore they are unlikely to be removed from the analysis. Hence, the next step is to apply a differencing procedure to all time independent variables in the dataset to convert them into their differences values, and rerun an ordinary least squared regression model from the origin by deleting the intercept from the model (Chatfield 1996). Differencing can be applied by simply subtracting the previous observation from the current observation, as follows:

$$D\ (Y_t) = Y_t - Y_{t-1} \qquad (3.4)$$

where,

$D\ (Y)$: the difference of variable Y at lag t,

Y_t : the value of Y at lag t,

Y_{t-1} : the value of Y at lag $t-1$.

The rho (i.e. the residual autocorrelation coefficient) is assumed to be (1.0) in the differencing procedure, which could overestimate the true rho value (Pindyck and Rubinfeld 1981).

When the differencing procedure cannot eliminate the temporal autocorrelation in a dataset, then the Cochrane-Orcutt procedure should be applied for the Autoregressive AR (1) term of this dataset (Wooldridge 2013). The procedure uses the ordinary least square residuals to obtain the value of rho which minimizes the sum of squared residuals. Rho is then used to transform the observations of the variables. The process continues until convergence is reached (Cochrane and Orcutt 1949; Wooldridge 2013). Considering the ordinary least squared regression model:

$$Y_t = \alpha + X_t \beta + \varepsilon_t \qquad (3.5)$$

where,

Y_t : the dependent variable at time (lag) t,

α: the intercept,

β : the vector of regression coefficients,

X_t : the vector of explanatory variables at time (lag) t,

ε_t : the error term of the model at time (lag) t

If the Durbin-Watson (*DW*) test revealed that the temporal autocorrelation exists among the model error terms, then the residuals must be modeled for the first order autoregressive term AR (1) such that:

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + e_t \qquad (3.6)$$

where,

ρ : the temporal autocorrelation coefficient (rho) between pairs of observations, $0 < \rho < 1$,

e_t : the error term of the residuals at time (lag) t.

The Cochrane-Orcutt procedure is obtained by taking a quasi-differencing or generalized differencing, such that the sum of squared residuals is minimized (Cochrane and Orcutt 1949; Wooldridge 2013):

$$Y_t - \rho\, Y_{t-1} = \alpha\,(1 - \rho) + \beta\,(X_t - \rho\, X_{t-1}) + e_t \tag{3.7}$$

The Cochrane-Orcutt iterative procedure starts by obtaining parameter estimates by the ordinary least square regression (OLS). Applying equation (3.6), the OLS residuals are then used to obtain an estimate of rho from the OLS regression. This estimate of rho is then used to produce transformed observations, and parameter estimates are obtained again by applying OLS to the transformed model. A new estimate of rho is computed and another round of parameter estimates is obtained. The iterations stop when successive parameter estimates differ by less than 0.001 (Wooldridge 2013).

3.3: Evaluation of Spatial Autocorrelation

In many vehicle crash datasets, geographic relationships among crashes can exists given that movement is confined to roadways which are traversed by many users. This phenomena is termed spatial autocorrelation and if not appropriately accounted for, can lead to incorrect parameter estimates (Lord and Persaud 2000; Wood 2002; Quddus 2004; MacNab 2004; El-Basyouny and Sayed 2009; Washington et al. 2010; Lord and Mannering 2010; Savolainen et al. 2011). This dissertation examines two spatial autocorrelation indices: 1) Moran's *I*; and 2) Getis-Ord G_i*statistic to differentiate between spatially clustered, dispersed, or random crash events. Some of these spatial autocorrelation statistics have been applied in the context of traffic safety data. For example, Manepalli et al. (2011) applied the G_i* to determine the road high crash areas in Arkansas, but without using the results in crash modeling. Truong and Somenahalli (2011) applied Moran's *I* and G_i* to identify the unsafe transit bus stops in south Australia. However, such autocorrelation statistics have not been previously investigated for their utility in crash severity modeling. As such, this dissertation introduces the application of these statistics in the crash modeling process as potential risk factors. In addition, this dissertation introduces a new hybrid method to evaluate spatial autocorrelation of crashes by combining both Moran's *I* and *Gi** statistic to examine the spatial clustering pattern of crashes.

For n observations on a variable x at locations i, j, Global Moran's *I* can be calculated as follows (Anselin 1995):

$$I = \frac{n}{S_0} \frac{\sum_i^n \sum_j^n w_{ij}\,(x_i - \bar{x})(x_j - \bar{x})}{\sum_i^n (x_i - \bar{x})^2} \tag{3.8}$$

where,

$\bar{x}$: the mean of the variable x;

x_i : the value of variable x at location i;

x_j : the value of variable x at location j;

w_{ij}: the elements of the weight matrix;

n : number of observations;

S_0 : is the sum of the elements of the weight matrix: $S_0 = \sum_i^n \sum_j^n \boldsymbol{w_{ij}}$

The local Moran's I for location i can be calculated as follows:

$$I_i = \frac{x_i - \bar{x}}{S_i^2} \sum_{j=1}^{n} w_{ij} \, (x_j - \bar{x}) \qquad (3.9)$$

$$S_i^2 = \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n-1} - \bar{x}^2 \qquad (3.10)$$

The weight matrix, w_{ij}, is a contiguity matrix. If zone j is adjacent to zone $_i$, wij= 1.0. Otherwise, wij= 0.0. Cliff and Ord (1975) generalized these definitions to include any type of weight, and in more current use, w_{ij}, is a distance-based weight which is the inverse distance between locations i and j ($1/d_{ij}$). Values of this index typically, range from -1.0 to +1.0, where a value of -1.0 indicates negative spatial autocorrelation, and a value of +1.0 indicates positive spatial autocorrelation. When nearby points have similar values, their cross product is high. Conversely, when nearby points have dissimilar values, their cross-product is low.

The General (Global) *G* statistic computes a single statistic for the entire study area, while the G_i statistic is an indicator for local spatial autocorrelation for each data point. The Global *G* statistic can be calculated as follows (Fischer and Wang 2011):

$$G = \frac{\sum_i \sum_{j \neq i} w_{ij} x_i x_j}{\sum_i \sum_{j \neq i} x_i x_j} \qquad (3.11)$$

where,

x_i : the value of variable x at location i;

x_j : the value of variable x at location j;

w_{ij} : the elements of the weight matrix;

There are two types of local G_i statistics, although almost the two types produce identical results (Getis and Ord 1996; Berglund and Karlstrom 1999). The first one, G_i, does not include the autocorrelation of a zone with itself, whereas the G_i* includes the interaction of a zone with itself (i.e. the G_i statistic does not include the value of X_i itself, but only the neighborhood values, but G_i* includes X_i as well as the neighborhood values), and both can be computed by the formulae (Fischer and Wang 2011):

$$G_i\ (d) = \frac{\sum_{j\neq i}^{n} w_{ij}\, x_j}{\sum_{j\neq i}^{n} x_j} \qquad (3.12)$$

$$G_i^*\ (d) = \frac{\sum_{j=1}^{n} w_{ij}\, x_j}{\sum_{j=1}^{n} x_j} \qquad (3.13)$$

where d is the neighborhood (threshold) distance, and w_{ij} is the weight matrix that has only 1.0 or 0.0 values, 1.0 if j is within d distance of i, and 0.0 if its beyond that distance. These formulae indicate that the cross-product of the value of X at location i and at another location j is weighted by a distance weight, w_{ij} which is defined by either a 1.0 if the two locations are equal to or closer than a threshold distance, d, or a 0.0 otherwise. The G statistic can vary between 0.0 and 1.0. The significance of the local autocorrelation between each point and its neighbors is assessed by the z-score and the p-value.

Both Moran's *I* and the Getis-Ord *Gi** statistic might be adapted to use values of x_i and x_j that represent any variable in the model, such as crash severity, number of vehicles involved, speeding, and accident type. In this dissertation, both are adapted to use values of x_i and x_j that represent the crash severity. A *Gi** value is computed for each crash location and subsequently used as a potential risk factor in the crash severity model with three possible indicators: high significant spatial autocorrelation values; low significant spatial autocorrelation values; or otherwise considered as insignificant random crashes.

A GIS can be used to compute the Moran's *I*, and G_i* statistics for a set of crashes using the following process:

- Spatially join the attributes of crash incidents to road segments based on their location relationship (i.e. latitude/longitude) using functionalities of a GIS that try to parse roads up into consistent analysis units and matching the two features according to their relative spatial locations;
- Build a network of roads from the crash attributed road segments;
- Generate spatial weights matrix for the network arcs;
- Compute the Global Moran's *I* available in the ArcMap 10.2.2 Spatial Statistics toolkit;
- Compute the Global (General) *Gi* statistic available in the ArcMap 10.2.2 Spatial Statistics toolkit;
- Compute Anselin local Moran's *I* available in the ArcMap 10.2.2 Spatial Statistics toolkit;

- Compute the local Getis-Ord local *Gi** statistic available in the ArcMap 10.2.2 Spatial Statistics toolkit;

Statistically significant high spatial autocorrelation locations will have a high *z*-value and be surrounded by other crashes with high *z*-values as well (referred to as high-high (HH)). Statistically significant low spatial autocorrelation locations (referred to as low-low (LL)) will be found in cases where a crash point will have a low *z*-value and be surrounded by other crashes with low *z*-values as well. If the *z*-value of a particular crash location is higher than the mean *z*-value of all crashes, then it would be considered high. If the *z*-value of a particular crash point is lower than the mean *z*-value of all crashes, then it would be considered low. The resultant *z*-scores and *p*-values indicate whether crashes with either high or low *z*-values are clustered. A high *z*-score and small *p*-value for a crash point indicates a spatial clustering of high values (i.e. HH). A low negative *z*-score and small *p*-value indicates a spatial clustering of low values (i.e. LL). The higher (or lower) the *z*-score, the more intense the clustering. A *z*-score near zero indicates no apparent spatial clustering. Both the Anselin local Moran's *I* and the local G_i*statistic can be computed by the ArcMap 10.2.2 Spatial Statistics toolkit (ESRI 2016).

Since the Anselin Moran's *I* and the Gi* can identify relatively different clustering patterns of crashes, therefore this dissertation recommends using a combination (hybrid) of these spatial autocorrelation indices to determine the clustering patterns. Using a combination of indices can make improvements on the clustering patterns. To couple the Moran's, and the Gi* autocorrelation indices into a new hybrid method, any combination maybe used by the user depending on his/her own interpretation of the results that produces the optimal measures. For instance, a combination of 30% Moran's *I*, and 70% Gi* in representing the final spatial autocorrelation measure of crashes is used in this dissertation to examine a new spatial clustering pattern of crashes.

3.4: Evaluation of Sight Distance

In this dissertation, a GIS-based viewshed analysis is developed to assess existing stopping and decision sight distances on multilane highways. This method can be used to identify locations along roadways that likely do not conform to the AASHTO (2011) criteria regarding stopping sight distance and decision sight distance. Moreover, this approach can also be used to compute actual sight distances at or near crash incidents which could then be used as a potential risk factor in crash prediction. This method provides a new technique for estimating the available stopping and decision sight distance and also presents a new method for estimating the passing sight distance on two-lane highways, and locating the passing zones and no-passing zones along two-lane highways.

3.4.1: Viewshed Analysis

Viewshed analysis in a GIS environment typically evaluates raster-based elevation data, such as digital elevation model (DEM), to determine which cells are visible from a particular location. Each raster cell is denoted by its column and row number, relative to a reference *X* and *Y* coordinate. Each raster cell is associated with a single attribute, such as elevation in the case of a DEM. The width of the raster cells denotes the spatial resolution of the dataset. To create a viewshed for determining the visibility between an observer location and a target point on the Earth's surface, all cells along the line-of-sight (LOS) from an observer's location and a target's location must be identified. Once the raster cells along the line-of-sight have been determined, the elevation value of each cell is loaded into an array, which holds the elevation values of the terrain profile between the two points. However, the LOS does not necessarily cross each cell at its center, with the exception of the beginning and end cells. Therefore, the terrain profile may be further refined by interpolating the elevation value at the approximate location at which the LOS enters and leaves each cell. After the cells underneath the LOS are selected and the elevation values are determined in a chosen geographic coordinate system, these values can then be used to create the terrain profile needed for determining visibility of the target from the observer, and a viewshed is created (De Floriani and Magillo 1994; Fisher 1996; Wang et al., 2000; Kim et al. 2004).

3.4.2: Generating Viewsheds

In order to conduct sight distance analysis of locations along highways, DEMs are needed as well as a representation of the road network to provide information on the trajectories that vehicle must follow when traveling along a highway. DEMs are often publicly available online in a variety of formats and spatial resolutions for most areas within the U.S. In this dissertation, the following steps are used to generate viewsheds:

- Derive a set of observer points along a roadway from which sight distance will be evaluated.
    - Densification of road segments: for each segment of the roadway, vertices are added such that the distance between each vertex and the next one is not more than the AASHTO recommended sight distance,

as shown in Figure 3.1. For instance, given a road segment that is 1100 m long, with a speed of 70 mph (110 km/h) and an AASHTO sight distance of 220 m, 5 vertices (i.e. 1100/220) would be added to the line so that there is at least one vertex every 220m.

    - Convert segment vertices to points: For each road segment, all vertices are extracted and rendered as point features. The resulting points are then used as possible locations from which drivers may view the landscape while driving.

- Determine the roadway analysis region: The region around each road to be analyzed should be sized according to how it is assumed features off the roadway are expected to impact visibility. For instance, it could be assumed that features more than 200m from a road segment likely wouldn't have a large impact on sight distance, etc.
    - To obtain this analysis region, the road segments can be transformed into polygons through a buffer transformation using a GIS. For example, the segments could be buffered by 200.0 m so as to define the areas of interest.
- Extract portions of DEMs within the analysis region: Given that, DEMs can be large and present a computational burden, only those portions of the DEMs corresponding with the analysis region are retained for analysis. This can be done by clipping the DEMs by the road buffer.
- Combine the portions of the DEMs within the analysis region: In cases where more than one DEM is needed to evaluate roadway sight distance, all of the

portions of the DEMs falling within the analysis region can be combined together into a single seamless DEM termed a Mosaic that can minimize the abrupt changes along the boundaries of the overlapping rasters.

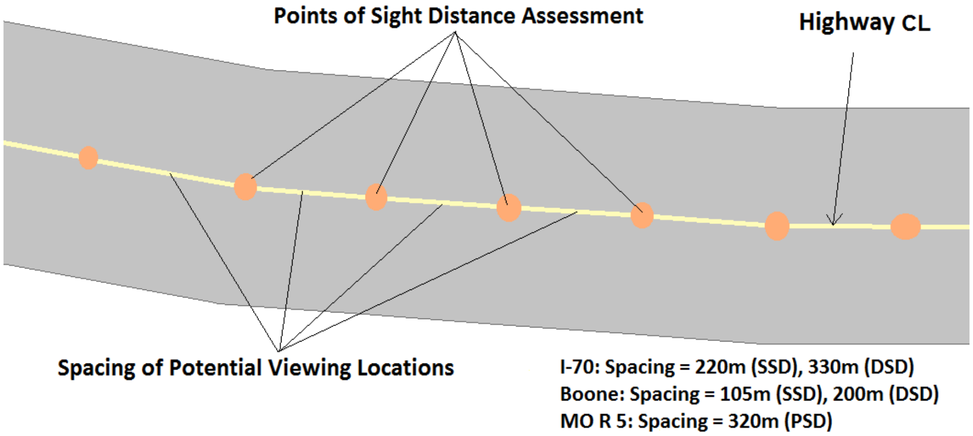
Points of Sight Distance Assessment
Highway CL
Spacing of Potential Viewing Locations
I-70: Spacing = 220m (SSD), 330m (DSD)
Boone: Spacing = 105m (SSD), 200m (DSD)
MO R 5: Spacing = 320m (PSD)

Figure 3.1: Observer points in viewshed analysis

- Creating viewsheds: Using the mosaicked DEMs, the observer points, and assuming the heights of the driver and an object on the road, viewsheds can be generated. In particular, the following parameters must be specified before creating the viewsheds, as shown in Figure 3.2 and Figure 3.3.

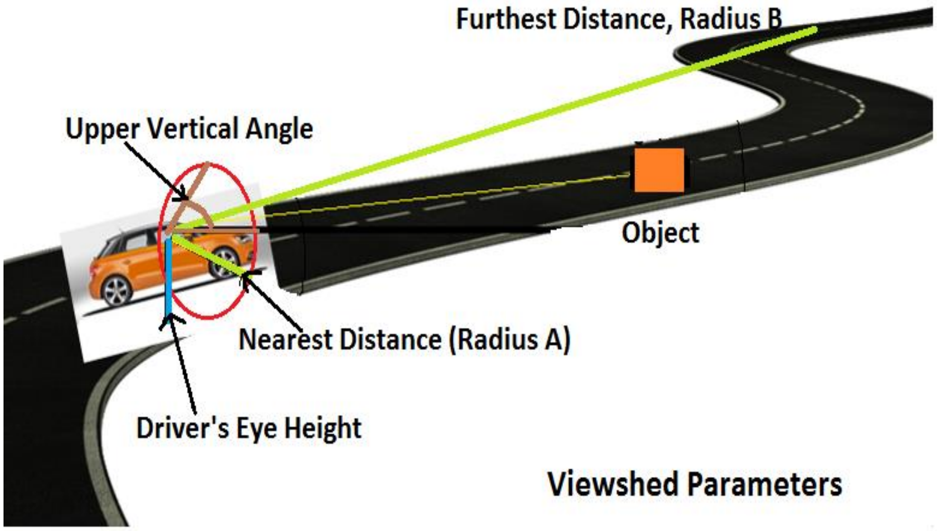

Figure 3.2: Viewshed parameters

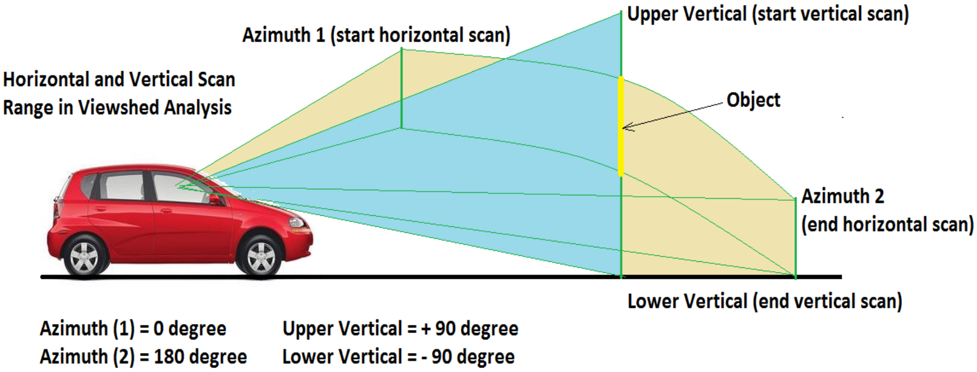
Azimuth 1 (start horizontal scan)
Upper Vertical (start vertical scan)
Horizontal and Vertical Scan
Range in Viewshed Analysis
Object
Azimuth 2
(end horizontal scan)
Lower Vertical (end vertical scan)
Azimuth (1) = 0 degree
Azimuth (2) = 180 degree
Upper Vertical = + 90 degree
Lower Vertical = - 90 degree

Figure 3.3: Viewshed horizontal and vertical scans

1. Observer Height = represents the height of the driver's eye above the road surface for each observer point. AASHTO (2011) recommends 1.08 meter (3.5 ft) for both stopping and decision sight distances. AASHTO (2011) also recommends 1.08 meter as an observer height for the passing sight distance.

2. Object Height = represents the height of a visible object above the road surface. AASHTO (2011) recommends 0.6 meter (2.0 ft) for both stopping and decision sight distances. AASHTO (2011) recommends 1.08 meter as an object height for the passing sight distance.

3. Start Azimuth = represents the start horizontal angle of the scan range for the observer. A 0.0 degree is used in this dissertation for stopping, decision, and passing sight distances.

4. End Azimuth = represents the end horizontal angle of the scan range for the observer. A 180.0 degree is used in this dissertation for stopping, decision, and passing sight distances.

5. Upper Vertical = represents the upper vertical angle of the scan for the observer. A 90.0 degree is used in this dissertation for stopping, decision, and passing sight distances.

6. Lower Vertical = represents the lower vertical angle of the scan for the observer. A negative 90.0 degree is used in this dissertation for stopping, decision, and passing sight distances.

7. Nearest Distance = represents the closest location that can be viewed by the observer. A 0.0 meter is used in this dissertation for stopping, decision, and passing sight distances.

8. Furthest Distance = represents the farthest location that can be viewed by the observer. This value could be infinity or any reasonable number that the driver's eye can see at farthest possible point. A value of 1000 meter is used for stopping, decision, and passing sight distances.

- Classifying the viewshed created in the previous step into segments that conform to AASHTO (2011) sight distance (denoted by AASHTO SD) and segments that may have visibility issues relative to AASHTO (2011) standards (denoted by NOT AASHTO SD). The conforming segments are those with available sight distances that are equal or greater than the AASHTO (2011) sight distance criteria, while the segments that may have visibility issues are those with available sight distance that are less than the AASHTO (2011) sight distance, as shown in Figure 3.4. The decision sight distance at segments with potential visibility issues was used as potential risk factor in the crash severity modeling.

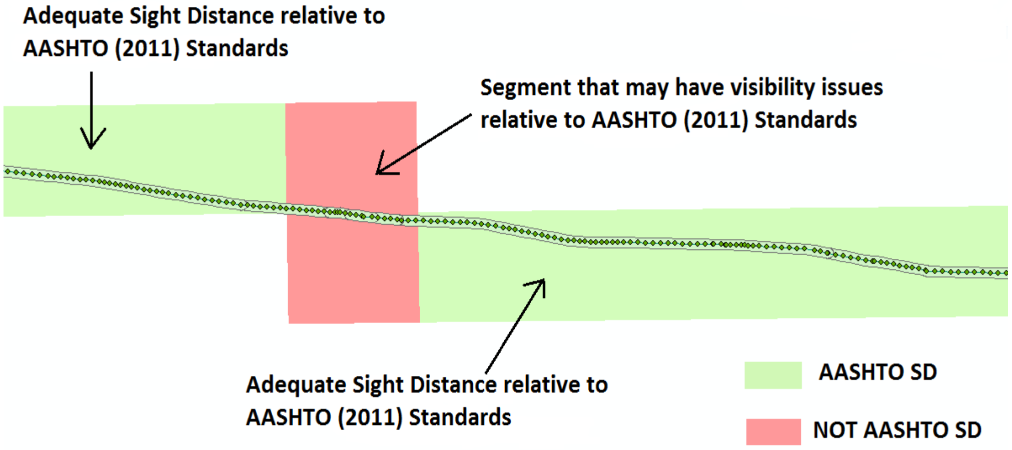
Adequate Sight Distance relative to AASHTO (2011) Standards
Segment that may have visibility issues relative to AASHTO (2011) Standards
Adequate Sight Distance relative to AASHTO (2011) Standards
AASHTO SD
NOT AASHTO SD

Figure 3.4: Classification of road segments

3.4.3: Incorporating Passing Sight Distance in Methodology

To incorporate the passing sight distance in the methodology, the viewsheds created in the previous steps should be classified into segments that conform to AASHTO (2011) passing sight distance (Passing Zones PZ) and segments that may not conform to

AASHTO passing sight distance (No-Passing Zones NPZ). PZs are those with available sight distances that are equal or greater than the AASHTO (2011) passing sight distance criteria, and the NPZs are those with available passing sight distance that are less than the AASHTO (2011) passing sight distance, as shown in Figure 3.5.

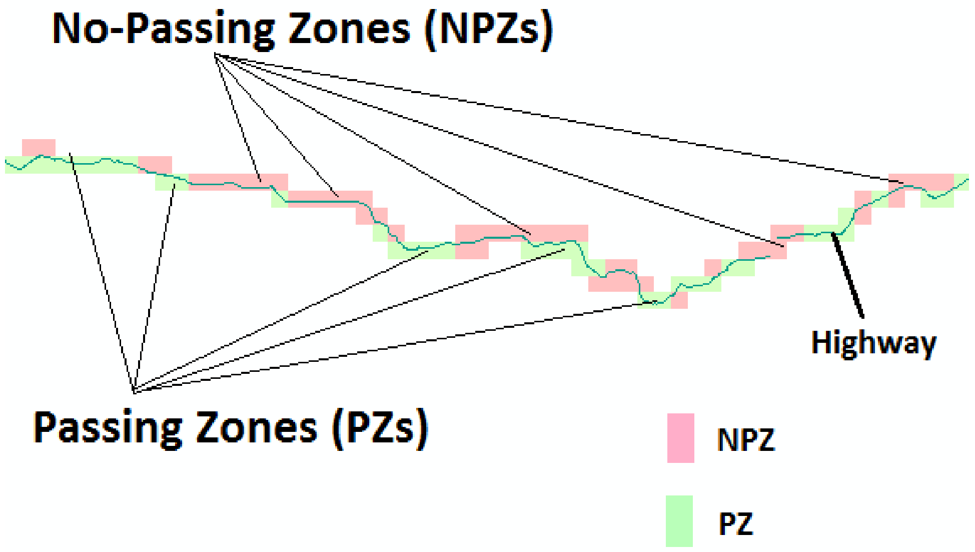

Figure 3.5: Passing and no-passing zones

3.5: Evaluation of Multinomial Logistic Regression (MNL)

Since the dependent variable in crash severity modeling (i.e. crash severity) usually has two or more outcome categories (i.e. fatal, injury, property-damage-only), therefore, logit and probit models are often used to model the severity of crash data. Discriminant analysis could also be used to model crash severity, but it assumes very restricted rules, so logit and probit are preferred due to their modeling flexibility (Washington et al. 2010; Greene 2012). Binary models consider two response outcomes (i.e. fatal vs. non-fatal or injury vs. property-damage-only), and multinomial models consider three or more response outcomes. There are many types of the multinomial models that can be used in modeling crash severity, such as, the multinomial logistic regression (MNL), the nested logistic regression, the mixed logistic regression, and the multinomial probit models, however, the MNL is the most popular and convenient model that can be used in the analysis of crash severity (Washington et al. 2010; Greene 2012). The dependent variable (i.e. crash severity) in this dissertation has four outcome categories (i.e. fatal, disabling injury, minor injury, property-damage-only), and is nominal (i.e. unordered), therefore it is modeled by the multinomial logistic regression (MNL). The MNL works by choosing one outcome category as the base (reference) category for the other categories. Then MNL contrasts all the outcomes of the dependent variable with this common reference category, which serves as the contrast point for all analyses, and the results of the analysis are always in reference to the contrast category (Greene 2012). In this dissertation, the property damage is considered as the reference group (i.e. base category), because it is the most frequent outcome of crash severity, and the other outcome levels (i.e. minor injury, disabling injury, and fatal) are estimated relative to the property damage.

There are very few applications of the MNL in crash modeling. For example, Abdel-Aty (2003) apply the ordered probit model and the ordered MNL to predict crash severity on roadway sections, signalized intersections and toll plazas by using the Florida crash database. Bham et al. (2012) apply a multinomial logistic regression to model the severity injury of different vehicle collision patterns in urban highways in Arkansas, and recommended the use of the MNL over other models. Despite these few applications of the MNL, this dissertation seeks to introduce a variety of new procedures in presenting the results of the MNL applications that have not been reported in other crash severity research. First, the use of odd ratios as regression estimates is explored to interpret the results of prediction instead of regression coefficients. Second, a greater focus is place on the assumption of the independence of irrelevant alternatives (IIA), which is very crucial in the MNL modeling, using the Hausman specification test. Third, the generalized Hosmer-Lemeshow test is used as an important goodness of fit measure to assess whether or not the observed incidents match the predicted incidents. Fourth, the concept of the classification table is evaluated as a measure of goodness of fit to determine the percent of corrected prediction cases. Next, tests for the multicollinearity among the independent variables as precondition assumption are conducted. The pseudo R square measure is used as a potential goodness of fit instead of the classical measures, such as the Deviance, the Akaike Information Criteria (AIC), and the Bayesian Information Criteria (BIC). Lastly, the marginal effects of all independent variables upon the dependent variable are presented. The following sections illustrate the assumptions of the MNL, the concept of logit functions and odd ratios, several methodological procedures that should be used in testing the assumptions of the MNL, and the MNL goodness of fit tests.

3.5.1: The Assumptions of MNL

The multinomial logistic regression uses the Maximum Likelihood Estimation (MLE) rather than the Ordinary Least Squared (OLS) estimation, therefore it avoids many of the typical assumptions tested in ordinary statistical analysis, such as the following (Greene 2012):

- Does not assume normal distribution of variables (both dependent variables DVs and independent variables IVs).
- Does not assume linearity between DV and IVs.
- Does not assume homoscedasticity (homogeneity of variances).
- Does not assume normal distribution of errors.

However, MNL does apply the following assumptions when used in the analysis:

- The dependent variable has to be categorical (i.e. it must be possible to divide the responses into different categories) but without intrinsic order (unordered).
- The independent variable may either be numerical (i.e. continuous) or categorical (i.e. discrete).
- Categories of DV must represent discrete units that are mutually exclusive and exhaustive.
- Categories of the DV must have a contrast reference/base category, otherwise, one must run all pair-wise contrasts between them.
- Large sample size must be used (not less than 30 observations).
- Multicollinearity must be checked, and is assumed to be relatively low, as it becomes difficult to differentiate between the impact of several variables if they are highly correlated.
- The assumption of independence of irrelevant alternatives (IIA) must hold. This assumption states that the odds of one class versus another do not depend on the presence or absence of other "irrelevant" alternatives.

3.5.2: The Logit Function and Odd Ratios of MNL

The MNL tries to find the best fitted model to describe the relationship between the polytomous dependent variable with more than two categories and a set of independent variables. The logistic regression model is a non-linear transformation of the linear regression model, as it consists of an S-shaped distribution function, and it's very easy to work with in most applications (Judge et al. 1985). The logit distribution constrains the estimated probabilities that lie between 0.0 and 1.0, as shown in Figure 3.6. The logistic regression function is bounded by 0.0 and 1.0, whereas the linear regression function may predict values above 1.0 and below 0.0.

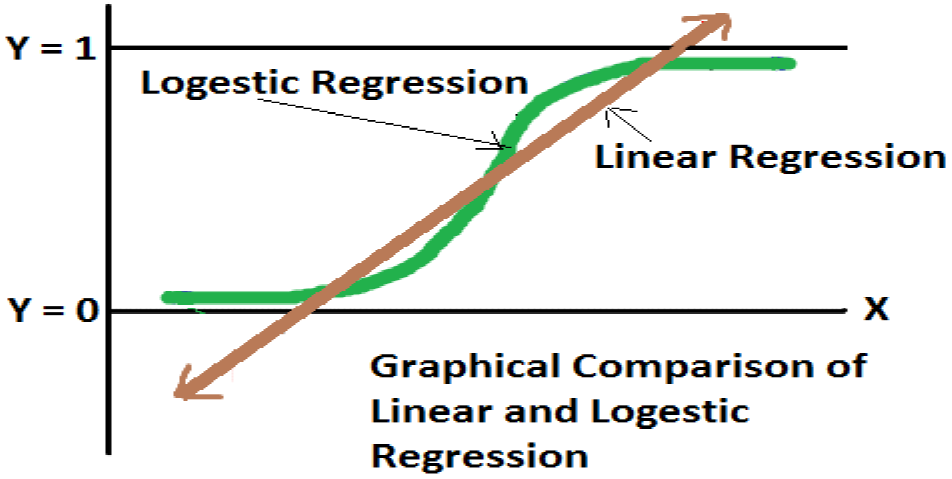

Figure 3.6: Comparison of linear and logistic regression

The logistic (logit) function can be expressed as:

$$Logit\ (p) = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k \qquad (3.14)$$

Where,

p: the probability of presence of an outcome of interest,

X_k: the vector of k independent variables,

b_0: the regression coefficient on the constant term (intercept),

b_k: the vector of regression coefficients on the independent variables X_k,

The odd ratio is the probability of the event divided by the probability of the nonevent, and is defined as follows (Judge et al. 1985; Greene 2012):

$$odd\ ratios = p\,/\,(1 - p) \qquad (3.15)$$

When p = 0, then odd (p) = 0, when p = 0.5, then odd (p) = 1.0, and when p = 1.0, then odd (p) = ∞.

The logit transformation is defined as the logged odds:

$$Logit\ (p) = ln\ [p\,/\,(1 - p)] \qquad (3.16)$$

The transformation from odds to log of odds is the log transformation, and this is a monotonic transformation. That is, the greater the odds, the greater the log of odds and vice versa.

Logit (p) can be back-transformed to p by the following formula:

$$p = \frac{1}{1+e^{-\,logit\ (p)}} \qquad (3.17)$$

The transformation from probability to odds is a monotonic transformation as well, meaning the odds increase as the probability increases or vice versa. Probability ranges from 0.0 and 1.0. Odds range from 0.0 and positive infinity (Judge et al. 1985; Baltagi 2011).

3.5.3: Maximum Likelihood Estimation (MLE)

Multinomial logistic regression uses the maximum likelihood estimation (MLE) to produce the regression parameters. Assuming that the random variables $X_1, X_2, \cdots,$ X_n form a random sample from a distribution $f\,(x\,|\theta)$; if X is continuous random variable, $f\,(x|\,\theta)$ is probability density function (pdf), if X is discrete random variable, $f\,(x|\,\theta)$ is point mass function (pmf). The distribution depends on a parameter θ, where θ could be a real unknown parameter or a vector of parameters. For every observed random sample $x_1, \cdots, x_n$, we define (Long 1996):

$$f\,(x_1, \cdots, x_n|\,\theta) = f\,(x_1|\,\theta) \cdots f\,(x_n|\,\theta) \qquad (3.18)$$

If $f\,(x|\,\theta)$ is pdf, $f\,(x1, \cdots, xn|\,\theta)$ is the joint density function; if $f\,(x|\,\theta)$ is pmf, $f\,(x1, \cdots, xn|\,\theta)$ is the joint probability. The function $f\,(x_1, \cdots, x_n|\,\theta)$ is the likelihood function, which depends on the unknown parameter θ, and it is denoted as $L(\theta)$. In order to get the maximum likelihood function, a value of θ for which the likelihood function $L(\theta)$ is a maximum is used as an estimate of θ. Maximizing $L(\theta)$ with a product of n terms is equivalent to maximizing $log\ L(\theta)$ because log is a monotonic increasing function. $log\ L(\theta)$ is a log likelihood function, and is denoted as $LL(\theta)$, as follows (Long 1996):

$$LL\ (\theta) = log\ (\theta) = log \prod_{i=1}^{n} f(Xi|\,\theta) = \sum_{i=1}^{n} f(Xi|\,\theta) \qquad (3.19)$$

3.5.4: The Effects of Independent Variables

The effect of any independent variable on the outcome can be tested using the likelihood ratio (*LR*) statistic test. If the dependent variable has *M* categories, then there are $M - 1$ non redundant coefficients (β_n) associated with each independent variable x_n.

The null hypothesis that x_n does not affect the dependent variable can be written as:

$$H0\colon \beta_{n,\, 1|\, Base} = \ldots = \beta_{n,\, M\, |\, Base} = 0 \qquad (3.20)$$

Where *Base* is the base category used in the model. The hypothesis can be tested with the *LR* test. First, the *LR* estimates the full model that contains all of the independent variables with the resulting *LR* statistic LR_F. Second, the *LR* estimates the restricted model formed by excluding the independent variable x_n with the resulting *LR* statistic LR_R. Finally, the *LR* estimates the difference between LR_F and LR_R which is distributed as chi-square with *n* degrees of freedom (the number of independent variables). The *LR* statistic is computed in terms of log likelihood (*LL*) as follows (Long 1996; Baltagi 2011):

$$LR = [-2\ LL\ (of\ full\ model)] - [-2\ LL\ (of\ restricted\ model)] \quad (3.21)$$

$$LR = LR_F - LR_R \qquad (3.22)$$

Alternatively, the null model is given by (-2 *log* (L_0)) where L_0 is the likelihood of obtaining the observations if the independent variables had no effect on the outcome (i.e. model with intercept alone). The full model is given by (-2 *log* (*L*)) where *L* is the likelihood of obtaining the observations with all independent variables incorporated in the model. The difference of these two yields a Chi-Squared statistic which is a measure of how well the independent variables affect the outcome or dependent variable (Greene 2012). If the *LR* statistic for the overall model is significant, then there is evidence that the independent variables have contributed to the prediction of the outcome.

3.5.5: The Independence of Irrelevant Alternatives (IIA)

The MNL assumes that the odd ratios for any pair of outcomes (i.e. any pair of the dependent variable categories) are determined without reference to the other categories that might be available (McFadden et al. 1976; Hausman 1978). This assumption is called the independence of irrelevant alternatives (IIA), which is very crucial in the MNL modeling. If the IIA holds, then the MNL model can be used, if the IIA does not hold, then the MNL cannot be used and alternative models should be utilized such as, the nested MNL. The IIA can be tested by the Hausman specification test, proposed by Hausman and McFadden (1984), which proceeds by estimating the error coefficients of the full model with all categories of the dependent variable included, then estimating the error coefficients of a restricted model by eliminating one or more outcome categories. The null hypothesis of the test is that the IIA does not exist and estimators of the full and restricted models are consistent, and under the alternative hypothesis the IIA does exist and only the estimators of the restricted model are consistent. The test statistic H_{IIA} is asymptotically distributed as chi square, and significant values of H_{IIA} indicate that the IIA assumption is violated (Hausman and McFadden 1984). The Hausman specification test involves the following steps:

1- Estimate the error coefficients of the full model with all *M* categories of the dependent variable included; these coefficients are contained in $\hat{E}_{f.}$

2- Estimate the error coefficients of a restricted model by eliminating one or more outcome categories; theses coefficients are contained in $\hat{E}_r$.

3- Let $\hat{E}^*_f$ represents $\hat{E}_f$ after eliminating all coefficients not estimated in the restricted model. The Hausman specification test of IIA is defined as (Hausman and McFadden 1984):

$$H_{IIA} = (\hat{E}_r - \hat{E}^*_f)'[Var(\hat{E}_r) - Var(\hat{E}^*_f)]^{-1} (\hat{E}_r - \hat{E}^*_f) \quad (3.23)$$

H_{IIA} is asymptotically distributed as chi square with degrees of freedom equal to the rows in $\hat{E}_r$. In this dissertation, the Hausman specification test will be applied on each outcome pair of the dependent variable (i.e. crash severity) separately, excluding the other category of the dependent variable. Since the property damage is assumed to be the base category, as it is the most frequent occurred category, therefore the test will be applied on the minor injury vs disabled injury first, and second; it will be applied on the minor injury vs fatal injury, and lastly; it will be applied on the disabled injury vs fatal injury. For each outcome pair, the test statistic H_{IIA} will be obtained and compared to the full model with all outcomes. If the value of H_{IIA} for any pair is significant, then the IIA assumption is violated and the MNL cannot be used in the modeling process. If the values of H_{IIA} for all pairs are insignificant, then the IIA assumption holds and the MNL can be used in the modeling process.

3.5.6: Multicollinearity

Multi-collinearity is the existence of linear relationships among the independent variables that can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, give false, non-significant *p*-values, and degrade the predictability of the model (Green 2012). The source of the multi-collinearity might come from data collection, sampling techniques, political or legal constraints, and outliers. Testing the multi-collinearity can be achieved by: (1) visual inspection of pairwise scatter plots of independent variables, and looking for near-perfect linear relationships between them; (2) Eigenvalues and Condition Indices; and (3) considering the variance inflation factors (VIF). The VIF is the most widely used test to measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. The VIF may be calculated for each predictor by doing a linear regression of that predictor on all the other predictors, and then obtaining the R^2 from that regression. The VIFs obtained by the linear regression can still be used in logistic regression models, because the concern is with the relationship among the independent variables included in the model, not with the functional form of the model (Menard 2002). Thus, a VIF of 1.6 tells us that the variance (the square of the standard error) of a particular coefficient is 60% larger than it would be if that predictor was completely uncorrelated with all other predictors.

The VIF has a lower value of 1.0 but no upper bound. As a rule of thumb, if VIF is more than 10.0, then multicollinearity is considered a serious problem, and must be corrected (Hoerl and Kennard 1970; Menard 2002; Green 2012). Variance inflation factors are scaled measures of the correlation coefficient between variable j and the rest of the independent variables. Specifically,

$$VIF_j = \frac{1}{1-{R_j}^2} \qquad (3.24)$$

where,

R^2_j: is the coefficient of determination of the regression model that includes all predictors except the jth predictor.

Variance inflation factors are often given as the reciprocal of the above formula. In this case, they are referred to as the tolerances. If R^2j equals zero (i.e. no correlation between j and the remaining independent variables), then VIF_j equals 1.0, and this is the minimum value.

3.5.7: The Generalized Hosmer-Lemeshow Statistic

The generalized Hosmer-Lemeshow test is used as an important goodness of fit measure to assess whether or not the observed events match expected events, by subgrouping the probabilities estimated from the data (Lemeshow and Hosmer 1982; Hosmer et al. 2013). The data set, of size n, is sorted according to the probabilities estimated from the final fitted MNL model. Then the data set is partitioned into several (Hosmer and Lemeshow recommend 10) equal-sized groups. The first group corresponds to the n/10 observations having the highest estimated probabilities. The next group corresponds to the n/10 observations having the next highest estimated probabilities, etc. A Pearson-like chi square statistic is constructed based on the observed and expected group frequencies. In order to get the generalized test statistic (HL), we suppose that we have a sample of n independent observations, (x_i, y_i), i = 1, . . ., n. Recoding y_i into binary indicator variables y_{ij}, such that $y_{ij} = 1$ when $y_i = j$ and $y_{ij} = 0$, otherwise (i = $1, . . ., n$ and $j = 0, . . ., c - 1$). After fitting the model, let π_{ij} denote the estimated probabilities for each observation ($i = 1, . . ., n$) for each possible outcome ($j = 0, . . ., c - 1$). By sorting the observations according to $1 - \pi_{i0}$, the complement of the estimated probability of the reference outcome. We then form g groups, each containing approximately n/g observations. For each group, we calculate the sums of the observed and estimated frequencies for each outcome category as follows (Fagerland and Hosmer 2012):

$$O_{kj} = \sum_{l \in \Omega_k} y_{lj} \qquad (3.25)$$

$$E_{kj} = \sum_{l \in \Omega_k} \pi_{lj} \qquad (3.26)$$

Where O_{kj} is the observed frequency, E_{jk} is the expected frequency, $k = 1, . . ., g$; $j = 0, . . ., c - 1$; and Ω_k denotes indices of the n/g observations in group k.

The multinomial goodness-of-fit (*HL*) test statistic is the Pearson's chi-squared statistic from the table of observed and estimated frequencies, and is given as (Fagerland and Hosmer 2012):

$$C_g = \sum_{k=1}^{g} \sum_{j=0}^{c-1} \frac{(O_{kj} - E_{kj})^2}{E_{kj}} \qquad (3.27)$$

The distribution of *Cg* is chi-squared and has $(g-2) \times (c-1)$ degrees of freedom (Fagerland et al. 2008). The null hypothesis is that the differences between the observed and predicted events are insignificant so the fitted model is correct, while the alternative hypothesis is that the differences are significant so the fitted model has deficiency and incorrect. If the test statistic *HL* is insignificant, then we will accept the null hypothesis, and conclude that the fitted model is a good fit. If the test statistic *HL* is significant, then we will reject the null hypothesis, and conclude that the data do not fit the hypothesized fitted MNL regression model.

3.5.8: The Classification Table of MNL

The classification table is another method to assess the goodness of fit of the MNL regression model. In this table the observed values for the dependent outcome and the predicted values (at a user defined cut-off value, for example p = 0.50) are cross-classified to indicate the correct % of predicted cases. This percent statistic assumes that if the estimated p is greater than or equal to 0.5 then the event is expected to occur and not occur otherwise. The bigger the % correct predictions, the better the model fit. We suppose for n observations that c (j, j') is the (j, j') th element of the classification table, $j, j' = 1, \ldots, J$. c (j, j') is the sum of the frequencies for the observations whose actual response category is j (as row) and predicted response category is j' (as column) respectively. Then, the percentage of total correct predictions of the model is given by (Kleinbaum and Klein 2010; Long and Freese 2014):

$$\%\ total\ correct\ prediction\ = \left(\frac{\sum_{j=1}^{n} c\left(j,j'\right)}{n}\right) * 100\% \quad (3.28)$$

The percentage of correct predictions for response category j is given by:

$$\%\ \text{correct}\ prediction\ of\ j\ = \left[\frac{c\ (j,j')}{\sum_{i=1}^{m} n_{ij}}\right] * 100\% \quad (3.29)$$

3.5.9: The Pseudo R-squares

In ordinary least squared (OLS) regression there is a non-pseudo R-square, which is often generated as a goodness-of-fit measure, and is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3.30)$$

where n is the number of observations in the model, y is the dependent variable, y-bar is the mean of the y values, and y-hat is the value predicted by the model. The numerator of the ratio is the sum of the squared differences between the actual y values and the predicted y values. The denominator of the ratio is the sum of squared differences between the actual y values and their mean.

When analyzing data with a multinomial logistic regression, there is no an equivalent statistic to R-squared. The estimates from a logistic regression are found by the maximum likelihood estimation rather than the least squared estimation, so the OLS approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squares have been developed. They are called "pseudo" R-squares because they are on a similar scale, ranging from 0 to 1 (though some pseudo R-squares never achieve 0 or 1) with higher values indicating better model fit, but they cannot be interpreted as one would interpret an OLS R-squared, and different pseudo R-squares can present different values (Menard 2000). Some of the popular pseudo R-squares are:

McFadden’s R-square, which is defined as (McFadden 1974):

$$R^2{}_{McF} = 1 - \frac{\ln L_M}{\ln L_o} \qquad (3.31)$$

Where L_0 is the value of the likelihood function for a model with no predictors (i.e. with intercept only), and L_M is the likelihood function for the model being estimated.

The ratio of the McFadden R-square indicates the level of improvement over the intercept model offered by the full model. Since a likelihood falls between 0.0 and 1.0, so the log of a likelihood is less than or equal to zero. If a model has a very low likelihood, then the log of the likelihood will have a larger magnitude than the log of a more likely model. Thus, a small ratio of log likelihoods indicates that the full model is a far better fit than the intercept model. When comparing two models on the same data, McFadden's would be higher for the model with the greater likelihood.

Another pseudo R-square is the Cox and Snell R^2 which is defined as (Cox and Snell 1989):

$$R^2_{C\&S} = 1 - \left(\frac{L_0}{L_M} \right)^{2/n} \qquad (3.32)$$

where n is the sample size. The Cox and Snell R-square indicates the level of improvement of the full model over the intercept model. This pseudo R-squared has a maximum value that is less than 1.0 when the full model predicts the outcome perfectly and has a likelihood of 1.0. The Nagelkerke R-square adjusts Cox & Snell's so that the range of possible values extends to 1.0 by dividing by its maximum possible value, $(1-L_0)^{2/n}$. If the full model perfectly predicts the outcome and has a likelihood of 1.0, then the Nagelkerke R-square = 1.0, which is defined as (Nagelkerke 1991):

$$R^2_{NK} = \frac{1-(\frac{L_0}{L_M})^{2/n}}{1-(L_0\ )^{2/n}} \qquad (3.33)$$

Pseudo R-squares are useful tools in evaluating multiple models predicting the same outcome on the same dataset, but they cannot be interpreted independently or compared across different datasets. In other words, a pseudo R-squared statistic without context has little meaning. A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome (Menard 2002; Tjur 2009). In this case, the higher pseudo R-squared indicates which model better predicts the outcome.

3.5.10: Estimation of Marginal Effects

Marginal effects are useful estimates of the impact of a one-unit change of an independent variable (predictor) on the dependent variable. The average marginal effects are interpreted as the effect of a one-unit change in an independent variable (keeping all other independent variables constant at their mean values) on dependent variable. It is common to use a single average marginal effect value for all observations of an independent variable. Elasticity analysis can also be used to interpret the effect of a specific independent variable on the dependent variable, but with a 1.0% change instead of a one-unit change. In MNL, the marginal effect of an explanatory variable (predictor) is the partial derivative of the event probability with respect to the predictor of interest (i.e. the change in the event probability for a unit change in the predictor). The marginal effect for a dummy independent variable is the difference of the predicted probability values at their different levels (Long and Freese 2014). The values of the marginal effects reflect the slopes of lines tangent to each of the predictors that is drawn tangent to the fitted probability curve at the selected point. The slope of the tangent line is the change in event probability, p, measured at two points one unit apart along this straight line. If the probability curve is linear (near p=0.5) at the selected point, then the marginal effect will approximate the probability change when changing the predictor by one unit. If the probability curve is nonlinear (near the smallest and largest values of p), the marginal effect might deviate from the change (Kleinbaum and Klein 2010; Long and Freese 2014). For multinomial logistic regression models, the possible response values are unordered with levels i=1, 2, ..., k. The probability of response level i is given by (Freese and Long 2000):

$$p_i = \frac{EXP\ (X'\beta i)}{\sum_j (EXP\ (X'\beta i))} \qquad (3.34)$$

where $X'$is the predictor of interest, and β_i is the regression coefficient (i.e. log

odd) of X'. The marginal effect of the jth predictor, X_j, on p_i is given by:

$$\frac{\partial p_i}{\partial X_j} = pi\ [\frac{\partial X'\beta i}{\partial X_j} - \sum_k (p_k \frac{\partial X'\beta k}{\partial X_j})] \qquad (3.35)$$

CHAPTER 4: APPLICATIONS OF MISSOURI CRASH DATA

4.1: Missouri Crash Data

To better illustrate the analysis framework presented in Chapter 3, Missouri crash data as reported by the Missouri State Highway Patrol (MSHP) and recorded in the Missouri Statewide Traffic Accident Records System (STARS) are analyzed. STARS is the primary source of crash data in the State of Missouri since 1978. The STARS crash data (in MS Excel format) can be publicly obtained from an online query portal at the MSHP website (MSHP 2016). In the state of Missouri, law enforcement agencies are required to investigate traffic crashes on public roadways if they involve a death or personal injury or property damage over $500.00, and to submit a Missouri Uniform Traffic Crash Report (MUCR) to STARS. Once the MUCR is approved by the Missouri State Highway Patrol, it will become a part of STARS data. In this application, three years of crash data (2013-2015) are considered. Each record in this database, is attributed with the number of persons killed and injured, property damage, latitude and longitude of each crash location, number of vehicles involved in the crash, type of the vehicle involved, speed involvement, alcohol involvement, the accident type, construction zone involvement, driver’s aggressiveness involvement, texting and cell phone involvement, light conditions, and the driver’s age.

4.2: Crash Data Terminology

STARS uses the following terms differentiate among crash types (MSHP 2016):

- Alcohol involved traffic crashes: any crash in which one or more drivers or pedestrians were drinking and, which in the opinion of the investigating officer, their intoxicated condition contributed to the cause of the crash.

- Speed involved traffic crashes: any crash where a contributing circumstance was either exceeding the speed limit or going too fast for conditions. Too fast for conditions indicates a vehicle's speed was too fast for conditions at the time of the crash which includes road, weather, and other conditions, and in the opinion of the investigating law enforcement officer, the driver error contributed to the cause of the crash.
- Aggressive driving: any driver of a motorized vehicle who has one or more of the following contributing circumstances: exceeding the speed limit, driving too fast for conditions, improper passing, violating a signal or sign, following too closely, improperly using a signal, improper lane usage or lane change, and failing to yield.
- Construction/Work Zone: an area of a road where construction, maintenance, or utility work activities are identified by warning signs, signals, or indicators on transport devices that mark the beginning and the end of a work activity. Work zones also include roadway sections where there is a moving work activity such as lane line painting or marking or roadside mowing if the beginning of the work activity is designated by warning signs or signals.

4.3: Missouri GIS Road Data

Missouri GIS road data was obtained from the Missouri Spatial Data Information Service (MSDIS). This road data includes the annual average daily traffic (AADT), the travel way direction, the county name, the city name, the road geometry (i.e. grade/level and number of lanes), the speed limits, and the road classification (MSDIS 2016).

4.4: Crash Severity Categories

There are different injury severity scales in the US. The most popular scale is the (KABCO) proposed by the National Safety Council (NSC) in 1990 and frequently used by law enforcement for classifying injuries. This acronym lists the five levels of injury- from most severe to least severe: K, killed (fatal); A, disabling injury or incapacitating injury; B, evident injury or non-incapacitating injury; C, possible injury; and O, no apparent injury or property damage only. Other types of injury severity data may include detailed information on trauma location and extent of injury that uses the Abbreviated Injury Scale (AIS), as proposed by the American Association for Automotive Medicine (Savolainen et al. 2011). AIS scores describe the severity of injury on a scale of 0.0 (no injury) to 6.0 (unsurvivable). Another scale is the Injury Severity Score (ISS) used by many hospitals, which is a measure of overall injury severity calculated by summing the squares of the AIS scores for each of the three most severely injured ISS body regions (head/neck, face, chest, abdomen and pelvic contents, extremities or pelvic girdle, and external), and its score ranges from 1.0 to 75.0 (Baker et al. 1974). In the US, each state uses different severity scale in reporting crashes. In the state of Missouri, the STARS data includes only four severity injury categories (i.e. property damage, minor injury, disabled injury, and fatal). As such, crash severity (i.e. the dependent variable) is modeled in this dissertation using the following four STARS severity categories:

- Property-Damage-Only: A property damage crash that includes any crash in which no person was killed or injured but property was damaged in the incident.

- Minor Injury: An injury crash in which one or more persons received an evident injury but not disabling in the incident.
- Disabled Injury: An injury crash in which one or more persons received a disabling in the incident.
- Fatal: A fatal crash includes any crash in which one or more persons were killed and their death occurred within 30 days of the incident.

If a crash result in more than one injury severity category, then the most severe category would be considered for reporting. For instance, if a crash resulted in fatal, and property damage, then this crash would be reported as fatal (MSHP 2016).

4.5: I-70 Corridor and Boone County Roads in Missouri

This dissertation models crash severity for two types of transportation corridors: 1) Interstate I-70 in Missouri (Figure 4.1), and 2) roadways in Boone County Missouri (MO) (Figure 4.2). The I-70 corridor in MO is a multi-lane divided highway that traverses the State of Missouri west to east with a total length of 403 km (250 mile). Boone County, MO has a total area of 1280.0 square km (496.0 square miles), and its county seat is Columbia, the fourth-largest city in Missouri and the home of the University of Missouri. The STARS and roadway data were carefully examined, labelled, filtered, and outliers and missing data were excluded from the analysis. The total numbers of the observed crashes within the three years 2013-2015 are 5869.0 along the I-70 corridor and 2348.0 along roads in Boone County, as shown in Table 4.1. A summary of the crash data by the level of severity is shown in Table 4.2.

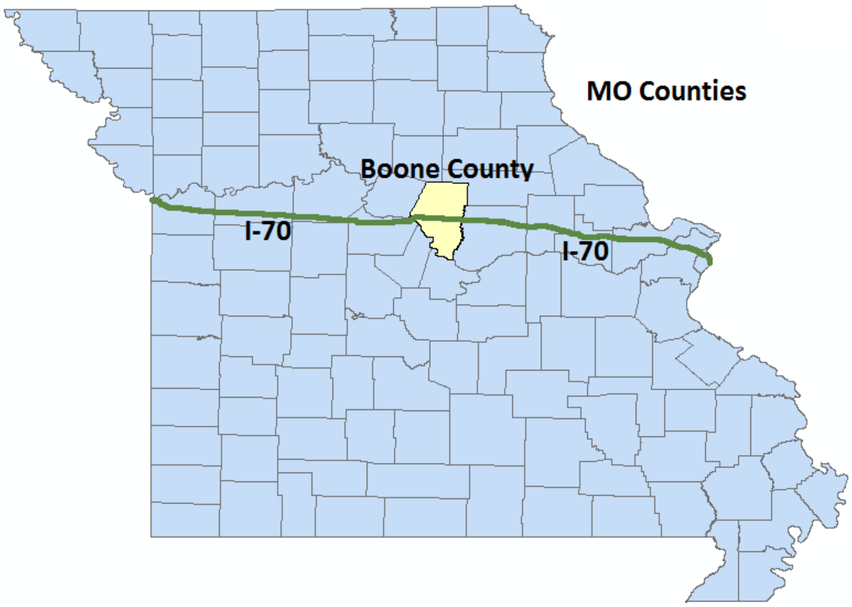

Figure 4.1: Interstate I-70 and Boone County in Missouri

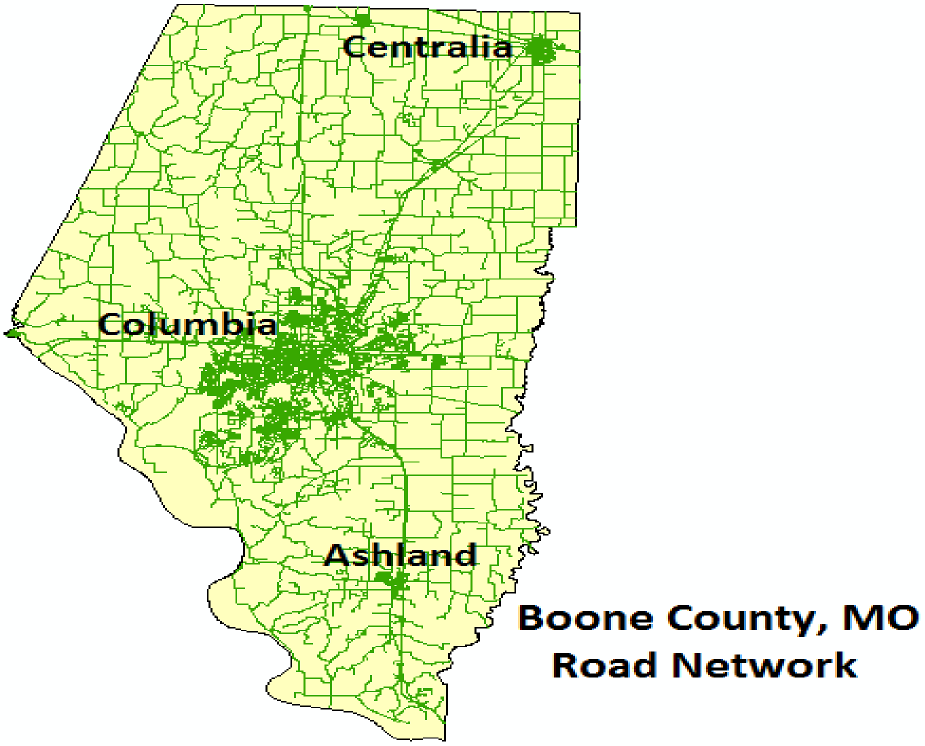

Figure 4.2: Boone County roads and major cities

Table 4.1: Observed crashes along I-70 and roads in Boone County, MO

Year	2013	2014	2015	Total
Interstate I-70, MO	1918	2009	1942	5869
Boone County roads, MO	794	811	743	2348

Table 4.2: Observed crashes by severity level

Severity Level	2013		2014		2015		Total	
	I-70	Boone	I-70	Boone	I-70	Boone	I-70	Boone
Property Damage Only	1479	547	1605	603	1474	530	4558	1680
Minor Injury	371	213	340	173	397	178	1108	564
Disabling Injury	63	31	46	28	57	30	166	89
Fatal Injury	5	3	18	7	14	5	37	15

The STARS system provides the latitude and longitude coordinates of each reported crash, rather than reporting the crash characteristics by road segment as is done by reporting agencies in other states. The latitude and longitude of each crash can be used to generate point features in a GIS so that crashes can be compared relative to other geographic features. All crashes that occurred within the boundaries of intersections are excluded from the analysis given the AASHTO (2011) sight distance used in this dissertation as a risk factor does not apply to the intersection boundaries, and therefore only crashes outside of intersections are considered. Given this criteria, 2164.0 crashes were excluded and 5869.0 retained for the I-70 datasets, 837.0 crashes were excluded and 2348.0 retained for Boone County dataset.

4.6: MO Route-5 Highway for Assessing the PSD

Since the passing sight distance (PSD) is only applied in practice on two-lane highways, MO Route-5 is used to assess the PSD as another study site given that the other two datasets include multilane and/or urban roadways. MO Route 5 is the longest two-lane highway in Missouri with a total length of 571 km (355 mile) that traverses the entire state from north to south, as shown in Figure 4.3. The GIS MO Route 5 data was obtained from the Missouri Spatial Data Information Service (MSDIS).

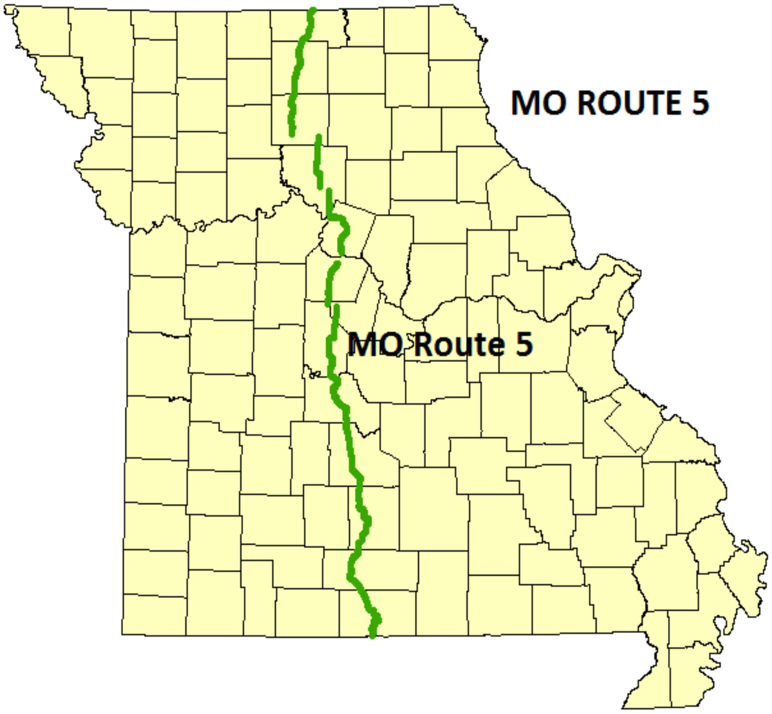

Figure 4.3: MO Route 5

4.7: The Digital Elevation Models (DEMs)

The Digital Elevation Models (DEMs) were used to assess the stopping and decision sight distances along the I-70 corridor and Boone roads. In addition, the DEMs were also used to assess the passing sight distance and locating the passing and no-passing zones along the MO Route 5. All DEMs were acquired from the Missouri Spatial Data Information Service (MSDIS) for the sight distance analysis. The resolution of the DEMs was 30 m x 30 m (with 2,100 row by 2,100 column). There were eleven DEMs that spanned the I-70 corridor, one DEM that covered Boone County, and thirteen DEMs that covered the length of MO Route 5 as shown in Table 4.3. The number of observer points generated for decision sight distance (using a maximum point spacing of 330.0 m) along I-70 corridor was 1247.0, and for Boone County (using a maximum point spacing of 200.0 m) was 9407.0. The number of observer points generated for passing sight distance (using a maximum point spacing of 320 m) along MO Route 5 was 2104.0.

Table 4.3: DEMs used in dissertation

I-70 Corridor, MO		MO Route 5	
MSDIS DEM's Code	County	MSDIS DEM's Code	County
jackson. e00.gz	Jackson	ozark. e00.gz	Ozark
lafayet. e00.gz	Lafayette	douglas. e00.gz	Douglas
saline. e00.gz	Saline	wright. e00.gz	Wright
cooper. e00.gz	Cooper	laclede. e00.gz	Laclede
boone. e00.gz	Boone	camden. e00.gz	Camden
callaway. e00.gz	Callaway	morgan. e00.gz	Morgan
montgom. e00.gz	Montgomery	moniteau. e00.gz	Moniteau
warren. e00.gz	Warren	cooper. e00.gz	Cooper
stchar. e00.gz	Saint Charles	howard. e00.gz	Howard
stlouis. e00.gz	Saint Louis	chariton. e00.gz	Chariton
stlcity. e00.gz	City of St. Louis	linn. e00.gz	Linn
-	-	sullivan. e00.gz	Sullivan
-	-	putnam. e00.gz	Putnam

4.8: Partitioning the Crash Data into Training and Testing

In order to assist in the evaluation of the performance of the crash prediction techniques to be applied later, the STARS crash data were first partitioned into training and testing datasets. The training dataset will be used to develop the prediction model, and the testing dataset will be used to evaluate or test the developed model against the observed data. In line with other analyses of crash data (Cameron and Trivedi 1998; Chang 2005; El-Basyouny and Sayed 2009), the STARS data for the entire period (2013-2015) was randomly partitioned into two parts, a training dataset that contains 70% of the observations, and a testing dataset that contains 30% of the observations. The training dataset includes 4,108 observed crashes for I-70 corridor and 1,761 for Boone County road network. The testing dataset includes 1,644 observed crashes for I-70 corridor and 704 for Boone County road network, as shown in Table 4.4.

Table 4.4: Training and testing datasets

Data type	% observations	Observed crashes (2013-2015)	
		I-70	Boone County
Entire dataset	100	5869	2348
Training dataset	70	4108	1644
Testing dataset	30	1761	704

4.9: Selection of Independent Variables

The occurrence of crashes and their degrees of severity can be attributed to different risk factors associated with road geometry, traffic operations, vehicle types, driver factors, and the environment. In general, focusing on a few independent variables, and leaving out some other important variables in crash modeling could generate simplified models that can produce incorrect parameter estimates and inferences (Arminger et al. 1995; Glenberg 1996; Lord and Persaud 2000; El-Basyouny and Sayed 2006; Caliendo et al. 2007; Geedipally et al. 2012). For example, Persaud and Dzbik (1993) consider only average annual daily traffic volume in modeling road crash frequency, and later added the road geometry variables to produce a more advanced model (Persaud et al. 2000). Caliendo et al. (2007) consider limited number of risk factors related to road geometry, traffic flow, weather, and road surface in modeling crash occurrence. Abdel-Aty (2003) consider only driver age and sex along with the road geometry, and traffic volumes in crash modeling.

Given that past research has only made use of limited numbers/types of independent variables, this dissertation seeks to investigate the use of a wide range of independent variables (i.e. risk factors) for estimating the parameters and inferences. The following group factors are included in the analysis:

- Road geometry (grade or level; number of lanes)
- Road classification (rural or urban; existing of construction zones)
- Environment (light conditions)
- Traffic operation (annual average daily traffic, AADT)
- Driver factors (driver's age; speeding; aggressive driving; driver intoxicated conditions; the use of cell phone or texting)
- Vehicle type (passenger car; motorcycles; truck)
- Number of vehicles involved in the crash
- Time factors (hour of crash occurrence; weekday; month)
- Accident type (animal; fixed object; overturn; pedestrian; vehicle in transport).

In addition, this dissertation explores integration of local spatial autocorrelation (Gi* statistic) of each crash occurrence and the AASHTO (2011) recommended sight distance as potential risk factors in models of crash severity prediction. Table 4.5 lists the group factors used in the analysis while Table 4.6 lists the risk factors included in the analysis, their interpretations, and the variable indicators.

Table 4.5: Analysis group factors

Group factors	Variables included in the analysis
Road Geometry	1 - grade or level 2 - number of lanes
Road Classification	1 - rural or urban 2 - construction zones
Time	1 - hour 2 - weekday 3 - month
Environment	1 - light conditions
Driver behavior	1 - driver age 2 - speeding 3 - aggressive driving 4 – driver's intoxicated condition 5 – use of cell phone or texting 6 - number of vehicles involved
Vehicle	1 - type of vehicle involved in the crash
Traffic operation	1 - AADT 2 - direction of travel
Accident type	1 - the type of accident occurred
Spatial autocorrelation index	1 - Gi*
Sight distance	1 - AASHTO Decision sight distance

Table 4.6: Risk factors and their interpretation

Variable name	Interpretation	Variable indicators
GRADE_LEVEL	The geometric condition of the road location where the crash occurred	0 - level 1 - grade
NO_LANES	The number of road lanes per each direction where the crash occurred	1 - one lane 2 - two lanes 3 - three lanes 4 - four lanes 5 - five lanes 6 - six lanes or more

RURAL_URBAN	The area classification where the crash occurred	0 - rural 1 - urban
CZONE	The ongoing work zone activity of the location where the crash occurred	0 - yes 1 - no
LOC_G_i*	The local spatial auto-correlation G_i* *z*-score of the crash occurred	0 - high-high (HH) 1 - low-low (LL) 2 - random
SIGHT_DIST	The conformity of road location where the crash occurred to the recommended AASHTO decision sight distance (DSD)	0 - conform to AASHTO DSD 1 - does not conform to AASHTO DSD
AADT	The Annual Average Daily Traffic of the location where the crash occurred	Numeric values in 1000s of vehicles.
HOUR	The hour at which the crash occurred	Values from 0.0 am to 23.0 pm
DAY_WEEK	The week-day on which the crash occurred	1 - Sun 2 - Mon 3 - Tues 4 - Wed 5 - Thurs 6 - Fri 7 - Sat
MONTH	The month of the year in which the crash occurred	Values from 1 to 12
LIGHT_COND	The light condition at the time of the crash occurrence	0 - Daylight 1 - Dark, lighted 2 - Dark, unlighted
DR_AGE	The age of the driver of the vehicle involved in the crash	0 - < 21 years 1 - (21 to 64) years 2 - > 64 years
VEH_TYPE	The type or body of the vehicle involved in the crash	0 - passenger car 1 - motorcycle 2 - truck
NO_VEHICLE	The total number of vehicles involved in the crash occurrence	1 - one vehicle 2 - two vehicles 3 - three vehicles 4 - four vehicles 5 - five vehicles 6 – six or more vehicles

ACC_TYPE	The type or the main cause of the crash occurrence	1 - animal 2 - fixed object 3 - overturn 4 - pedestrian 5 – vehicle in transport
DR_DRINK	The driver has intoxicated condition contributed to the cause of the crash.	0 - yes 1 - no
SPEED	Exceeding the speed limit of the road section at which the crash occurred	0 - yes 1 - no
DR_AGRESSIVE	Aggressive driving due to one or more of the following conditions: improper passing, violating a sign, following too closely, improperly using a signal, improper lane change, and failing to yield	0 - yes 1 - no
CELL_TEXT	The use of cell phone or texting by the driver at the time of the crash occurrence	0 - yes 1 - no

4.10: Applications of Temporal Autocorrelation (TA)

The Durbin Watson (*DW)* test is applied to the I-70 corridor and Boone County roads at two temporal levels; aggregation by year, and aggregation over all three years. Data for each year in aggregate is separately tested using (month, weekday, and hour) as the independent temporal variables, and then the aggregate three-year period is tested using the same independent variables.

The Breusch-Godfrey (*LM)* test is applied to the I-70 corridor and Boone County roads for the first 36 lags at two temporal levels; aggregation by year, and aggregation over all three years. Data for each year in aggregate is separately tested using (month, weekday, and hour) as the independent temporal variables, and then the aggregate three-year period is tested using the same independent variables. The *LM* test is applied with degrees of freedom equal to the number of lags (i.e. 36 degrees of freedom). The minimum recommended number of lags that should be considered for the *LM* and *LBQ* tests is roughly taken as the natural logarithm of the number of observations within the dataset (Tsay 2010), and larger values are recommended to detect the existence of temporal autocorrelation. For I-70 corridor, the number of observations of the aggregated three years (2013-2015) is 5869, and the minimum recommended number of lags = ln (5869) = 8.7. For Boone County roads, the number of observations of the aggregated three years (2013-2015) is 2348, and the minimum recommended number of lags = ln (2348) = 7.7. This dissertation uses 36 lags in both the *LM* and *LBQ* tests instead of the minimum recommended number.

The Box-Ljung *Q* statistic (*LBQ*) is applied to the I-70 corridor and Boone County roads for the aggregated three-year period (2013-2015) using the time independent variables (month, weekday, and hour) and for the first 36 lags. In addition, correlograms of the autocorrelation function (ACF) and partial autocorrelation function (PACF) for the I-70 corridor and the Boone County roads for the aggregated three-year period (2013-2015) are presented.

4.10.1: The Significant TA of I-70 (2014) Dataset

Both the Durbin-Watson and the Breusch-Godfrey tests indicated the existence of significant temporal autocorrelation within the I-70 (2014) dataset. For the I-70 (2013) and I-70 (2015) the temporal autocorrelation was not significant. For the Boone County roads, the temporal autocorrelation was not significant for all three years (2013, 2014, 2015). The complete results are presented in Chapter 5.

4.11: Applications of Spatial Autocorrelation

Crashes along MO I-70 corridor and along the roads in Boone County, MO are analyzed to assess whether they are spatially clustered, dispersed, or random. First, the aggregated level of the crash data for the three years' period (2013 – 2015) was analyzed.

Second, the data for only one year (2015) was analyzed. The one-year analysis was conducted in order to discover any hidden effects that might exist within the three years' level regarding the spatial autocorrelation.

The Global Moran's *I* and the Global (General) *Gi**for the entire I-70 corridor and the entire Boone County roads, were first calculated at both the aggregated three years' level (2013-2015) and the one-year level (2015). The Global Moran's *I* and the General *Gi** evaluate whether the overall highway crashes are clustered, dispersed, or random, and assesses the overall pattern and trend of the data. The ArcMap 10.2.2 Spatial Statistics toolkit was applied to compute the Global Moran's *I*, and the General *Gi** for the two datasets. For determining the Global Moran's *I* and the General *Gi** for I-70 corridor, the layer of spatial join of crash incidents to road segments of I-70 (2013-2015) was used as the input feature class, the crash severity was used as the input field, and the inverse distance was chosen as the conceptualization method for the spatial relationships. Likewise, for determining the Global Moran's *I* and the General *Gi** for Boone County roads, the layer of spatial join of crash incidents to road segments of Boone roads (2013-2015) was used as the input feature class, the crash severity was used as the input field, and the inverse distance was employed as the conceptualization method for the spatial relationships. The significant high and low spatial autocorrelation crashes, and outliers are identified using the Anselin Local Moran's *I*, and the local G_i* statistic. The *z*-scores and *p*-values are used to evaluate the statistical significance of the computed values. In order to determine the Anselin local Moran's *I* and the local *Gi** for the I-70 corridor, the layer of spatial join of crash incidents to road segments was used as the input feature class, the crash severity was used as the input field, the Euclidean distance, and the inverse distance was employed as the conceptualization method for the spatial relationships. Likewise, for determining the Anselin local Moran's *I* and the local $Gi*$ for Boone County roads, the layer of spatial join of crash incidents to road segments was used as the input feature class, the crash severity was used as the input field, the Euclidean distance, and the inverse distance was employed as the conceptualization method for the spatial relationships.

Since the $Gi*$ statistic has identified a larger number of the significant (HHs) and significant (LLs) compared to the Moran's *I*, for both the I-70 corridor and the Boone County roads, therefore the G_i* statistic (i.e. the G_i*'s z-scores) was used in modeling the crash severity in this dissertation. In addition, the clustering pattern of both the three-year period (2013-2015) and the one-year period (2015) was almost identical for both the I-70 corridor and the Boone County roads, which implies that the three years' (2013-2015) level has no hidden or unobserved effects of spatial clustering, therefore, the three years' dataset was chosen to be included in the analysis, and each crash point was assigned a G_i*'s z-score spatial autocorrelation value at the three years' level.

Since the Anselin local Moran's *I* has identified different clustering patterns than the local $Gi*$ statistic for both the I-70 corridor and the Boone County roads, therefore this dissertation recommends using a combination (hybrid) of these methods in hot spot analysis. Using a combination of indices can improve the clustering patterns. To couple the Moran's, and $Gi*$ autocorrelation indices into a new hybrid method, any combination maybe used to depend on the user's interpretation of the results that produces the optimal measures. For instance, a combination of 30% Moran's *I*, and 70% G_i* in determining the final spatial autocorrelation measure of crashes is presented in this dissertation to show a new spatial clustering pattern of crashes.

4.12: Applications of Sight Distance

To model the stopping sight distance (SSD) along a road using viewsheds, observation locations must first be generated along roads in the study areas. In this application, vertices are added to each road segment along the I-70 corridor such that each vertex is not more than 220.0 m from the next, and vertices are added to roads in Boone County such that no vertex is more than 105 m apart. Once the vertices have been added to the road segments, they are then extracted as point features from which visibility can be evaluated. The 220.0 m distance represents the recommended AASHTO stopping sight distance, which corresponds to the mostly assigned speed limit of 110 km/h (70 mph) at the I-70 in MO. The 105.0 m distance represents the recommended AASHTO stopping sight distance, which corresponds to the average assumed speed limit of 70 km/h (45 mph) at the Boone County roads in MO.

To model the decision sight distance (DSD) along a road using viewsheds, vertices are added to road segments along the I-70 corridor such that vertices are not more than 330.0 m from the next, while vertices are added to roads in Boone County such that they are not further than 200m apart. Once the vertices have been added to the road segments, they are then extracted as point features from which visibility can be evaluated. The 330.0 m distance represents the recommended AASHTO decision sight distance (Avoidance Maneuver C on rural highways), which corresponds to the speed limit of 110 km/h (70 mph) at the I-70 in MO. The 200.0 m distance represents the recommended AASHTO decision sight distance (Avoidance Maneuver C on rural highways), which corresponds to the average assumed speed limit of 70 km/h (45 mph) at the Boone County roads in MO.

The viewshed analysis revealed that the available stopping sight distance (SSD) conforms to the AASHTO (2011) standards throughout the I-70 corridor in MO. The SSD was equal to or more than the 220.0 m sight distance. However, the viewshed analysis also revealed that some segments at the I-70 do not conform to the AASHTO (2011) decision sight distance standards of 330.0 m, and they may have visibility issues relative to AASHTO (2011) standards. The decision sight distance at these segments was used as potential risk factor in the crash severity modeling.

Likewise, the viewshed analysis revealed that the available stopping sight distance throughout Boone County roads conforms to the AASHTO (2011) standards. An average speed limit of 70 km/h (45 mph) was assumed for roads in this area, which would yield a corresponding AASHTO (2011) stopping sight distance of 105 m. The SSD was equal to or more than the 105.0 m sight distance throughout the Boone roads. However, the viewshed analysis also revealed that some segments of the Boone roads do not conform to the AASHTO standards of 200.0 m decision sight distance that represents the (Avoidance Maneuver C on rural highways), and they may have visibility issues relative to AASHTO (2011) standards. The decision sight distance at these segments was used as a potential risk factor in the crash severity modeling.

Since the passing sight distance (PSD) is only applied in practice to two-lane highways, therefore, MO Route-5 was examined to evaluate the AASHTO (2011) PSD criteria. To incorporate PSD in the methodology, observer points must be generated from which visibility along the roadway can be evaluated. To accomplish this, vertices are added to MO Route 5 segments such that each vertex is no more than 320.0 m from the next vertex. This distance represents the recommended AASHTO (2011) passing sight distance, which corresponds to an average speed limit of 100 km/h (60 mph) along MO Route 5, as shown in Figure 4.4.

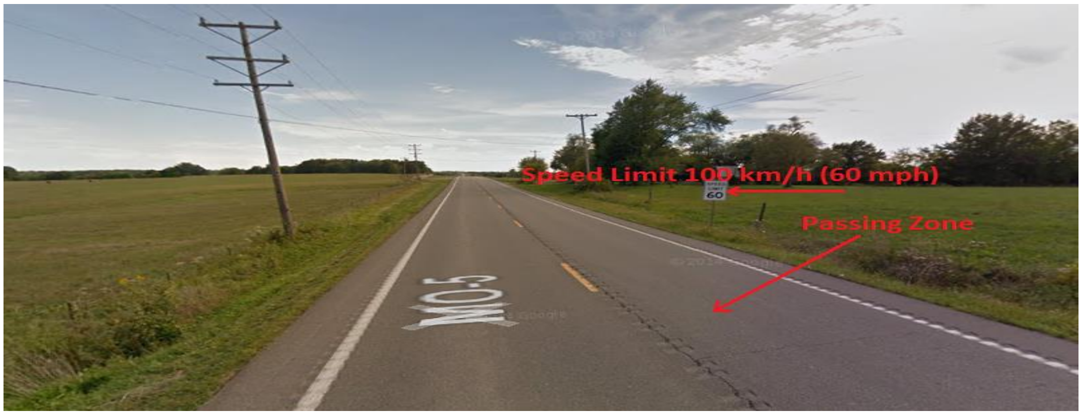

Figure 4.4: Speed limit of MO Route 5

The viewshed analysis resulted in two classifications of segments regarding the PSD: 1) segments having passing sight distance that conform to the AASHTO (2011) PSD standards throughout MO-5 (i.e. Passing zones PZs); and 2) segments that might not conform to the AASHTO PSD standards and may have visibility issues (i.e. No-passing zones NPZs).

4.13: Applications of Multinomial Logistic Regression

This dissertation applies multinomial logistic regression (MNL) to model the relationships of the crash severity categories with the independent variables. The I-70 corridor and Boone County crash datasets are tested under the assumptions of the MNL. The categories of the dependent variable in this dissertation (i.e. fatal, disabling injury, minor injury, property-damage-only) is considered nominal (i.e. cannot be ordered in any logical way). This research seeks to investigate the use of a wider range of independent variables (i.e. risk factors) in crash severity modeling, given that past research has only made use of limited numbers/types of independent variables. In addition, this dissertation seeks to introduce a variety of new procedures in presenting the results of the MNL applications that have not been reported in other crash severity models, including: 1) the use of the odd ratios as regression estimates instead of using regression coefficients to interpret the results of prediction; 2) a focus on the assumption of the independence of irrelevant alternatives (IIA) that is very important in the MNL modeling, using the Hausman specification test; 3) consideration of the generalized Hosmer-Lemeshow test as an important goodness of fit measure to assess whether or not the observed incidents match the predicted incidents; 4) use of the classification table as a measure of goodness of fit to determine the percent of corrected prediction cases; 5) testing for the multicollinearity among the independent variables as precondition assumption; 6) se of the pseudo R squares as potential goodness of fits instead of classical measures of goodness of fit, such as the Deviance, the Akaike Information Criteria (AIC), and the Bayesian Information Criteria (BIC); and 7) presenting the marginal effects of all independent variables upon the dependent variable. The next sections illustrate the testing procedures that were applied to both the I-70 and Boone County datasets.

4.13.1: Testing the Effects of Independent Variables

Multinomial logistic regression (MNL) is usually conducted using maximum likelihood estimation, which is an iterative procedure. The first iteration (called iteration zero) is the log likelihood of the null or empty model; that is, a model with no predictors. At the next iteration, the predictors are included in the model. At each iteration, the log likelihood decreases as the goal is to minimize the log likelihood. When the difference between successive iterations is very small, the model is said to have converged, the iterating stops, and the final log likelihood (*LR*) statistic is computed. The log likelihood ration (*LR*) test statistic is obtained for the I-70 corridor and the Boone County road network for both the training and testing data, using the Stata 14 software package and reported in Table 4.7.

Table 4.7: The LR statistic results

Dataset	# Observations	LR statistic	p-value
I-70 Training data	4108	339.12	0.0000
I-70 Testing data	1761	122.44	0.0000
Boone Training data	1644	125.03	0.0000
Boone Testing data	704	89.74	0.0061

The effect of any independent variable on the outcome can be tested using the likelihood ratio (*LR*) statistic test. The null hypothesis of this test is that the independent variables do not affect the dependent variable. The null model is calculated by obtaining the log likelihood of the observations with just the response variable in the model from iteration zero (i.e. model with intercept alone). The final fitted model is calculated by obtaining the log likelihood of observations with all the independent variables in the model from the final iteration after convergence. The difference of these two yields a chi-squared *LR* statistic which is a measure of how well the independent variables affect the outcomes or dependent variable categories (Greene 2012). If the *LR* statistic for the overall model is significant, then there is evidence that the independent variables are effective and they have contributed to the prediction of the outcome. Table 4.7 shows that the Likelihood Ratio (*LR*) test statistic for both the I-70 corridor and Boone County datasets is significant at the 95% confidence level with *p*-values less than 0.05 for the training and testing datasets, implying that all the independent variables included in the models are not equal to zero, and this indicates that they are effectively contributing to modeling the crash severity for all categories. Thus, it can be concluded that the overall chosen models for the I-70 corridor and Boone County data are good fits.

4.13.2: Testing the IIA Assumption

The Independence of Irrelevant Alternatives (IIA) assumption in multinomial logistic regression means that adding or deleting alternative outcome categories does not affect the odd ratios among the remaining outcomes (McFadden et al. 1976; Hausman 1978). The Hausman specification test is used to test the IIA assumption for both the I-70 dataset and the Boone County dataset (both training and testing datasets). The results of this test are shown in Table 4.8, as computed using the Stata 14 software package.

Table 4.8: The IIA Assumption results

Dataset	Minor Injury vs. Disabled		Minor Injury vs. Fatal		Disabled vs. Fatal	
	H_{IIA}	p-value	H_{IIA}	p-value	H_{IIA}	p-value
I-70 Training	1.46	0.5461	1.39	0.6725	1.73	0.7748
I-70 Testing	1.08	0.6726	1.14	0.7453	1.24	0.6833
Boone Training	1.27	0.4655	1.53	0.4973	1.31	0.5376
Boone Testing	0.72	0.4994	0.83	0.5274	0.55	0.4775

The null hypothesis of the test is that the IIA does not exist and under the alternative hypothesis the IIA does exist. The Hausman specification test statistic H_{IIA} is asymptotically distributed as chi square, and significant values of H_{IIA} indicate that the IIA assumption is violated (Hausman and McFadden 1984). The Hausman specification test was run on each outcome pair of the dependent variable (i.e. crash severity) separately, excluding the other category of the dependent variable. The base category was assumed to be the records were property damage was reported. First, the test was run on the second vs the third categories (i.e. minor injury vs disabled), second; it was run on the second vs the fourth categories (i.e. minor injury vs fatal), and lastly; it was run on the third vs the fourth categories (i.e. disabled vs fatal). Table 4.7 shows that for all cases the H_{IIA} statistic was insignificant at the 95% confidence level with their *p*-values greater than 0.05 for both the I-70 corridor and the Boone County datasets. Therefore, the null hypothesis can be accepted and it can be concluded that the IIA assumption has not been violated so that the odd ratios of any outcome pair of the dependent variable are determined without reference to the other category.

4.13.3: Testing the Generalized Hosmer-Lemeshow Statistic

The generalized Hosmer-Lemeshow statistic assesses whether or not the observed events match the predicted events, by subgrouping the probabilities estimated from the data (Lemeshow and Hosmer 1982; Hosmer et al. 2013). This test works by sorting the data according to the probabilities estimated from the final fitted MNL model. Then the sorted dataset is partitioned into several equal-sized groups. Then, the *HL* test statistic that follows a chi-square distribution is constructed based on the observed and predicted group frequencies. The null hypothesis is that the differences between the observed and predicted events are insignificant so the fitted model is correct, while the alternative hypothesis is that the differences are significant so the fitted model has deficiency and incorrect. If the test statistic *HL* is insignificant, then we will accept the null hypothesis, and conclude that the fitted model is a good fit. If the test statistic *HL* is significant, then we will reject the null hypothesis, and conclude that the data do not fit the hypothesized fitted MNL regression model. The generalized Hosmer-Lemeshow test is applied to both the I-70 dataset and the Boone County dataset (both training and testing datasets) with ten groups for each dataset. This test was again conducted using the Stata 14 software package and the results of this test are summarized in Table 4. 9.

Table 4.9: The Generalized Hosmer-Lemeshow test results

Dataset	# Observations	# Groups	HL statistic	p-value
I-70 Training	4108	10	27.406	0.286
I-70 Testing	1761	10	27.134	0.298
Boone Training	1644	10	20.384	0.675
Boone Testing	704	10	19.743	0.752

Table 4.9 shows that the *HL* test statistic for both the I-70 corridor and Boone County road network is insignificant at the 95% confidence level with *p*-values larger than 0.05 for the training and testing datasets. Therefore, the null hypothesis cannot be rejected and it can be concluded that the overall models of I-70 corridor and Boone County road network are good fit, and there is a good match between the predicted events and the observed events for all categories of the dependent variable.

4.13.4: Testing the Multicollinearity

Multicollinearity occurs when two or more predictors in the model are highly correlated that can create inaccurate estimates of the regression coefficients, and inflate the standard errors. The MNL model requires that multicollinearity be low between predictors in the model. To test for this assumption, the variance inflation factor (VIF) is used to detect multicollinearity among all predictors in our MNL logistic regression models, as it is the most widely used test for multicollinearity (Greene 2008). The VIF measures how much the variance of the estimated regression coefficients is inflated as compared to when the predictors are not linearly related. The VIF may be calculated for each predictor by doing a linear regression of that predictor on all the other predictors. The VIFs obtained by the linear regression can still be used in logistic regression models, because the concern is with the relationship among the independent variables included in the model, not with the functional form of the model (Menard 2002). The VIF has a lower value of 1.0 but no upper bound. As a rule of thumb, if VIF is more than 10.0, then multicollinearity is considered a serious problem, and must be corrected (Hoerl and Kennard 1970; Menard 2002; Green 2008). The VIF statistic is obtained for the I-70 corridor and the Boone County road network data using the Stata 14 and the results are reported in Table 4.10.

Table 4.10: VIF results

Independent Variable	VIF	
	I-70 corridor	Boone County road network
MONTH	1.023	1.044
DAY_WEEK	1.013	1.009
HOUR	1.026	1.054
NO_VEHICLE	2.099	2.339
DIRECTION	6.397	1.060
LIGHT_COND	1.113	1.258
ACC_TYPE	2.264	2.635
DR_DRINK	1.046	1.131
SPEED	1.408	1.406
CZONE	1.072	1.014
DR_AGGRESSIVE	1.373	1.372
CELL_TEXT	1.008	1.008
DR_AGE	1.015	1.035
VEH_TYPE	1.044	1.082
RURAL_URBAN	2.455	1.515
NUMBER_LANES	3.504	1.600
AADT	4.896	1.653
GRADE_LEVEL	6.457	1.016
SIGHT_DIST	1.054	1.044
Gi *	1.218	1.289

The VIFs of all the independent variables are considerably less than 10.0 for both the I-70 and Boone County datasets as can be seen from Table 4.10. The VIFs of the independent variables (Direction and Grade-Level) of the I-70 dataset are 6.397 and 6.457 respectively, but they are still less than 10.0. The VIFs of the other predictors are even less than 5.0. Based on this, it can be concluded that multicollinearity is not a serious problem in both datasets, and this implies that the assumption of low multicollinearity is achieved in the MLN model.

4.13.5: The Classification Table

The classification table is used to assess the goodness of fit of the MNL regression model. In this table the observed values for the dependent outcomes and the predicted values (at a user defined cut-off value) are cross-classified to indicate the correct % of predicted cases. This percent statistic assumes that if the predicted probability is greater than or equal to the (cut-off value) then the event is expected to occur and not occur otherwise. The bigger the % correct predictions, the better the model fit. The classification tables for the I-70 corridor dataset and for the Boone County road network (for both training and testing data) are obtained using the SPSS 23 and the results are detailed in Tables 4.11 and 4.12 respectively.

Table 4.11: I-70 classification table results

Severity Categories	I-70 Training Data			I-70 Testing Data		
	# obs.	% correct	Overall % correct	# obs.	% correct	Overall % correct
Property Damage	3186	99.6%	92.2%	1372	97.3%	91.5%
Minor Injury	785	65.4%		323	69.8%	
Disabled	114	72.8%		52	76.2%	
Fatal	23	77.1%		14	83.6%	

Table 4.12: Boone roads classification table results

Severity Categories	Boone County Training Data			Boone County Testing Data		
	# obs.	% correct	Overall % correct	# obs.	% correct	Overall % correct
Property Damage	1175	95.4%	86.5%	505	98.2%	91.6%
Minor Injury	397	62.9%		167	73.6%	
Disabled	61	68.3%		28	79.4%	
Fatal	11	81.8%		4	86.9%	

Table 4.12 shows how many cases are correctly predicted for each category of the dependent variable. For example, for the I-70 training data, there are 3,168 observed incidents involving property damage and the percent correctly predicted is 99.6%, 785 observed incidents involving minor injury with 65.4% correctly predicted, 114 observed incidents involving disabled with 72.8% correctly predicted, and 23 observed incidents involving fatal crashes and the percent correctly predicted is 77.1%. The overall percentage gives the overall percent of cases that are correctly predicted by the full model, which is 92.2% for the I-70 training data and 91.5% for testing data. This overall percentage is an important goodness-of-fit measure that indicates how well the data have fitted the full model. Likewise, for the Boone County road network, the overall percentage of the training data is 86.5% and 91.6% for testing data. These overall percentages of correctly predicted cases demonstrate that our MNL models are good fit, confirming the results obtained by the generalized Hosmer-Lemeshow test statistic that there is a good match between the predicted events and the observed events for all categories of the dependent variable.

4.13.6: The Pseudo R-squares

Multinomial logistic regression does not have an equivalent to the R-squared that is found in ordinary least square regression; however, there are some pseudo-R-square statistics that have been developed for MNL. The McFadden R-square treats the log likelihood of the intercept model as a total sum of squares, and the log likelihood of the full model as the sum of squared errors, the Cox and Snell's R-square reflects the improvement of the full model over the intercept model through the ratio of log likelihood, and the Nagelkerke R-square try to adjust the Cox and Snell's so that the range of possible values extends to 1.0. Pseudo R-squares are generally useful tools in evaluating multiple models predicting the same outcome on the same dataset, but they cannot be interpreted independently or compared across different datasets (Menard 2002; Tjur 2009). In this case, the higher pseudo R-squared indicates which model better predicts the outcome. Three types of pseudo R-squares (McFadden’s, Cox and Snell’s, and Nagelkerke’s) are obtained for the I-70 corridor and the Boone County road network (both training and testing datasets), using SPSS 23, as shown in Table 4.13. First, these pseudo R-squares are applied to the intercept only model for each dataset, and then they are applied to the full model with all predictors to capture any improvement in the fitted full model.

Table 4.13: The pseudo R-squares results

Pseudo R-square	I-70 Training		I-70 Testing		Boone Training		Boone Testing	
	Intercept	Full	Intercept	Full	Intercept	Full	Intercept	Full
McFadden	0.025	0.118	0.028	0.138	0.062	0.128	0.066	0.211
Cox - Snell	0.031	0.123	0.047	0.147	0.076	0.172	0.052	0.256
Nagelkerke	0.046	0.132	0.054	0.166	0.085	0.223	0.068	0.332

The improvement of the full model over the intercept model through the three types of pseudo R-squares is clear for both the training and testing datasets of I-70 and Boone County road network. For example, the McFadden R-square value for the I-70 training dataset is increased from 0.025 for the intercept to 0.118 for the full model, the Cox and Snell R-square value is increased from 0.031 for the intercept to 0.123 for the full model, and the Nagelkerke R-square is also increased from 0.046 for the intercept to 0.132 for the full mode. Likewise, the three R-squares are increased in their values from the intercept to the full model for the I-70 testing dataset, the Boone training and testing datasets. The higher pseudo R-squared values for the full models compared to the intercept models indicate that the fitted full models better predict the outcomes of the dependent variable, and the predictors are effective in modeling the different outcomes of the crash severity.

CHAPTER 5: RESULTS

This dissertation focuses on modeling four outcome categories of severity injury of vehicular crash data, namely; property-damage-only, minor-injury, disabled-injury, and fatal injury. In this chapter, the results of detecting the temporal autocorrelation among the independent variables that are related to the time by the *DW*, *LM*, and *LBQ* tests are first presented, followed by the correction results of the differencing, and the Cochrane-Orcutt procedure that were applied to the (2014) I-70 dataset. Next, the spatial autocorrelation results are presented using the Moran's *I*, the *Gi**. Next, the results of the available decision sight distance of the roadway compared to AASHTO (2011) criteria are presented, and roadway segments with inadequate sight distance that do not conform to the AASHTO (2011) standards are identified and then incorporated in the modeling process as risk factors. Next, the results of the passing sight distance on the MO Route 5 highway are presented. Lastly, the odd ratios, the risk factors that contributed to the crash severity, and marginal effects of the multinomial logistic regression are revealed.

5.1: Impacts of Temporal Autocorrelation

5.1.1: DW Test Results

Table 5.1, shows the results of the Durbin-Watson (*DW*) test for the I-70 and the Boone County datasets at the one-year aggregate level. It can be seen that the temporal autocorrelation of the I-70 dataset for the year 2013 is found to be 3.64% with *p* value of 0.0512 (non-significant at alpha of 0.05); for the year 2014 year is found to be 7.19% with *p*-value of 0.0002 (significant at alpha of 0.01); and for the year 2015 is found to be 2.38% with *p*-value of 0.1371 (non-significant at alpha of 0.05). The temporal autocorrelation of the Boone County dataset for the year 2013 is found to be -1.18% with *p*-value of 0.5518 (non-significant at alpha of 0.05); for the year 2014 year is found to be 0.15% with *p*-value of 0.7437 (non-significant at alpha of 0.05); and for the year 2015 is found to be -1.17% with *p*-value of 0.5896 (non-significant at alpha of 0.05). So, the only significant temporal autocorrelation is existed within the I-70 (2014) data, which should be removed before using this dataset in the modeling process.

Table 5.1: *DW* statistic for I-70 and Boone County roads

Year	Durbin-Watson		Autocorrelation		P-value		Decision	
	I-70	Boone	I-70	Boone	I-70	Boone	I-70	Boone
2013	1.927	2.013	0.0364	-0.0118	0.0512	0.5518	non-sig	non-sig
2014	1.843	1.997	0.0719	0.0015	0.0002	0.7437	sig.	non-sig
2015	1.952	2.021	0.0238	-0.0117	0.1371	0.5896	non-sig	non-sig

5.1.2: LM Test Results

Table 5.2, shows the results of the *LM* test for the I-70 and the Boone County datasets at the one-year aggregate level. The *LM* value (using 36 lags or 36 degrees of freedom) of the I-70 dataset for the year 2013 is found to be 31.022 with *p*-value of 0.7042 (non-significant at alpha of 0.05); for the year 2014 is found to be 60.129 with *p*-value of 0.0071 (significant at alpha of 0.01); and for the year 2015 is found to be 50.876 with *p*-value of 0.0512 (non-significant at alpha of 0.05). The *LM* value (using 36 lags or 36 degrees of freedom) of the Boone County dataset for the year 2013 is found to be 30.016 with *p*-value of 0.7482 (non-significant at alpha of 0.05); for the year 2014 year is found to be 40.229 with *p*-value of 0.2884 (non-significant at alpha of 0.05); and for the year 2015 is found to be 22,289 with *p*-value of 0.9642 (non-significant at alpha of 0.05). The results of the *LM* test confirms the results of the *DW* test that the I-70 dataset for the year 2014 contains a significant temporal autocorrelation as shown in Table 5.2.

Table 5.2: LM statistic for I-70 and Boone County roads

Year	LM statistic		p-value		Decision	
	I-70	Boone	I-70	Boone	I-70	Boone
2013	31.022	30.016	0.7042	0.7482	non-sig	non-sig
2014	60.129	40.229	0.0071	0.2884	Sig.	non-sig
2015	50.876	22.289	0.0672	0.9642	non-sig	non-sig

5.1.3: Differencing Results

Since a significant temporal autocorrelation is found to be existed within the I-70 (2014) data, then this should be removed before using this dataset in the modeling process (Washington et al. 2010; Lord and Mannering 2010; Savolainen et al. 2011). In order to remove any significant temporal autocorrelation that may be existed in a dataset, one of the first remedial measures should be to investigate the omission of one or more of the explanatory variables, especially variables that are related to time. In this dissertation, the three time variables in the datasets (month, weekday, hour) have potential influence on the dependent variable (i.e. crash severity), therefore they are unlikely to be removed from the analysis. Hence, the next step is to apply a differencing procedure to all time independent variables in the dataset to convert them into their differences values, and then rerun an ordinary least squared regression model from the origin by deleting the intercept from the model (Chatfield 1996). The first order differencing is applied to the I-70 (2014) dataset, and the ordinary least square residuals were obtained, then the Durbin-Watson (*DW*) test is calculated to check for the temporal autocorrelation. The result of the *DW* statistic showed that the temporal autocorrelation was still existed even after applying the first order differencing. Although the first order differencing is enough to show whether the differencing procedure can be used to remove the serial (temporal) correlation or not (Pindyck and Rubinfeld 1981), however, more differencing orders (up to 7 orders) are applied to the I-70 (2014) dataset, and the Durbin-Watson test (*DW* statistic) is calculated each time to check for the temporal autocorrelation. The results showed that the temporal autocorrelation was not removed by this method. Table 5.3 shows seven differencing orders that were applied to the data and their *DW* statistics.

Table 5.3: Differencing results for 2014 I-70 data

Difference order	DW statistic	Auto correlation	p-value	Decision
D1	1.841	0.0731	0.0002	sig.
D2	1.833	0.0724	0.0001	sig.
D3	1.831	0.0722	0.0001	sig.
D4	1.823	0.0812	0.0001	sig.
D5	1.821	0.0822	0.0001	sig.
D6	1.829	0.0781	0.0001	sig.
D7	1.820	0.0825	0.0001	sig.

5.1.4: Cochrane-Orcutt Results

Since the differencing procedure was not effective in eliminating the temporal autocorrelation from the I-70 dataset of the year 2014, then it was necessary to apply the Cochrane-Orcutt procedure for the Autoregressive AR (1) term. The iterative Cochrane-Orcutt procedure was applied to the I-70 (2014) dataset, and an optimized rho (i.e. the residual autocorrelation coefficient) value of 0.07333 was obtained using the Stata 14 software that minimizes the estimated sum of squared residuals (ESS), then the *DW* statistic was calculated for the transformed residuals. The results showed that the temporal autocorrelation was removed from the I-70 (2014) dataset, as shown in Table

5.4. The *DW* statistic for the I-70 (2014) dataset is changed after applying the Cochrane-Orcutt procedure from 1.843 (with a significant *p*-value of 0.0002) to 1.992 (with a non-significant *p*-value of 0.7167).

Table 5.4: Cochrane-Orcutt results for 2014 I-70 data

Iteration #	rho	ESS	DW	p-value	Decision
1	0.07295	568.242	1.992	0.7167	non-sig
2	0.07333	568.241			
3	0.07333	568.241			

After removing the temporal autocorrelation from the I-70 (2014) dataset, the *DW* test and the *LM* test were applied for the three years' period (2013-2015) for both the I-70 and Boone County datasets. The *DW* statistic for the three years' period (2013-2015) is 1.971 with temporal autocorrelation of 1.47% for the I-70 dataset and 2.017 with temporal autocorrelation of (- 0.87%) for the Boone County dataset respectively, both of which were non-significant, as shown in Table 5.5.

Table 5.5: Overall *DW* statistic for I-70 and Boone County roads

Year	Durbin-Watson		Autocorrelation		p-value		Decision	
	I-70	Boone	I-70	Boone	I-70	Boone	I-70	Boone
2013-2015	1.971	2.017	0.0147	-0.0087	0.1289	0.6601	non-sig	non-sig

The *LM* value for the three years' period (2013-2015) using 36 lags is 41.203 for the I-70 dataset and 34.352 for the Boone County dataset, both of which were non-significant, as shown in Table 5.6. The results from the *DW* test and the *LM* test indicate that there is no significant temporal autocorrelation between each of the temporal independent variables (i.e. month, weekday, and hour) and the dependent variable (i.e. crash severity) in the (2013-2015) dataset.

Table 5.6: Overall LM statistic for I-70 and Boone County roads

Year	LM statistic		p-value		Decision	
	I-70	Boone	I-70	Boone	I-70	Boone
2013-2015	41.203	34.352	0.2534	0.5471	non-sig	non-sig

5.1.5: The LBQ Test Results

The Box-Ljung *Q* statistic (*LBQ*) is applied to the aggregated three-year period (2013-2015). Table 5.7 shows the Box-Ljung *Q* statistic, the auto correlation function (*ACF)* and the partial autocorrelation function *(PACF*) with their *p*-values for the I-70 dataset for the first 36 lags. Table 5.8 shows the Box-Ljung *Q* statistic, the auto correlation function (*ACF)* and the partial autocorrelation function *(PACF*) with their p-values for the Boone County dataset for the first 36 lags. The LBQ statistic, the *ACF*, and the PACF for all 36 lags were non-significant for both the I-70 and the Boone County datasets. The *LBQ* statistic increases with the lag progress, indicating no temporal autocorrelation within these two datasets, confirming the results of the *DW* test and the *LM* test. Figure 5.1 and Figure 5.2 show correlograms of ACF and PACF for the I-70 dataset and the Boone County datasets for the three years' period (2013-2015) respectively.

Table 5.7: LBQ test results for I-70

Lag #	ACF	PACF	LBQ-Statistic	p-value
1	0.015	0.015	1.2720	0.259
2	-0.009	-0.009	1.7093	0.425
3	-0.024	-0.024	5.1985	0.158
4	0.021	0.021	7.7212	0.102
5	-0.006	-0.007	7.9130	0.161
6	-0.013	-0.013	8.9711	0.175
7	0.016	0.018	10.564	0.159

8	0.018	0.017	12.576	0.127
9	0.001	0.001	12.588	0.182
10	-0.002	-0.000	12.608	0.246
11	-0.001	-0.001	12.612	0.319
12	-0.013	-0.013	13.555	0.330
13	0.011	0.012	14.215	0.359
14	-0.007	-0.007	14.469	0.415
15	0.008	0.008	14.876	0.460
16	-0.022	-0.022	17.683	0.343
17	0.006	0.006	17.875	0.397
18	0.003	0.003	17.937	0.460
19	-0.001	-0.002	17.946	0.526
20	0.002	0.003	17.963	0.590
21	0.003	0.003	18.011	0.648
22	0.012	0.011	18.804	0.657
23	-0.010	-0.010	19.441	0.675
24	-0.018	-0.017	21.297	0.621
25	-0.025	-0.024	24.926	0.467
26	-0.019	-0.020	27.163	0.401
27	-0.017	-0.017	28.857	0.368
28	-0.005	-0.006	29.012	0.412
29	-0.005	-0.005	29.160	0.457
30	0.011	0.010	29.869	0.472
31	-0.006	-0.005	30.071	0.514
32	-0.028	-0.028	34.843	0.334
33	0.002	0.005	34.877	0.379
34	0.029	0.030	39.955	0.223
35	0.018	0.016	41.843	0.198
36	0.000	0.002	41.843	0.232

Table 5.8: LBQ test results for Boone County roads

Lag #	ACF	PACF	LBQ-Statistic	p-value
1	-0.009	-0.009	0.1767	0.674
2	-0.031	-0.031	2.3798	0.304
3	-0.020	-0.021	3.3379	0.342
4	0.026	0.025	4.9545	0.292
5	0.048	0.048	10.918	0.053
6	0.038	0.041	14.380	0.076
7	-0.007	-0.002	14.496	0.063
8	-0.015	-0.012	15.035	0.058
9	0.020	0.019	16.016	0.067
10	-0.022	-0.027	17.163	0.071
11	0.012	0.008	17.486	0.094
12	-0.019	-0.020	18.328	0.106
13	0.032	0.032	20.742	0.078
14	0.007	0.007	20.861	0.105
15	-0.007	-0.006	20.991	0.137
16	-0.041	-0.038	24.918	0.071

17	-0.018	-0.019	25.664	0.081
18	-0.009	-0.015	25.852	0.103
19	0.009	0.004	26.030	0.129
20	0.015	0.015	26.550	0.148
21	-0.003	0.005	26.576	0.185
22	-0.026	-0.022	28.231	0.168
23	-0.002	0.001	28.242	0.207
24	0.005	0.001	28.305	0.247
25	-0.008	-0.010	28.468	0.287
26	-0.008	-0.010	28.617	0.329
27	0.011	0.013	28.902	0.366
28	0.009	0.009	29.077	0.409
29	-0.030	-0.028	31.241	0.354
30	-0.032	-0.030	33.654	0.295
31	-0.007	-0.008	33.787	0.334
32	-0.004	-0.011	33.830	0.379
33	0.012	0.008	34.174	0.411
34	-0.010	-0.009	34.436	0.447
35	-0.017	-0.009	35.143	0.461
36	-0.021	-0.018	36.226	0.458

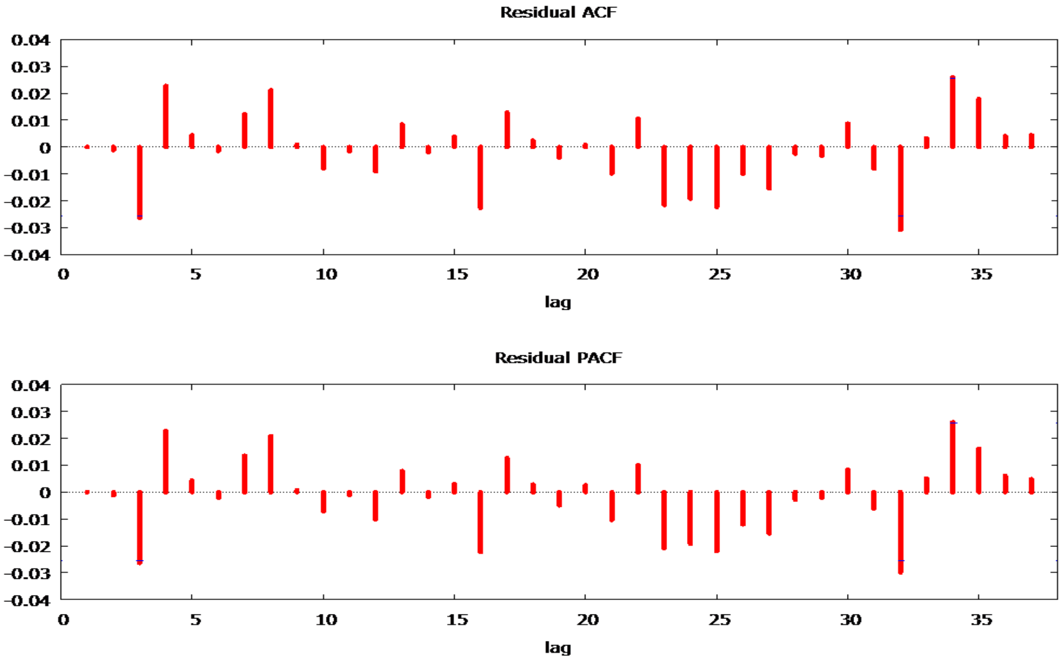

Figure 5.1: Correlogram of I-70 (2013-2015) dataset

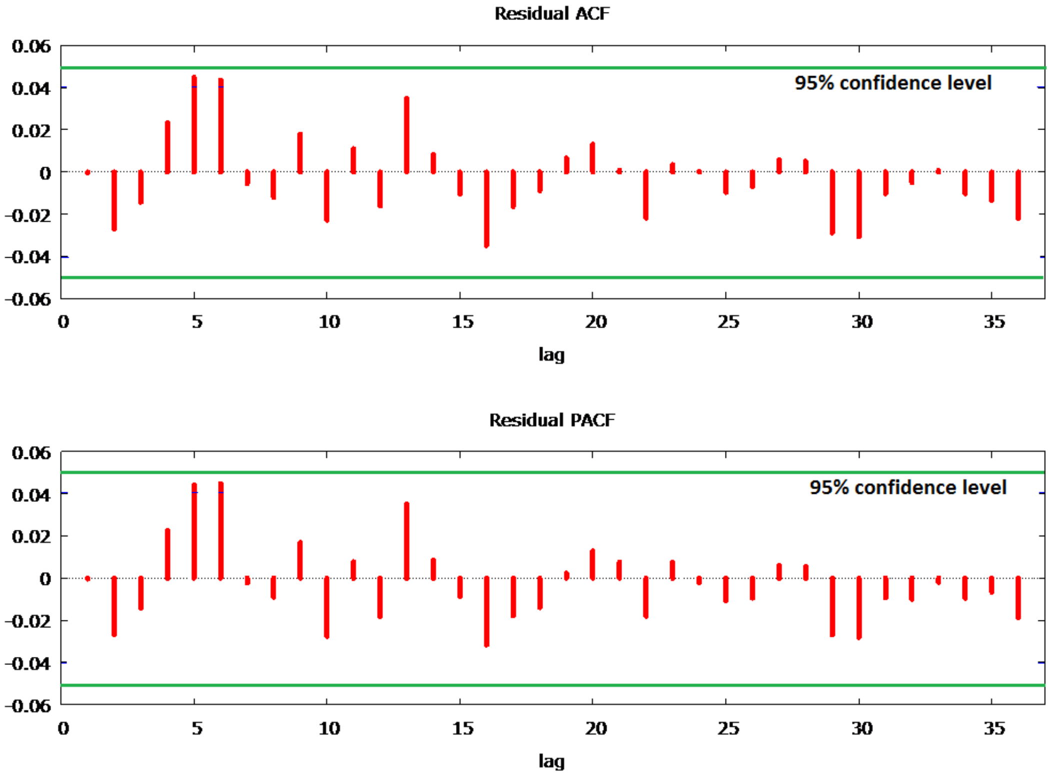

Figure 5.2: Correlogram of Boone County roads (2013-2015) dataset

5.2: Impacts of Spatial Autocorrelation

5.2.1: Global Moran’s I and General Gi*

The Global Moran’s *I* and the General *Gi** statistic evaluate whether the overall highway crashes are clustered, dispersed, or random, and assess the overall clustering patterns of the data. The Global Moran’s *I*, *z* scores, and *p*-values are reported in Table 5.9 for both the I-70 corridor and the Boone County roads.

Table 5.9: Global Moran's *I* for I-70 and Boone County, MO

Dataset	Year	Global Moran's *I*	z-score	p-value	Decision
I-70	2013-2015	0.594	415.09	0.0000	sig.
	2015	0.589	136.23	0.0000	sig.
Boone County	2013-2015	0.636	485.62	0.0000	sig.
	2015	0.583	177.65	0.0000	sig.

The results of the analysis are interpreted within the context of the null hypothesis. For the Global Moran's *I* statistic, the null hypothesis states that the attributes (i.e. the I-70 and Boone County roads) being analyzed are randomly distributed among the features in the study area (i.e. there is no global spatial autocorrelation exists for the entire area). Since the *p*-values in Table 5.9 for both the I-70 and the Boone County roads are smaller than 0.05 (using a confidence level of 95%), then this indicates that the Global Moran's *I* is significant for the three years' level and the one year level as well, and hence, we can reject the null hypothesis, and state that it is quite possible that the spatial distribution of the overall I-70 and Boone County road crashes is the result of clustered spatial processes, and this observed spatial patterns of the crashes could very well be one of many possible versions of complete spatial clustering.

The General *Gi** statistic, *z* scores, and *p*-values are shown in Table 5.10 for both the I-70 corridor and the Boone County road network. The results of the analysis again are interpreted within the context of the null hypothesis. For the General *Gi** statistic, the null hypothesis states that the attributes (i.e. the I-70 and Boone County) being analyzed are randomly distributed among the features in the study area (i.e. there is no global spatial autocorrelation exists for the entire area).

Table 5.10: General Gi* for I-70 and Boone County, MO

Dataset	Year	General Gi*	z-score	p-value	Decision
I-70	2013-2015	0.098	- 3.952	0.0000	sig.
	2015	0.029	+ 5.638	0.0000	sig.
Boone County	2013-2015	0.055	- 18.366	0.0000	sig.
	2015	0.073	- 11.346	0.0000	sig.

Since the *p*-values in Table 5.10 for both the I-70 and the Boone County were smaller than 0.05 (using a confidence level of 95%), then this indicates that the General *Gi** spatial autocorrelation is significant for the three years' level and the one year level as well, and hence, we can reject the null hypothesis, and state that it is quite possible that the spatial distribution of the overall I-70 and Boone County crashes is the result of clustered spatial processes, and this observed spatial patterns of the crashes could very well be one of many possible versions of complete spatial clustering. This confirms the result of the Global Moran's *I* for both I-70 and Boone County.

5.2.2: Anselin Local Moran's I and Local Gi* Statistic

Table 5.11 and 5.12 show the results of the significant high spatial autocorrelation crashes, the significant low spatial autocorrelation crashes, outliers, and the non-significant random crashes of the I-70 and the Boone County by Anselin Moran's *I* and local G_i* statistic respectively at the three years' (2013-2015) and the one year (2015). The Anselin Moran's *I* identified significant clusters of high values (HH), significant clusters of low values (LL), significant outliers in which a high value is surrounded primarily by low values (HL), significant outliers in which a low value is surrounded primarily by high values (LH), and non-significant random crashes. The *Gi** identified significant HHs and LLs, and non-significant random crashes. The Gi* does not identify the outliers HLs or LHs. The Moran's *I* identified (2442) significant HHs crashes for the I-70 dataset at the three years', whereas the G_i* statistic identified (3298) HHs crashes for the I-70 dataset at the three years. The Moran's *I* identified (798) significant HHs crashes for the I-70 dataset at the one year (2015), whereas the G_i* statistic identified (979) significant HHs crashes for the I-70 dataset at the one year (2015). The Moran's *I* identified (1735) significant LLs crashes for the I-70 dataset at the three years', whereas the G_i* statistic identified (1957) significant LLs crashes for the I-70 dataset at the three years. The Moran's *I* identified (568) significant LLs crashes for the I-70 dataset at the one year (2015), whereas the G_i* statistic identified (665) significant LLs crashes for the I-70 dataset at the one year (2015). The Moran's *I* identified (147) significant HLs and (329) significant LHs for the I-70 dataset at the three years', compared to the G_i* that identified (0) significant HLs and (0) significant LHs as outliers. The Moran's *I* identified (55) significant HLs and (69) significant LHs for the I-70 dataset at the one years (2015), compared to the G_i* that identified (0) significant HLs and (0) significant LHs as outliers. So, it is clear that the Gi* statistic has identified a larger number of the (HHs) and (LLs) compared to the Moran's *I*. The Moran's *I* identified (808) significant HHs crashes for the Boone County dataset at the three years', whereas the G_i* statistic identified (1040) significant HHs crashes for the Boone County dataset at the three years. The Moran's *I* identified (220) significant HHs crashes for the Boone County dataset at the one year (2015), whereas the G_i* statistic identified (291) significant HHs crashes for the Boone dataset at the one year (2015). The Moran's *I* identified (1280) significant LLs crashes for the Boone dataset at the three years', whereas the G_i* statistic identified (1162) significant LLs crashes for the Boone dataset at the three years. The Moran's *I* identified (316) significant LLs crashes for the Boone dataset at the one year (2015), whereas the G_i* statistic identified (382) significant LLs crashes for the Boone dataset at the one year (2015). The Moran's *I* identified (61) significant HLs and (14) significant LHs for the Boone dataset at the three years', compared to the G_i* that identified (0) non-significant HLs and (0) non-significant LHs as outliers. The Moran's *I* identified (18) significant HLs and (16) significant LHs for the Boone dataset at the one years (2015), compared to the G_i* that identified (0) non-significant HLs and (0) non-significant LHs as outliers. So again, it is clear that the Gi* statistic has identified a larger number of the (HHs) compared to the Moran's *I*.

Table 5.11: Crash clustering patterns of I-70, MO

Index	Year	High-High HH	Low-Low LL	Outliers HL	Outliers LH	Random
Anselin Moran's *I*	2013-2015	2442	1735	147	329	1216
	2015	798	568	55	69	452
G_i* statistic	2013-2015	3298	1957	0	0	614
	2015	979	665	0	0	298

Table 5.12: Crash clustering patterns of Boone County roads, MO

Index	Year	High-High HH	Low-Low LL	Outliers HL	Outliers LH	Random
Anselin Moran's *I*	2013-2015	808	1280	61	14	185
	2015	220	316	18	16	173
G_i* statistic	2013-2015	1040	1162	0	0	146
	2015	291	382	0	0	70

Figure 5.3 and Figure 5.4 show the clustering pattern identified by Moran's *I* for the I-70 corridor at the three years (2013-2015) and the one year (2015) respectively. Figure 5.5 and Figure 5.6 show the clustering pattern identified by *Gi** statistic for the I-70 corridor at the three years (2013-2015) and the one year (2015) respectively. The three years clustering pattern is very close to the one-year clustering pattern in both Moran's *I* and *Gi** statistic. For example, in Moran's *I*, clusters # 1, # 2, # 4, # 6, and # 7 are almost similar in the three years and one-year level. Cluster # 3, and # 5 are slightly different (i.e. it shows only outliers, and random crashes in the three years' level, and it shows outliers, random crashes, and LLs in the one-year level). Likewise, in *Gi**statistic, clusters # 1, # 2, # 3, # 4, # 5, # 6, and # 7 are almost similar in the three years and one-year level. The only difference is between cluster # 2 and # 3 (i.e. it shows random crashes, and LLs in the three years' level, and it shows random crashes, but without LLs in the one-year level). In addition, the extent and type of hot spots, cold spots, outliers, and random crashes differ from one method to the other. For example, cluster # 2, # 4, and # 5 is identified by Moran's *I* as outliers or random crashes, while it has been identified as mixed LLs and random crashes by Gi*. Clusters # 1, and # 7 are identified

by both methods as mostly HHs. Cluster # 2, and # 6 are identified by Moran's as purely LLs, while it has been identified by Gi* as mixed LLs and random crashes.

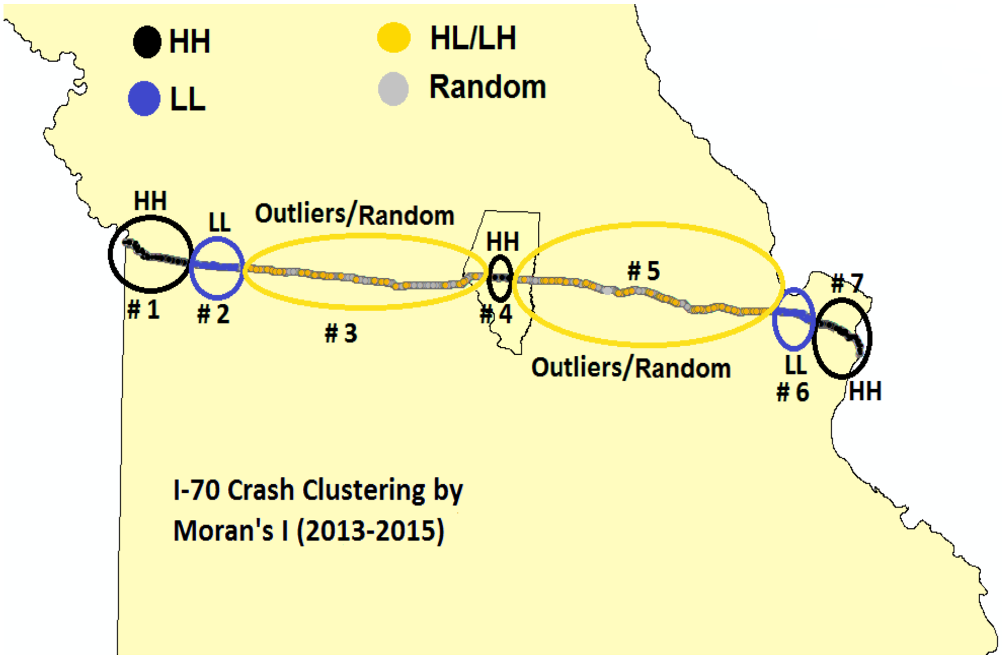

Figure 5.3: I-70 crash clustering (2013-2015) by Moran's I

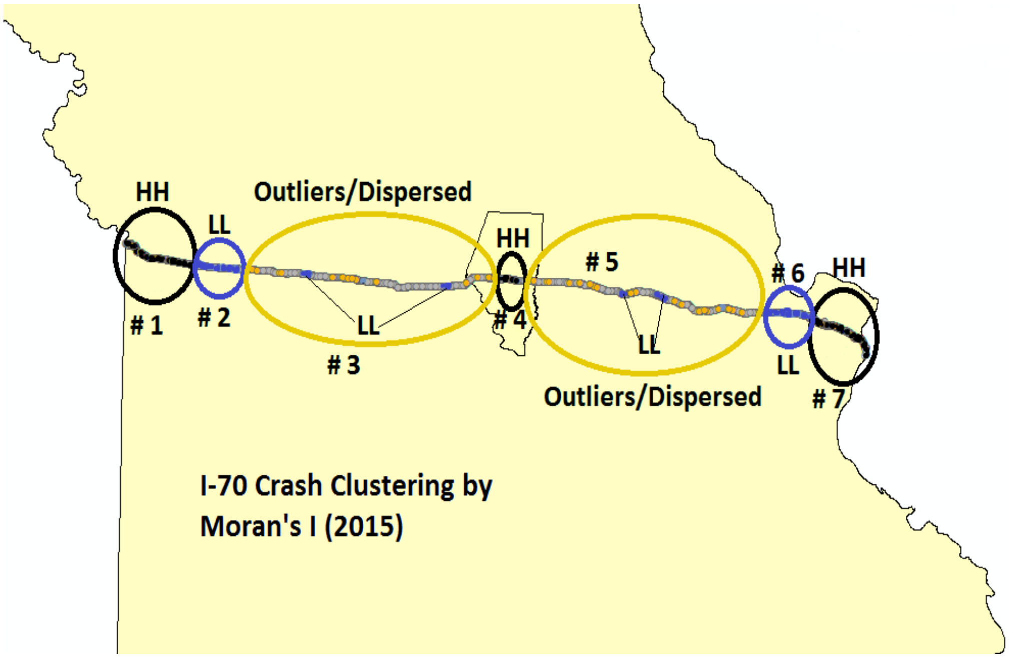

Figure 5.4: I-70 crash clustering (2015) by Moran's I

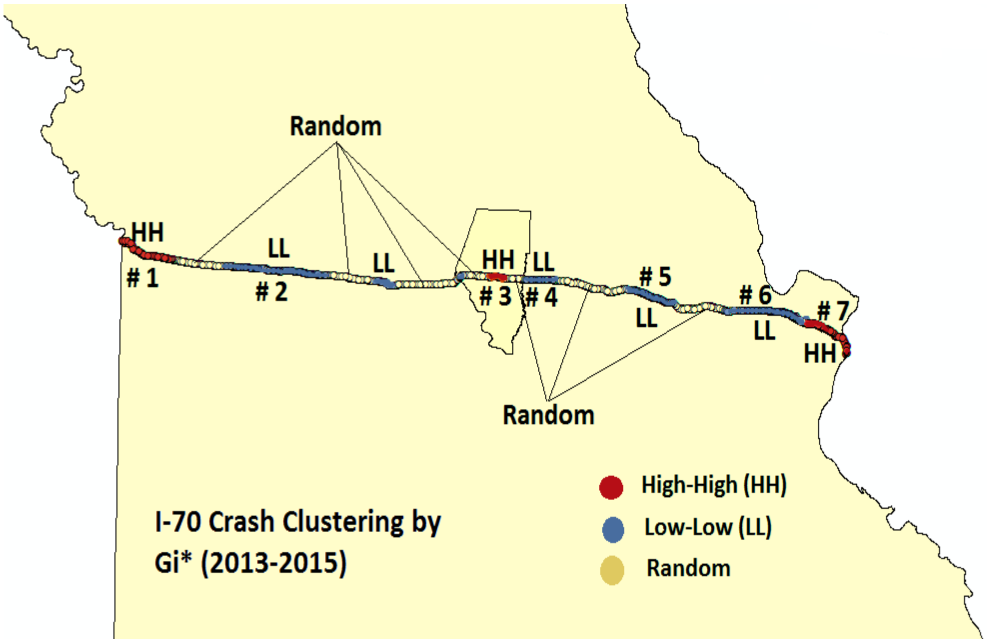

Figure 5.5: I-70 crash clustering (2013-2015) by Gi*

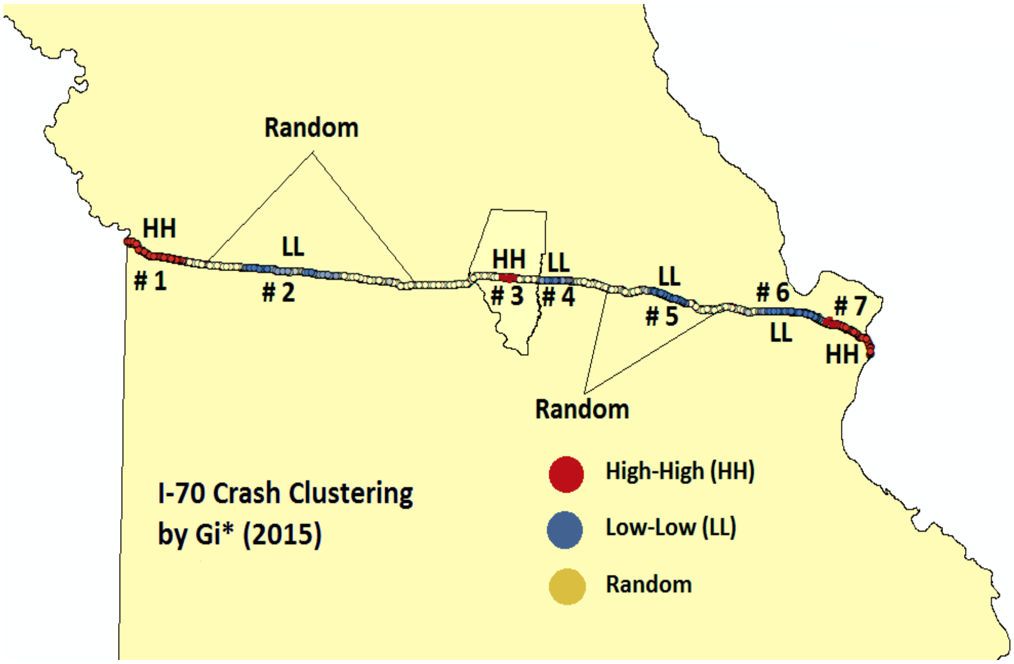

Figure 5.6: I-70 crash clustering (2015) by Gi*

Figure 5.7 and Figure 5.8 show the clustering pattern identified by Moran's *I* for the Boone County roads at the three years (2013-2015) and the one year (2015) respectively. Figure 5.9 and Figure 5.10 show the clustering pattern identified by *Gi** statistic for the Boone County roads at the three years (2013-2015) and the one year (2015) respectively. The three years' clustering pattern is very close to the one-year clustering pattern in both Moran's *I* and *Gi** statistic. For example, in Moran's *I*, cluster # 1contains HHs, LLs, HLs, LHs, and random crashes in both the three years and one-year dataset. Cluster # 2 is almost similar in the three years and one-year level except that the three years contains almost purely random crashes, while the one-year contains mixed HLs and random dispersed crashes. Cluster # 3 is slightly different (i.e. it shows LLs, HLs, and random crashes in the three years' level, and it shows random crashes, LLs, HLs, and LHs in the one-year level). In Gi*statistic, cluster # 1 is almost similar in the three years and one-year level. Cluster # 2 is slightly different (i.e. it shows LLs, and small number of HHs in the three years' level, and it shows LLs, and some random crashes in the one-year level). Cluster # 3 is almost similar in the three years and one year datasets. In addition, the extent and type of hot spots, cold spots, outliers, and random crashes differ from one method to the other. For example, cluster # 1 is identified by Moran's *I* as mixed HHs, LLs, HLs, LHs and random crashes while it has been identified as mostly HHs, LLs and random crashes by *Gi**. Clusters # 2 is identified by Moran's *I* as mostly random crashes, while it has been identified by *Gi** as mostly LLs. Cluster # 3 is identified by Moran's *I* as mixed LLs, HLs, and random crashes while it has been identified as mostly LLs by *Gi**.

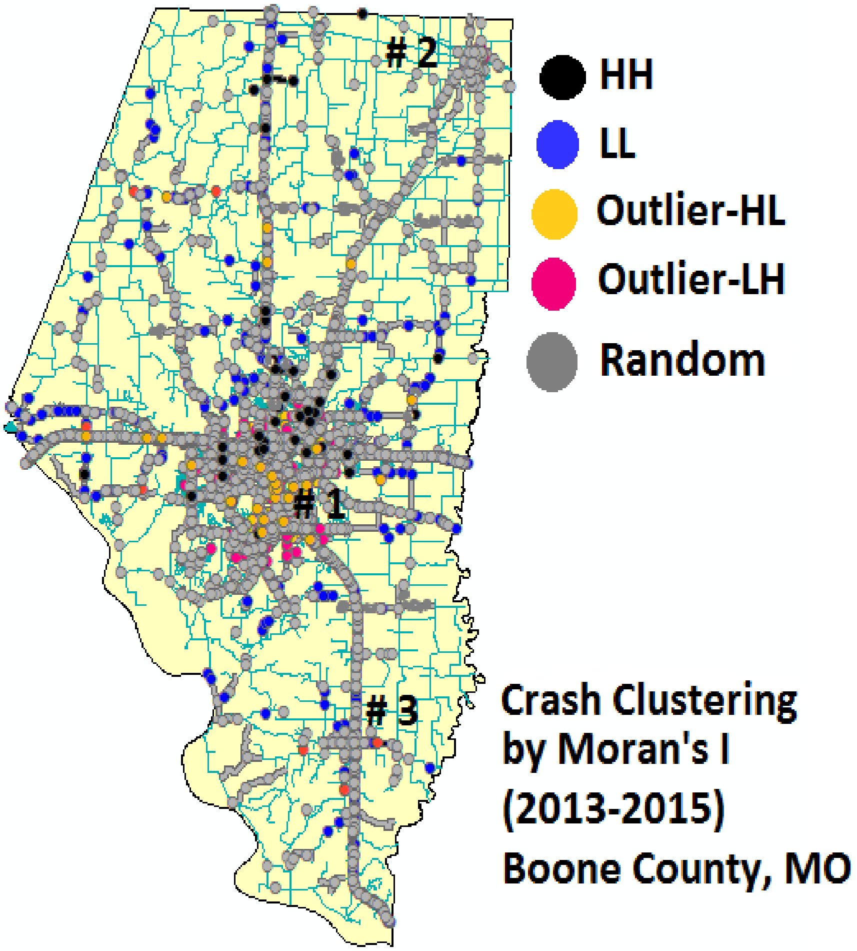

Figure 5.7: Boone crash clustering (2013-2015) by Moran’s I

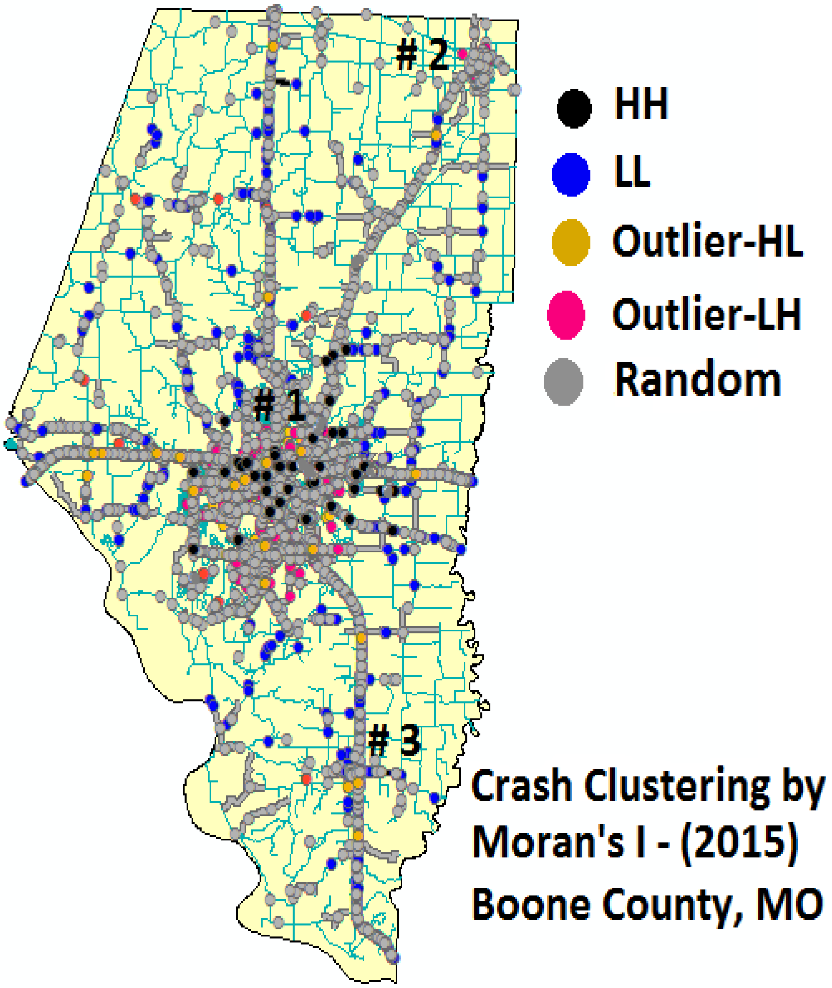

Figure 5.8: Boone crash clustering (2015) by Moran’s I

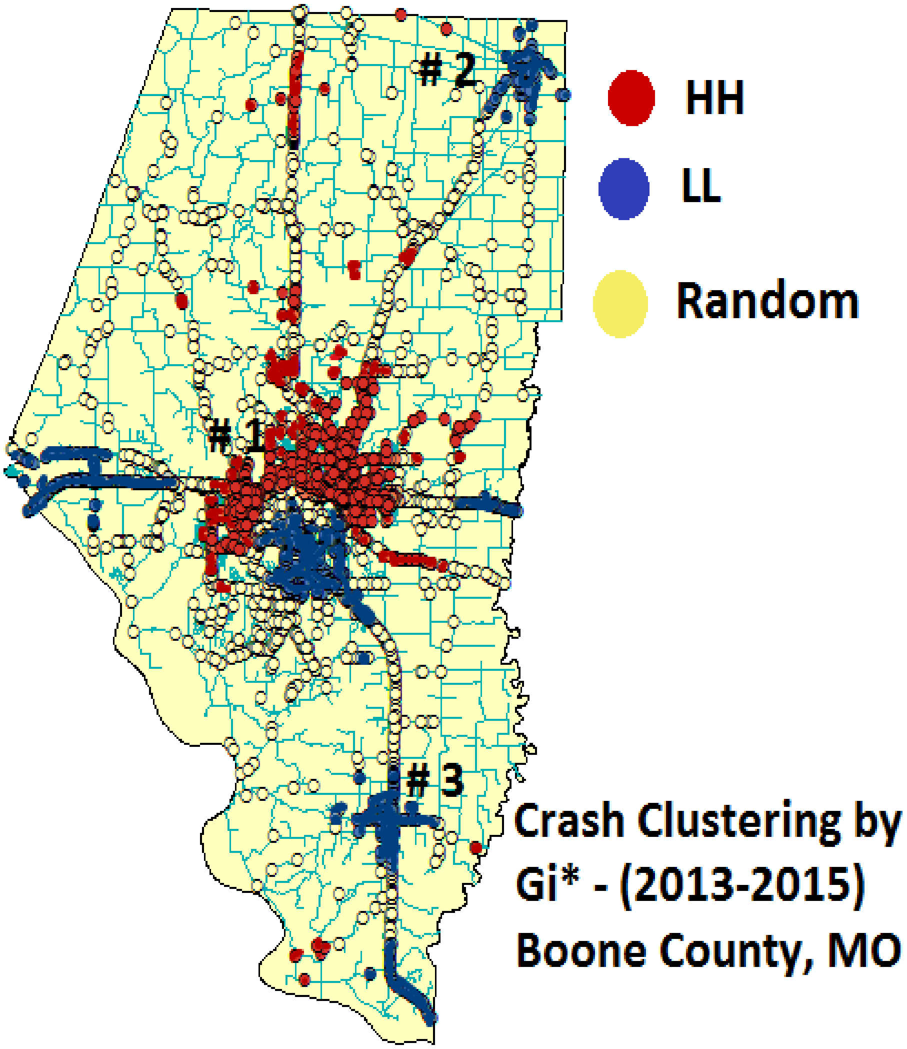

Figure 5.9: Boone crash clustering (2013-2015) by Gi*

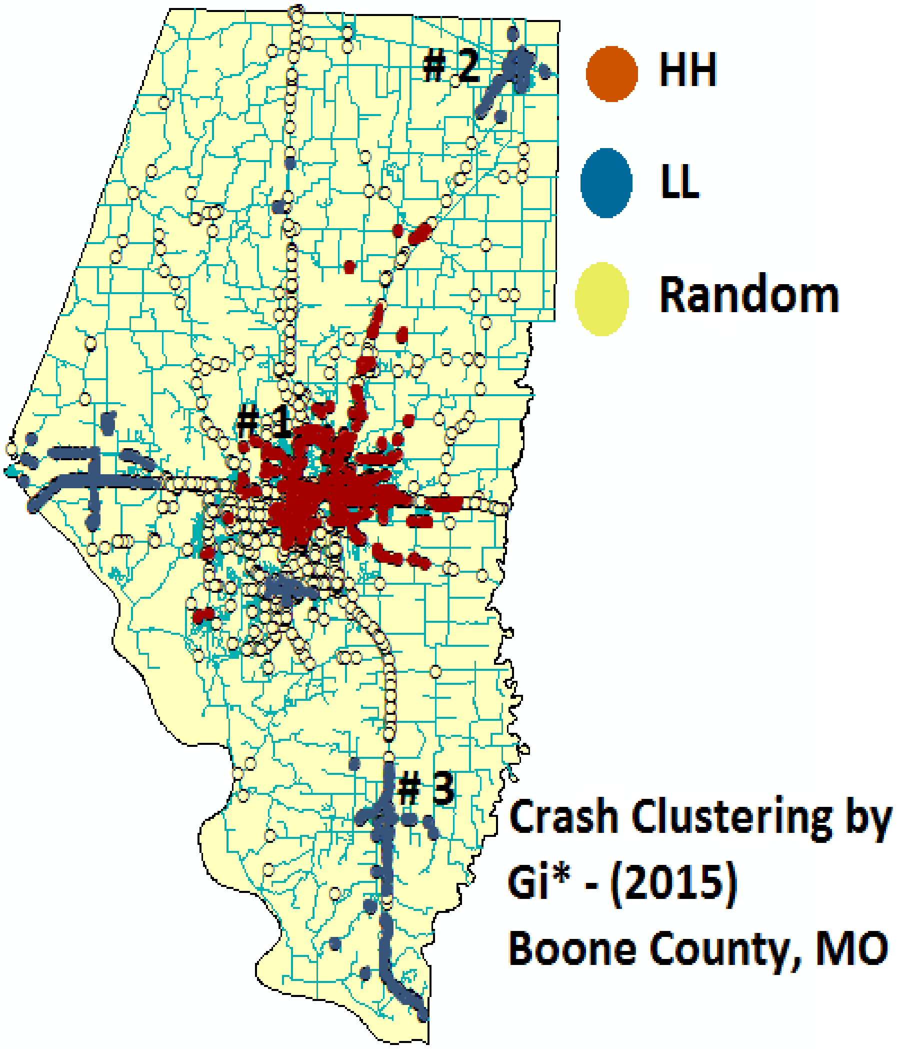

Figure 5.10: Boone crash clustering (2015) by Gi*

Since the *Gi** statistic has identified a larger number of the significant (HHs) and significant (LLs) compared to the Moran's *I*, for both the I-70 corridor and the Boone County roads, therefore the G_i* statistic (i.e. the G_i*'s *z*-scores) were used in modeling the crash severity in this dissertation. In addition, the clustering pattern of both the three-year period (2013-2015) and the one-year period (2015) was almost identical for both the I-70 corridor and the Boone County roads, which implies that the three years' (2013-

2015) level has no hidden or unobserved effects of spatial clustering, therefore, the three years’ datasets were chosen to be included in the analysis, and each crash was assigned a corresponding G_i*’s *z*-score spatial autocorrelation value at the three years’ level.

5.2.3: Spatial Autocorrelation by a Hybrid Method

A combination of 30% Moran’s *I*, and 70% G_i* is used to render yet another measure of spatial autocorrelation as shown in Figure 5.11 for the I-70 corridor and Figure 5.12 for the Boone County roads. A new combined index was created that consists of (30% Moran’s *I* + 70% $Gi*$) for both the I-70 corridor and the Boone County road network, then the hybrid method is applied using the Getis-Ord $Gi*$ index in ArcMap 10.2.2 toolkits for all crashes. The results produced new statistically significant spatial clusters of high spatial autocorrelation values and low spatial autocorrelation values. From Figure 5.11 for the I-70 corridor, it can be seen that cluster # 1 in Kansas City area and cluster # 5 in St. Louis area become mostly LLs compared to HHs in Moran’s *I* and $Gi*$. This change makes sense because clusters of LLs and random crashes are more likely happen in big cities as they consist of few clustered crashes (Myers et al. 2013). Cluster # 2 and # 4 become mostly HHs compared to LLs in Moran’s *I* and $Gi*$. This change again makes sense because clusters of HHs are more likely happen in the suburban areas of big cities as they consist of large clustered crashes (Myers et al. 2013). Cluster # 3 remains HHs in Columbia area, which is the same as in Moran’s *I* and $Gi*$. In addition, the new hybrid method resulted in the removal of all the LLs in the central portion of the I-70 corridor compared to Moran’s *I* and $Gi*$.

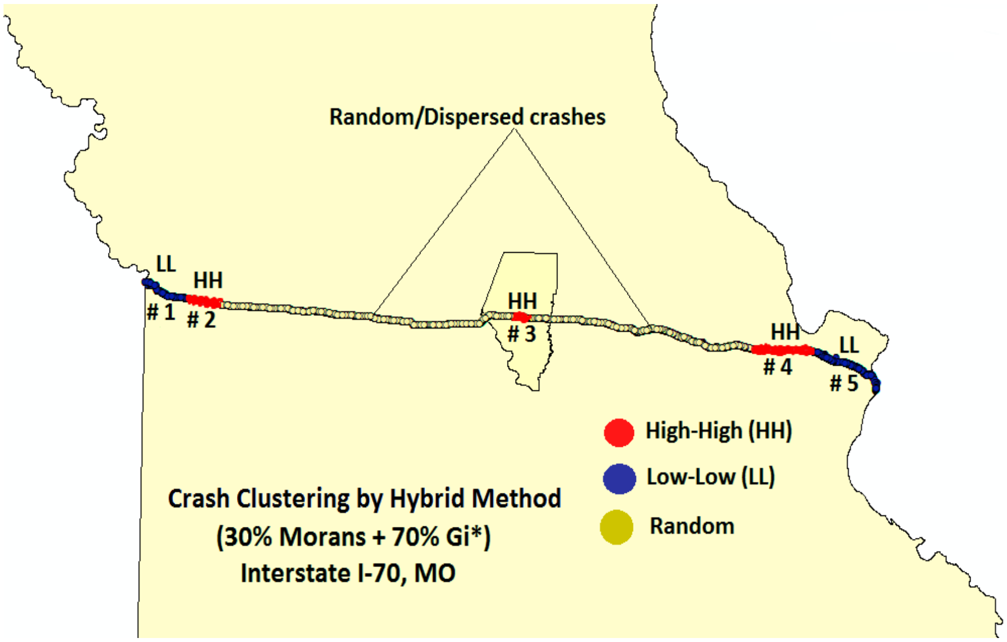

Figure 5.11: I-70 crash clustering via Hybrid method

From Figure 5.12 for the Boone County road network, is can be seen that cluster # 1 near the city of Columbia area is mixed of HHs, LLs, and random crashes compared to mostly HHs and LLs in *Gi** and mostly outliers, HHs, and LLs in Moran's *I*. Cluster # 2 now presents insignificant random crashes compared to mostly LLs in *Gi** and mostly outliers in Moran's *I*. Clusters # 3 becomes mostly insignificant random crashes with some LLs compared to mixed LLs, and outliers in Moran's *I* and LLs in *Gi**. This change makes sense because clusters of LLs and random crashes are more likely happen in big cities (i.e. Columbia) (Myers et al. 2013), and clusters of HHs are more likely happen in the suburban areas of big cities as they consist of large clustered crashes (Myers et al. 2013). The results of the new method of hybrid clustering patterns for both the I-70 corridor and Boone County have shown improvements in the new cluster mapping depending on the user's interpretation of the new patterns. Using different combination of Moran's, and *Gi** (i.e. 50% Moran's + 50% *Gi** for example) could result in different cluster mapping. In addition, the new hybrid method could produce new clusters if Moran's *I* is used in determining the hybrid results instead of the *Gi** statistic.

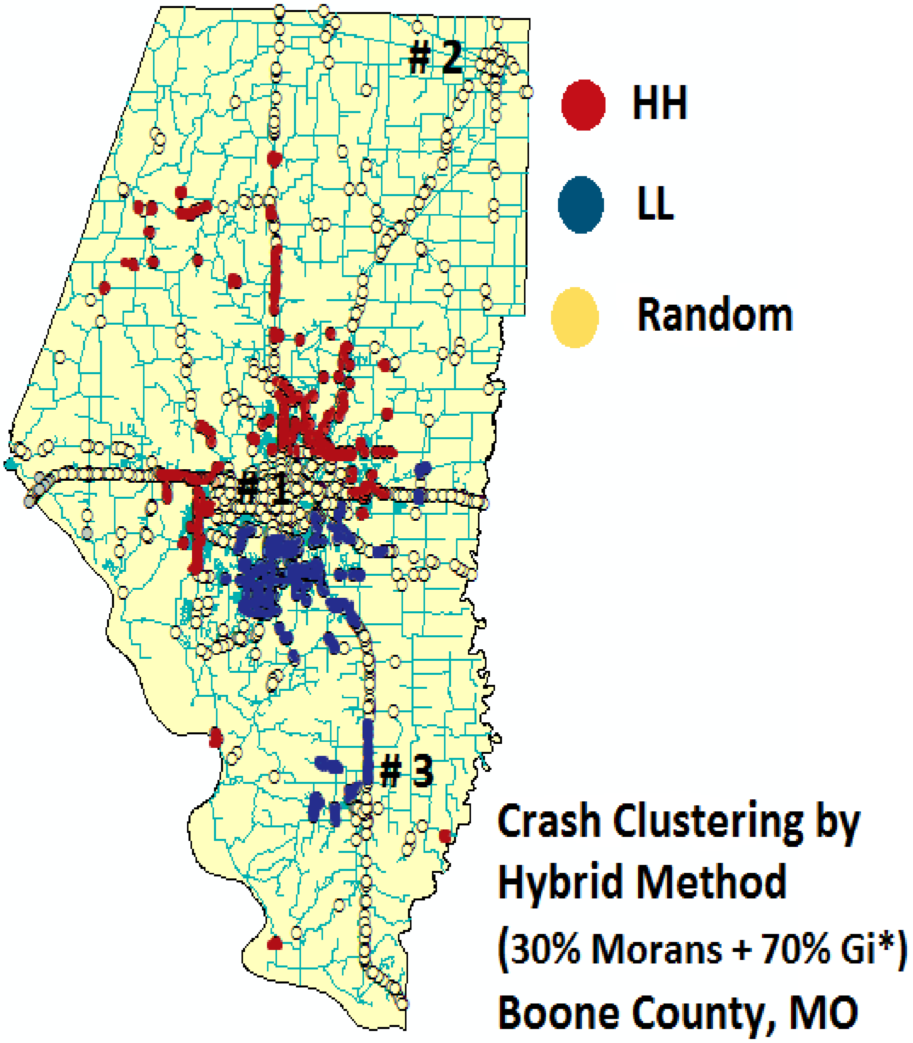

Figure 5.12: Boone crash clustering via Hybrid method

Table 5.13 details the HHs, LLs, HLs, LHs, and random crashes identified by Moran's *I*, *Gi**, and the new hybrid method for the I-70 corridor. Table 5.14 details the HHs, LLs, HLs, LHs, and random crashes identified by Moran's *I*, *Gi**, and the new hybrid method for the Boone County roads.

Table 5.13: I-70 crash clustering by Moran’s I, Gi*, and Hybrid method

Index	High-High HH	Low-Low LL	Outliers HL	Outliers LH	Random
Anselin Moran’s *I*	2442	1735	147	329	1216
G_i* statistic	3298	1957	0	0	614
Hybrid	918	1459	0	0	3492

Table 5.14: Boone crash clustering by Moran’s I, Gi*, and Hybrid method

Index	High-High HH	Low-Low LL	Outliers HL	Outliers LH	Random
Anselin Moran’s *I*	808	1280	61	14	185
G_i* statistic	1040	1162	0	0	146
Hybrid	322	874	0	0	1152

From Table 5.13 and Table 5.14, we can see that the number of the significant hot spots (HHs) and (LLs) identified by the hybrid method have decreased compared to Moran’s *I* and *Gi**. However, the number of insignificant random crashes identified by this method has increased compared to the other two methods.

5.3: Impacts of Sight Distance

5.3.1: Sight Distance Along I-70 Corridor

Table 5.15 shows the longitude, latitude, and lengths of I-70 segments with potential visibility issues relative to AASHTO (2011) decision sight distance (DSD) of 330.0 m, and Figure 5.13 shows their approximate locations along I-70 corridor.

Table 5.15: I-70 segments with potential visibility issues

Segment	From		To		Length km
	Long.	Lat.	Long.	Lat.	
1	-94.355665	39.040854	-94.331322	39.037485	2.14
2	-94.210207	39.022696	-94.184787	39.020336	2.21
3	-93.905340	39.006390	-93.870017	39.004924	3.06
4	-93.554500	38.989800	-93.530170	38.988139	2.11
5	-93.263254	38.971130	-93.239222	38.969708	2.08
6	-93.093837	38.953528	-93.069387	38.951666	2.12
7	-92.972242	38.933410	-92.948582	38.933872	2.05
8	-92.390708	38.970300	-92.366791	38.966601	2.11
9	-92.342165	38.969390	-92.318084	38.968274	2.09
10	-90.578508	38.797010	-90.554344	38.790528	2.21
11	-90.204219	38.672989	-90.194226	38.653716	2.31

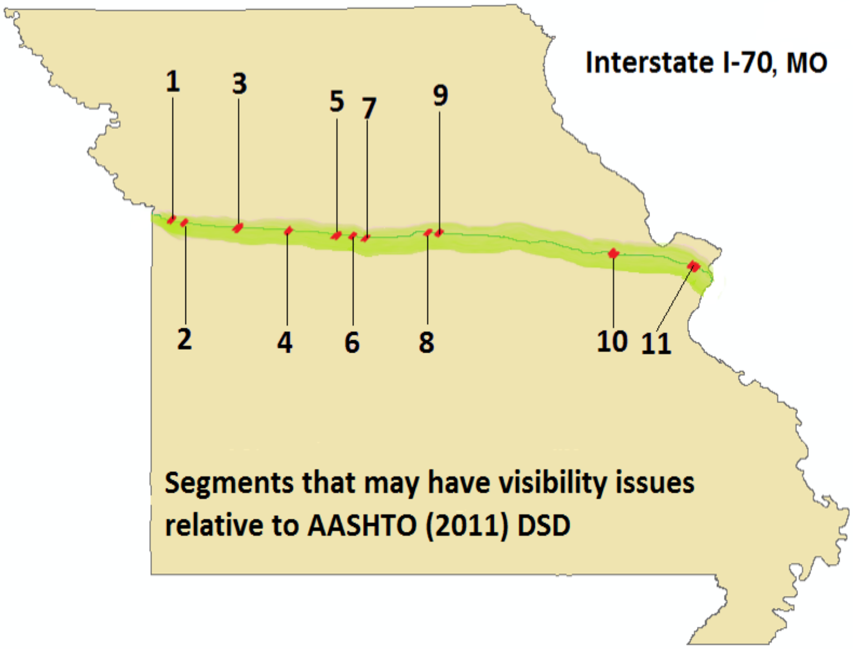
Interstate I-70, MO
1
3
5
7
9
2
4
6
8
10
11
Segments that may have visibility issues
relative to AASHTO (2011) DSD

Figure 5.13: I-70 segments with potential visibility issues

Segment number 1, for instance, may have visibility issues relative to AASHTO (2011) decision SD as shown in Figure 5.14. Looking at the image of segment number1, using Google Earth, we can see that there is an interchange (also called exit) between the start location of this segment (39.040854, - 94.355665) and the end location of the segment (39.037485, - 94.331322), as shown in Figure 5.14. It is possible that the vertical clearance under the interchange bridge could have an adverse effect on the available decision sight distance at this segment, because it could decrease the vertical scan range of the upper vertical angle of the driver's eye. Another possible interpretation is that the entrance and exit ramp maneuvers could have an adverse effect on the available decision sight distance, because it could decrease the horizontal scan range (i.e. the azimuth scan range) of the driver's eye at this location.

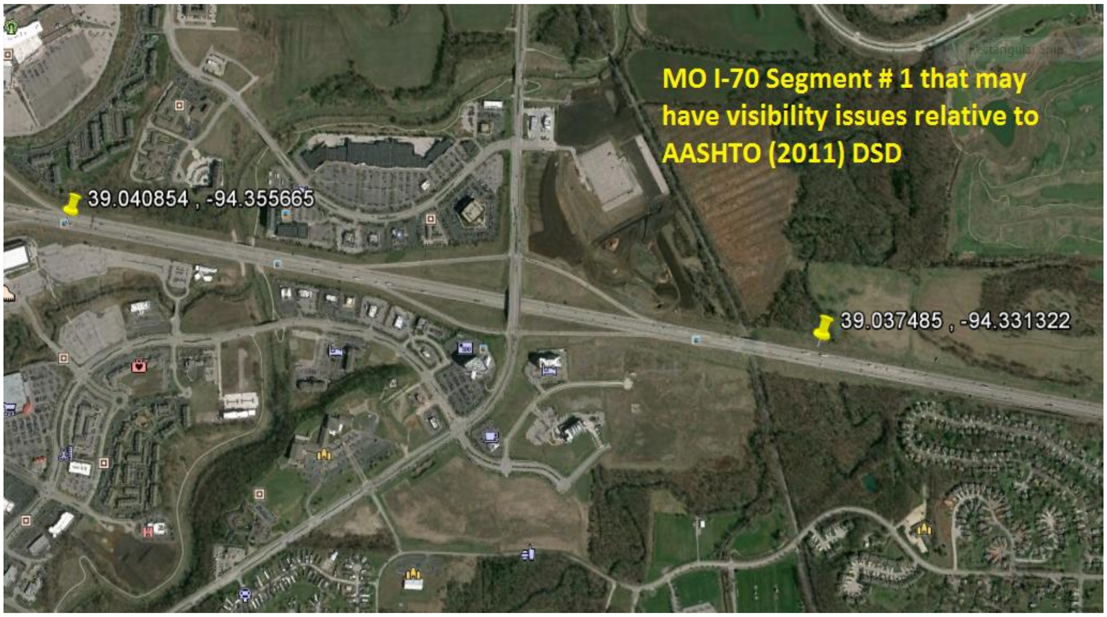
MO I-70 Segment # 1 that may have visibility issues relative to AASHTO (2011) DSD
39.040854 , -94.355665
39.037485 , -94.331322

Figure 5.14: MO I-70 segment 1 with potential visibility issues

Another example is the image of segment number 8. We can see that there is an interchange between the start location of this segment (38.970300, - 92.390708) and the end location of the segment (38.966601, - 92.366791), as shown in Figure 5.15. It is again possible that the vertical clearance under the interchange bridge could have an adverse effect on the available decision sight distance at this segment, because it could decrease the vertical scan range of the upper vertical angle of the driver’s eye, or it is possible that the entrance and exit ramp maneuvers could have an adverse effect on the available decision sight distance, because it could decrease the horizontal scan range (i.e. the azimuth scan range) of the driver’s eye at this location.

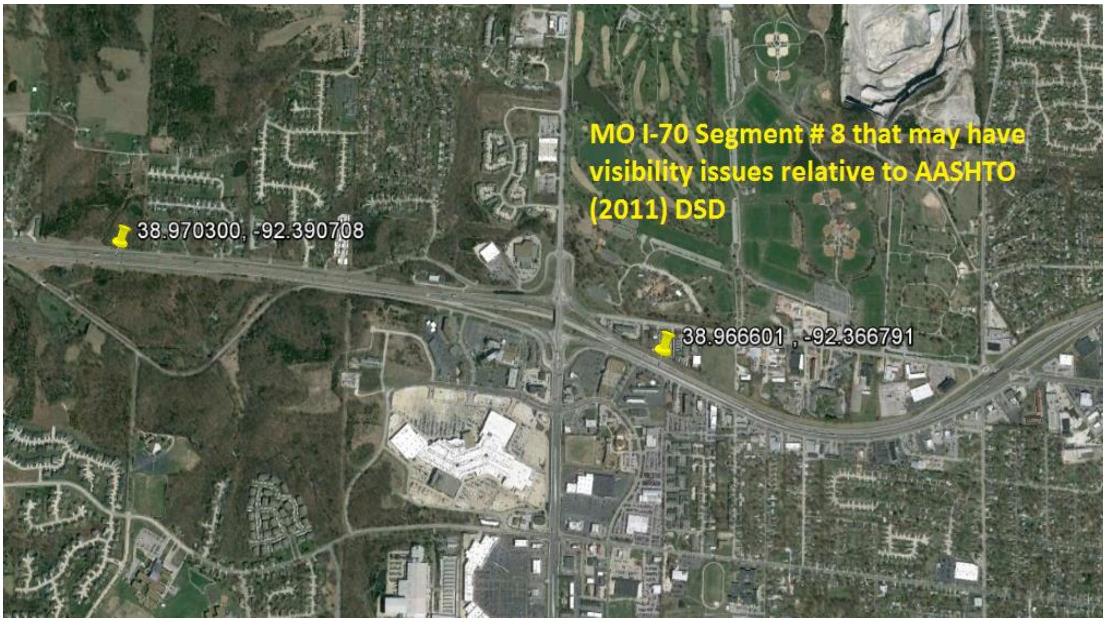
MO I-70 Segment # 8 that may have visibility issues relative to AASHTO (2011) DSD
38.970300, -92.390708
38.966601 , -92.366791

Figure 5.15: MO I-70 segment 8 with potential visibility issues

Another example is the segment number 3. We can see that there are two interchanges between the start of this segment (39.006390, -93.905340) and the end of the segment (39.004924, -93.870017), as shown in Figure 5.16. It is again possible that the vertical clearance under the interchange bridges could have an adverse effect on the available decision sight distance at this segment, because it could decrease the vertical scan range of the upper vertical angle of the driver’s eye, or it is possible that the entrance and exit ramp maneuvers of any of these two interchanges could have an adverse effect on the available decision sight distance, because it could decrease the horizontal scan range (i.e. the azimuth scan range) of the driver’s eye at this location.

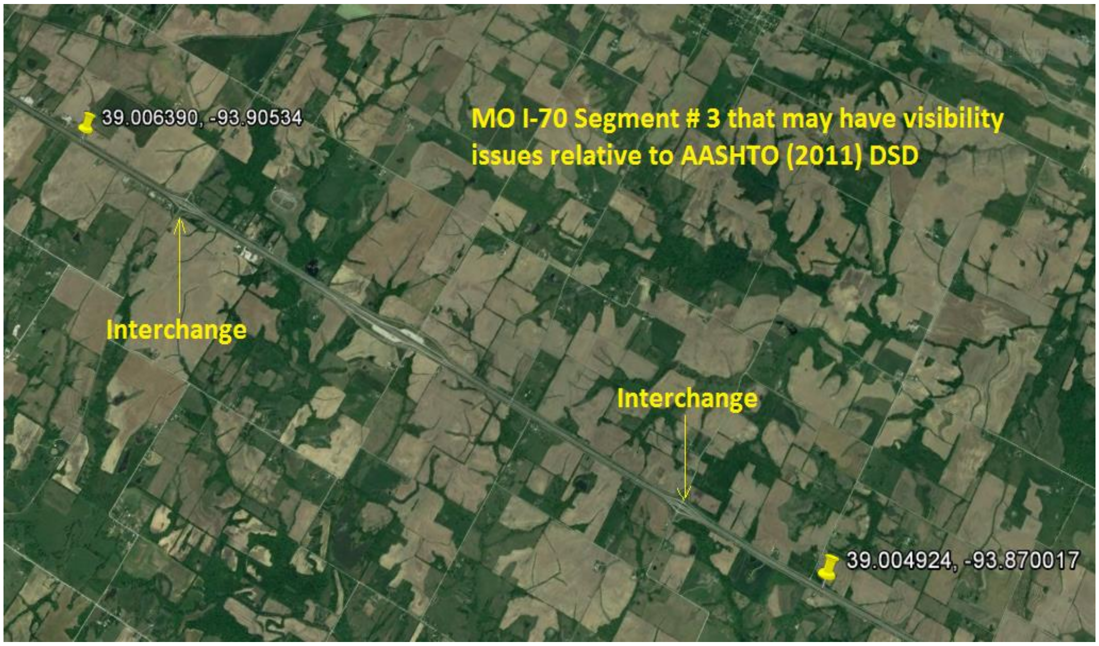
39.006390, -93.90534
MO I-70 Segment # 3 that may have visibility issues relative to AASHTO (2011) DSD
Interchange
Interchange
39.004924, -93.870017

Figure 5.16: MO I-70 segment 3 with potential visibility issues

The remaining segments of I-70 (i.e. segments number 2, 4, 5, 6, 7, 9, 10) with potential visibility issues relative to AASHTO (2011) DSD can be interpreted in a similar way as there are interchanges between the start location and the end location of these segments.

5.3.2: Sight Distance Along Boone County Roads

Table 5.16 shows the longitude, latitude, and lengths of Boone County road segments with potential visibility issues relative to AASHTO (2011) decision sight distance of 200.0 m, and Figure 5.17 illustrates their approximate locations throughout the Boone County, MO.

Table 5.16: Boone roads segments with potential visibility issues

Road Name	From		To		Length km
	Lat.	Long.	Lat.	Long.	
1-Sydow	39.205990	-92.257740	39.197193	-92.258112	0.98
2-Dinwiddie	39.243401	-92.228303	39.236184	-92.227750	0.80
3-Angell	39.241531	-92.169930	39.234541	-92.170834	0.78
4-Rangeline	39.187692	-92.105960	39.178168	-92.106369	1.06
5-Varnon	39.087813	-92.233221	39.081695	-92.233503	0.68
6-Mcdonald	39.114700	-92.118440	39.109900	-92.118627	0.53
7-Nature	39.051783	-92.436955	39.048180	-92.437122	0.40
8-Hickory G.	38.950704	-92.472984	38.946218	-92.473131	0.50
9-Purdy	38.930070	-92.171844	38.927075	-92.172013	0.33
10-Warren	38.862713	-92.410867	38.859155	-92.409102	0.42
11-Smith H.	38.821463	-92.385533	38.817382	-92.383916	0.47
12-Clinken B.	38.781785	-92.183016	38.770873	-92.180108	1.24
13-Raitt	38.708705	-92.259812	38.706569	-92.260963	0.26

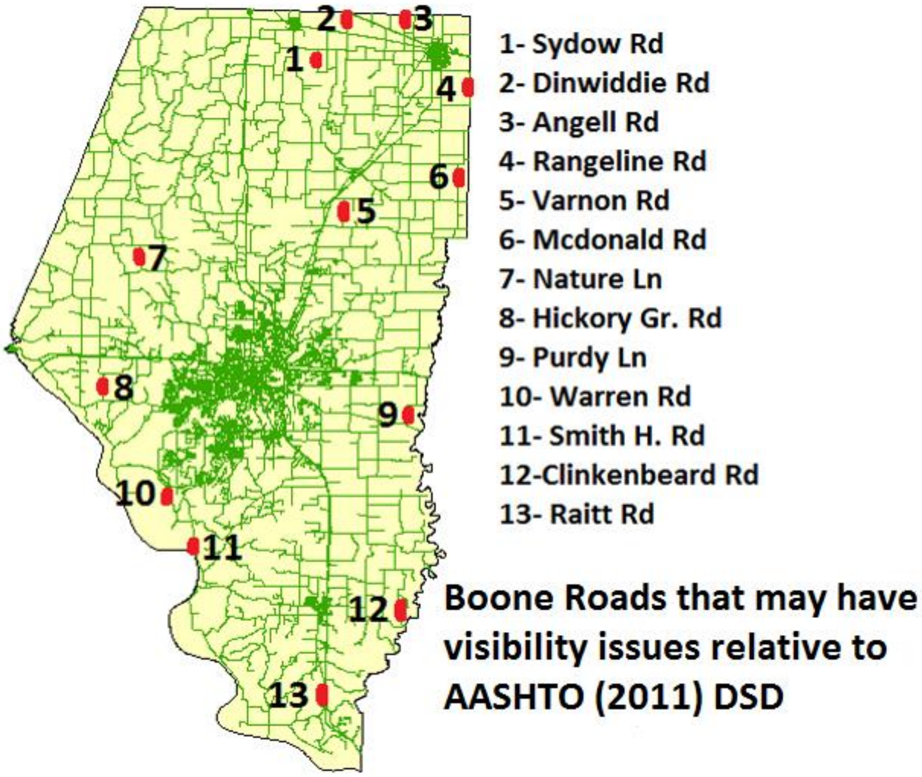

Figure 5.17: MO Boone roads with potential visibility issues

Looking at segment number 7, for example, using Google Earth, we can see that there are two horizontal curves between the start location of this segment (39.051783, -92.436955) and the end location of the segment (39.048180, -92.437122), as shown in Figure 5.18. It is possible that the distance ahead from the start or end positions up to the nearest horizontal curve is less than the AASHTO (2011) DSD requirement, implying that the existing curves could be the reason for the inadequate DSD at this segment.

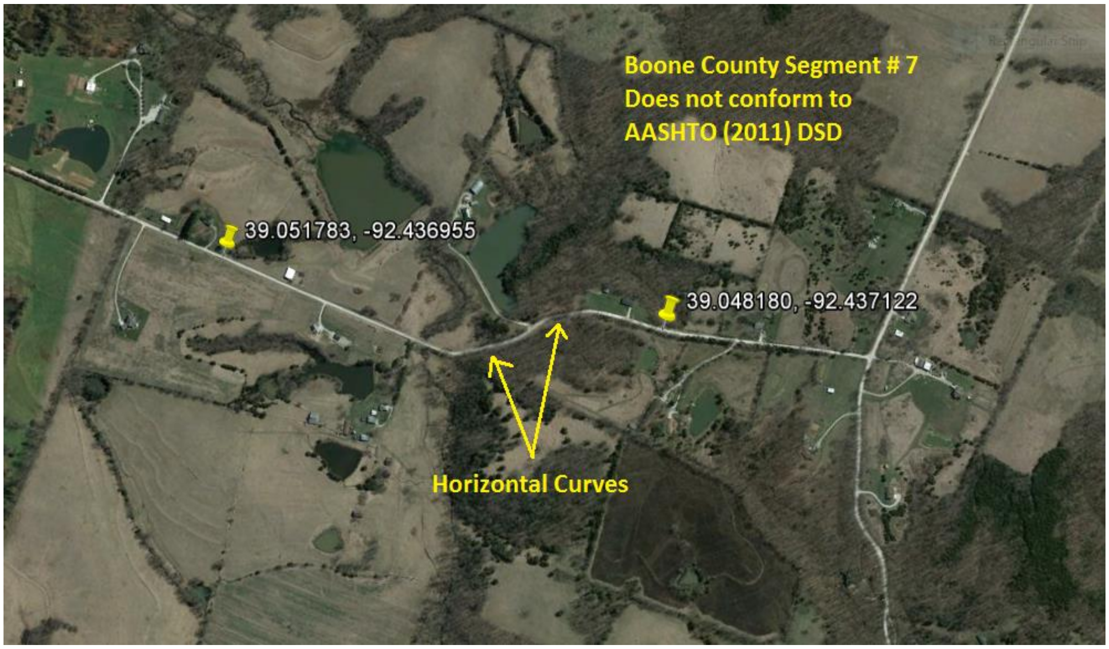

Figure 5.18: MO Boone roads segment 7 with potential visibility issues

Another example is segment number 11. We can see that there are two horizontal curves between the start location of this segment (38.821463, -92.385533) and the end location of the segment (38.817382, -92.383916), as shown in Figure 5.19. It is possible that the distance ahead from the start or end positions up to the nearest horizontal curve is less than the AASHTO (2011) DSD requirement, implying that the existing curves could be the reason for the inadequate DSD at this segment.

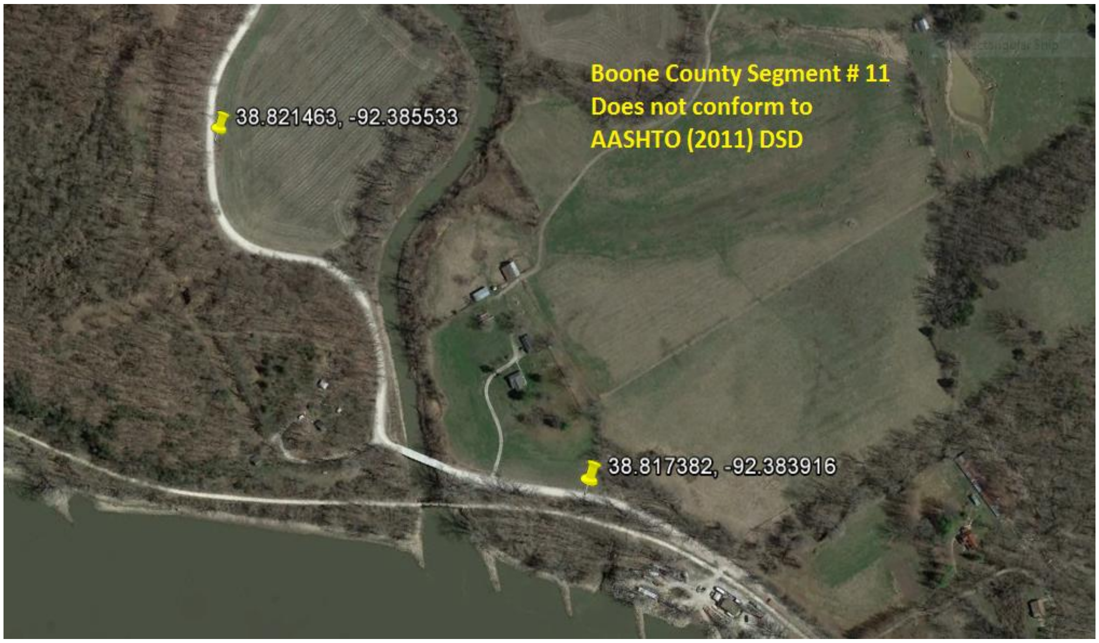

Figure 5.19: MO Boone roads segment 11 with potential visibility issues

The remaining segments of Boone County that may have visibility issues relative to AASHTO (2011) DSD can be interpreted in a similar way as there are curves or blocking features (i.e. side trees) between the start location and the end location of these segments.

5.3.3: Passing Sight Distance along MO Route 5

Table 5.17 shows the latitude, longitude, and the lengths of the no-passing zones (NPZs) throughout MO Route 5, and Figure 5.20 shows the locations of PZs and NPZs.

Table 5.17: The longitude, latitude, and lengths of NPZs at MO-R5

Segment	From		To		Length, km
	Long.	Lat.	Long.	Lat.	
1	-93.004359	40.484525	-93.039366	40.429639	6.78
2	-93.095767	40.370752	-93.103407	40.352586	2.11
3	-93.104301	40.238526	-93.108841	40.220032	2.20
4	-93.143679	40.163192	-93.151739	40.144075	2.39
5	-93.153424	40.106560	-93.154823	40.086905	2.22
6	-93.167665	39.955366	-93.167010	39.917416	4.22

7	-93.173719	39.822029	-93.175213	39.777525	4.95
8	-92.948035	39.712250	-92.960232	39.538017	19.40
9	-92.924321	39.519018	-92.931916	39.431446	9.76
10	-92.849713	39.405615	-92.850153	39.368578	4.24
11	-92.848705	39.273705	-92.848248	39.254711	2.37
12	-92.737894	39.008679	-92.738758	38.998677	1.32
13	-92.776593	38.934922	-92.788225	38.913645	2.88
14	-92.826733	38.875756	-92.839573	38.857380	2.61
15	-92.852559	38.803348	-92.858902	38.671686	14.65
16	-92.783911	38.651579	-92.789150	38.592526	6.58
17	-92.804933	38.501483	-92.809033	38.460242	4.60
18	-92.828179	38.444716	-92.840909	38.440861	1.29
19	-92.852193	38.421307	-92.852686	38.336142	9.47
20	-92.841158	38.327190	-92.828479	38.317812	1.79
21	-92.783003	38.043133	-92.780999	38.030541	1.48
22	-92.766330	38.023961	-92.757785	38.016843	1.39
23	-92.695300	37.834654	-92.686848	37.812360	2.85
24	-92.677993	37.777665	-92.663442	37.680213	10.91
25	-92.649409	37.664414	-92.592571	37.540851	14.62
26	-92.590049	37.475200	-92.593311	37.392065	9.25
27	-92.546268	37.308944	-92.521418	37.264932	5.36
28	-92.571119	37.135156	-92.605843	37.077511	7.11
29	-92.637784	37.021681	-92.650352	37.002673	2.32
30	-92.662056	36.983856	-92.666214	36.940981	5.27
31	-92.680214	36.888024	-92.667394	36.878627	1.63
32	-92.653373	36.868902	-92.643354	36.863384	1.38
33	-92.626141	36.850418	-92.620128	36.847817	1.22
34	-92.604948	36.831432	-92.596547	36.825270	1.25
35	-92.557487	36.774156	-92.547966	36.755395	2.66
36	-92.538684	36.736304	-92.522083	36.698760	4.92
37	-92.512340	36.679130	-92.503572	36.671346	1.46
38	-92.479960	36.666402	-92.503176	36.670698	1.15
39	-92.466909	36.641292	-92.456152	36.632850	1.33
40	-92.439636	36.622524	-92.431578	36.600425	3.22
41	-92.474234	36.565489	-92.492366	36.546843	2.84
42	-92.481389	36.512929	-92.482681	36.499052	2.10

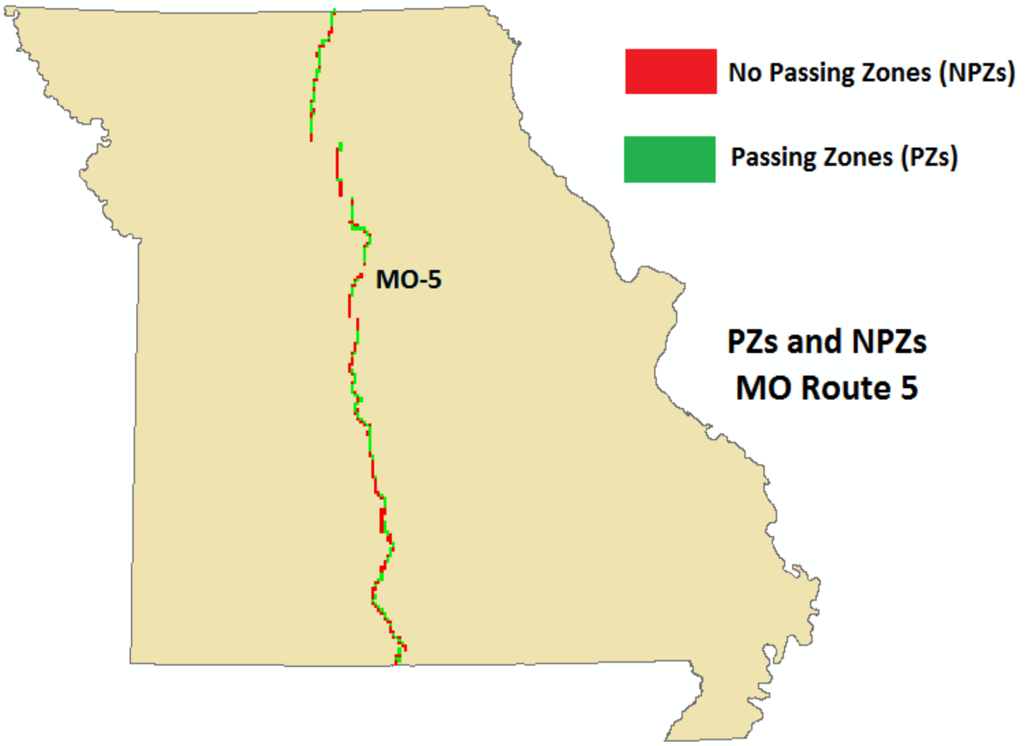

Figure 5.20: Locations of PZs and NPZs at MO Route 5

Looking at the end location (39.538017, -92.960232) of segment number 8, for example, using Google Earth, we can see that the surrounding area is hilly with an upgrade ahead, as sown in Figure 5.21. This implies that the existence of upgrades and/or curves could be the reason for the inadequate AASHTO (2011) passing sight distance at this segment. Another example is the start location (37.308944, -92.546268) of segment number 27. We can see that the surrounding area is also hilly with an upgrade ahead, as sown in Figure 5.22. This again implies that the existence of upgrades and/or curves could be the reason for the inadequate AASHTO (2011) passing sight distance at this segment.

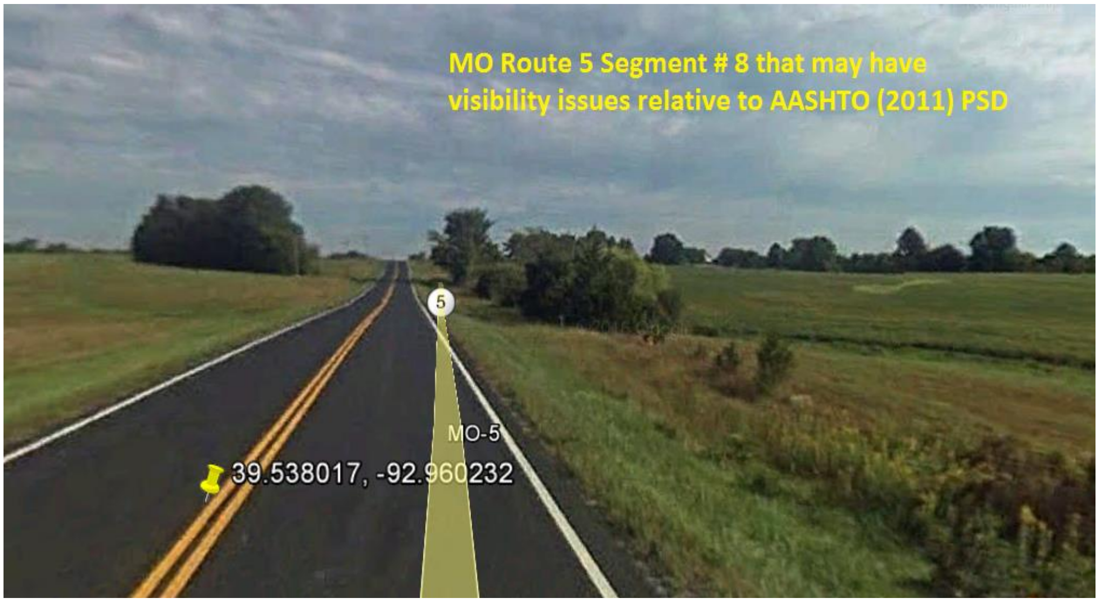
MO Route 5 Segment # 8 that may have visibility issues relative to AASHTO (2011) PSD
5
MO-5
39.538017, -92.960232

Figure 5.21: End location of segment 8 of MO Route

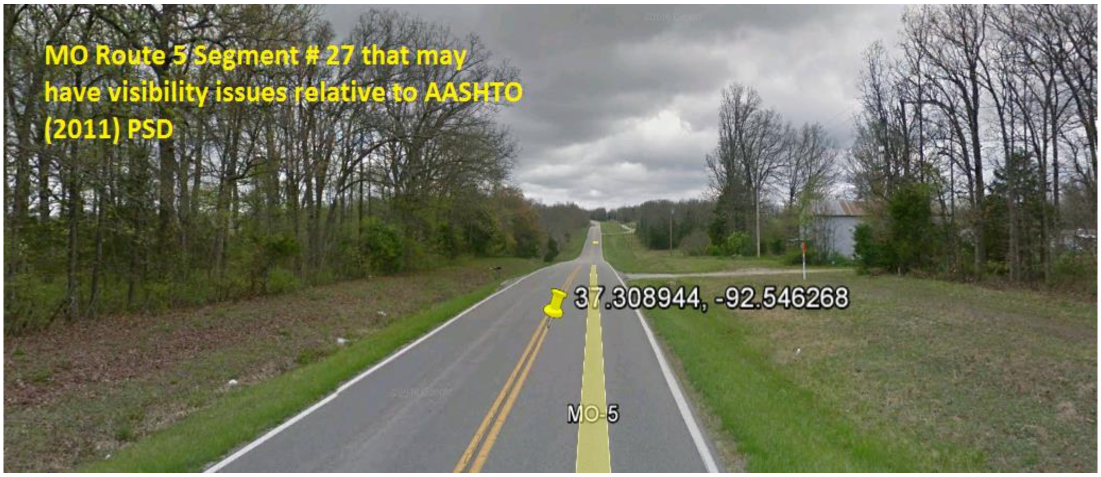
MO Route 5 Segment # 27 that may have visibility issues relative to AASHTO (2011) PSD
37.308944, -92.546268
MO-5

Figure 5.22: Start location of segment 27 of MO Route 5

5.4: Impacts of Multinomial Logistic Regression

The prediction results of the MNL are shown in the following sections:

5.4.1: Predicted Odd Ratios for I-70 Corridor

The odd ratios in MNL models present the probability of the event divided by the probability of the nonevent, and they can be obtained by exponentiating the multinomial logit coefficients (i.e. $e^{(coef.)}$). The multinomial logistic regression model estimates (k-1) models, where k is the number of outcome levels of the dependent variable, and the kth equation is relative to the referent group. In our model, the property damage is considered as the referent group (i.e. base level), because it is the most frequent outcome of crash severity, and the other outcome levels (i.e. minor injury, disabled, and fatal) are estimated relative to the property damage. The standard interpretation of the multinomial logistic regression is that for a unit change in the predictor variable, the odd ratio of outcome *m* relative to the referent group is expected to change by its respective parameter estimate given the other predictors in the model are held constant (Greene 2012, Judge et al. 1985). The predicted odd ratios for the I-70 corridor (for both training and testing data) are obtained using Stata 14 and reported in Table 5.18. The odd ratios are significant when their related *p*-values at the 95% confidence level are less than 0.05. If the odd ratios are greater than 1.0, then the predictors are positively correlated with the dependent variable (i.e. crash severity), and if the odd ratios are smaller than 1.0, then the predictors are negatively correlated with the dependent variable. In other words, if the odd ratios are greater than 1.0, then the predictors would increase the likelihood of the crash severity occurrence at the specified level, indicating positive contribution to the crash severity occurrence at that level, and if the odd ratios are smaller than 1.0, then the predictors would decrease the likelihood of the crash severity occurrence at the specified level, indicating negative contribution to the crash occurrence at that level.

Table 5.18: Predicted odd ratios for I-70, MO

Variable	I-70 Training Data			I-70 Testing Data		
	Odd Ratio	Std. error	p-value	Odd Ratio	Std. error	p-value
Crash Severity-Case 1: Minor Injury relative to base level (property damage)						
MONTH	1.015594	0.0121626	0.196	1.098245	.0187836	0.326
DAY_WEEK	.9868066	.0201894	0.516	.9911457	.0322842	0.421
HOUR	1.002493	.0069472	0.719	1.017365	.0112121	0.118
NO_VEHICLE	2.013444	.1528305	0.000	1.603548	.1673633	0.000
DIRECTION	1.001714	.204671	0.993	1.299179	.4009112	0.396
LIGHT_COND	1.018658	.0539375	0.727	1.079072	.0817907	0.800
ACC_TYPE	.7646827	.0322156	0.000	.827777	.0512309	0.002
DR_DRINK	.4393219	.0827939	0.000	.4566945	.1487597	0.016
SPEED	.7628727	.0832404	0.013	.7331396	.1258309	0.021
CZONE	.8728007	.1914342	0.882	.8306115	.4002926	0.384
DR_AGGRESSIVE	.6820784	.1231692	0.044	.6812309	.1762853	0.046
CELL_TEXT	.5149235	.1742725	0.049	.3814188	.2081773	0.047
DR_AGE	1.037926	.3769126	0.158	1.078291	.2189271	0.183
VEH_TYPE	.8286522	.1593428	0.462	0.857681	.1783352	0.413
RURAL_URBAN	1.21414	.1662723	0.157	1.194506	.2581555	0.411
NUMBER_ LANES	1.043295	.0714342	0.536	1.009117	.1109496	0.081
AADT	1.000573	.0018531	0.757	1.000707	.0028542	0.804
GRADE_LEVEL	.9969032	.2049085	0.988	.9728124	.3983592	0.425
SIGHT_DIST	.7887534	.1578979	0.563	.7821214	.2694931	0.387
Gi*	1.411994	.0908021	0.000	1.538624	.1471736	0.000
CONSTANT	.504704	.3106637	0.267	.3146406	.3145004	0.247
Crash Severity-Case 2: Disabled relative to base level (property damage)						
MONTH	1.04566	.0294898	0.113	1.052662	.044181	0.221
DAY_WEEK	.9849045	.055887	0.004	.9713375	.0714767	0.019
HOUR	1.0907501	.0153366	0.548	1.0921144	.0225957	0.067
NO_VEHICLE	2.325778	.346116	0.000	1.303495	.3296267	0.029
DIRECTION	1.0244691	.102775	0.141	1.0231048	.7614965	0.314
LIGHT_COND	1.0325387	.1239202	0.836	1.0277047	.2202536	0.156
ACC_TYPE	.77145632	.061105	0.000	.79145609	.1310232	0.006
DR_DRINK	.1758408	.0543585	0.000	.2855924	.1548372	0.021
SPEED	.6718398	.1729933	0.122	.5888928	.2284686	0.172
CZONE	.8159377	.5622705	0.760	.81661143	.3387404	0.375
DR_AGGRESSIVE	.79283284	.3286617	0.251	.72908047	.4614627	0.475
CELL_TEXT	.6839739	.518411	0.016	.6161388	.1346915	0.029
DR_AGE	1.098286	.482946	0.243	1.08442	.4398022	0.283
VEH_TYPE	.7338291	.172765	0.389	0.672993	.1798307	0.317
RURAL_URBAN	1.154855	.3503854	0.635	1.1281573	.4360739	0.274
NUMBER_ LANES	1.0623837	.1297353	0.035	1.0729747	.327797	0.041
AADT	1.0900496	.0048217	0.302	1.0993353	.0064666	0.306
GRADE_LEVEL	.99225575	.095807	0.000	.92474128	1.210746	0.015

SIGHT_DIST	.79043115	.3559404	0.733	.7477588	.8220375	0.102
Gi*	1.4794826	.1600518	0.899	1.48589779	.3452499	0.033
CONSTANT	.4430657	5.736671	0.273	.42062742	.40145	0.417
Crash Severity-Case 3: Fatal relative to base level (property damage)						
MONTH	1.204367	.0806321	0.005	1.204406	.0831499	0.014
DAY_WEEK	.9863804	.105922	0.737	.9828217	.1155401	0.177
HOUR	1.023859	.0319797	0.450	1.036516	.0377693	0.365
NO_VEHICLE	2.232134	.5612323	0.001	1.707896	.4912682	0.009
DIRECTION	1.099131	1.869167	0.515	1.042631	1.473025	0.231
LIGHT_COND	1.042018	.6304126	0.001	1.038765	.766612	0.007
ACC_TYPE	.7563569	.3752455	0.063	.6287748	.3629575	0.370
DR_DRINK	.1747316	.1104344	0.006	.2648509	.4530978	0.033
SPEED	.3108948	.2162619	0.093	.3551321	.334089	0.271
CZONE	.82678563	.1350873	0.081	.8429472	.244088	0.291
DR_AGGRESSIVE	.8619844	3.320254	0.003	.8827105	3.191887	0.008
CELL_TEXT	.2562309	.2849574	0.021	.0714367	.0850799	0.027
DR_AGE	1.0981655	.7690331	0.295	1.0616548	.2628931	0.319
VEH_TYPE	.7822954	.1692881	0.284	0.781194	.1672393	0.342
RURAL_URBAN	1.3862095	.4994118	0.605	1.4874849	.4314113	0.217
NUMBER_ LANES	1.0718678	.3198925	0.404	1.0565231	.9127193	0.344
AADT	1.002445	.0111129	0.226	1.0876658	.0134131	0.361
GRADE_LEVEL	.75853	.7540079	0.781	.82517107	2.630318	0.377
SIGHT_DIST	.7223947	.8266755	0.765	.7313107	.0126497	0.189
Gi*	1.93349	.6259672	0.042	2.124357	1.128395	0.045
CONSTANT	.68610	2.9507	0.916	.677015	1.40901	0.487

For example, when inspecting the MONTH predictor in the 1st case of crash severity (i.e. minor injury relative to property damage) in Table 5.18 for the training dataset, the odd ratio is greater than 1.0 (i.e. 1.015594), which indicates that this predictor is positively contributing to the crash severity at this level (i.e. minor injury), however it is not significant at the 95% confidence as its p-value is greater than 0.05. In other words, the contribution of the predictor MONTH to the crash severity of the level of minor injury, would be expected to increase by a factor of 1.015594 given the other variables in the model are held constant. When inspecting the DAY_WEEK predictor in the 1st case of crash severity (i.e. minor injury relative to property damage) in Table 5.18 for the training dataset, the odd ratio is smaller than 1.0 (i.e. 0.9868066), which indicates that this predictor is negatively contributing to the crash severity at this level (i.e. minor injury), and it is not significant at the 95% confidence as its p-value is greater than 0.05. When inspecting the NO_VEHICLE predictor in the 1st case of crash severity (i.e. minor injury relative to property damage) in Table 5.18 for the training dataset, the odd ratio is greater than 1.0 (i.e. 2.013444), which indicates that this predictor is positively contributing to the crash severity at this level (i.e. minor injury), and it is significant at the 95% confidence as its p-value is less than 0.05. So, the contribution of the predictor NO_VEHICLE to the crash severity of the level of minor injury, would be expected to increase by a factor of 2.013444 given the other variables in the model are held constant. Likewise, when inspecting the MONTH predictor in the 2nd case of crash severity (i.e. disabled relative to property damage) in Table 5.18 for the training dataset, the odd ratio is greater than 1.0 (i.e. 1.04566), which indicates that this predictor is positively contributing to the crash severity at this level (i.e. disabled), however it is not significant at the 95% confidence as its p-value is greater than 0.05. In other words, the contribution of the predictor MONTH to the crash severity of the level of “disabled”, would be expected to increase by a factor of 1.04566 given the other variables in the model are held constant. When inspecting the MONTH predictor in the 3rd case of crash severity (i.e. fatal relative to property damage) in Table 5.18 for the training dataset, the odd ratio is greater than 1.0 (i.e. 1.204367), which indicates that this predictor is positively contributing to the crash severity at this level (i.e. fatal), and it is significant at the 95% confidence as its p-value is less than 0.05. When inspecting the NO_VEHICLE predictor in the 2nd and 3rd cases of crash severity (i.e. disabled relative to property damage, and fatal relative to property damage) in Table 5.18 for the training dataset, the odd ratios are greater than 1.0 (i.e. 2.325778, 2.232134 respectively), which indicates that this predictor is positively contributing to the crash severity at these two levels (i.e. disabled, and fatal), and it is significant at the 95% confidence as its p-values are less than 0.05. So, the contribution of the predictor NO_VEHICLE to the crash severity of the levels of "disabled" and "fatal", would be expected to increase by a factor of 2.325778 and 2.232134 respectively given the other variables in the model are held constant.

5.4.2: Significant Risk Factors for I-70 Corridor

The statistically significant risk factors (i.e. predictors or independent variables) of the I-70 corridor in Missouri at the 95% confidence level are shown in Table 5.19.

Table 5.19: Significant risk factors for I-70, MO

Crash Severity Level	INTERSTATE I-70, MO	
	Significant Risk Factors	Significant Group Factors
Case 1: Minor Injury	1. NO_VEHICLE 2. ACC_TYPE 3. DR_DRINK 4. SPEED 5. DR_AGGRESSIVE 6. CELL_TEXT 7. Gi*	1. Driver Behavior 2. Accident Type 3. Spatial Autocorrelation
Case 2: Disabled	1. DAY_WEEK 2. NO_VEHICLE 3. ACC_TYPE 4. DR_DRINK 5. CELL_TEXT 6. NUMBER_LANES 7. GRADE_LEVEL	1. Time 2. Driver Behavior 3. Accident Type 4. Road Geometry
Case 3: Fatal	1. MONTH 2. NO_VEHICLES 3. LIGHT_COND 4. DR_DRINK 5. DR_AGGRESSIVE 6. CELL_TEXT 7. Gi*	1. Time 2. Driver Behavior 3. Environment 4. Spatial Autocorrelation

For the 1st case of crash severity level (i.e. minor injury relative to property damage), the number of vehicles involved in the crashes, the accident type, the driver drink, the speed, the driver aggressiveness, the cell-text, and the spatial autocorrelation index $Gi*$ are significant at the 95% confidence level. For the 2nd case of crash severity level (i.e. disabled relative to property damage), the day of the week, the number of vehicles involved in the crashes, the accident type, the driver drink, the cell-text, the number of lanes, and the grade of the road are significant at the 95% confidence level. For the 3rd case of crash severity level (i.e. fatal relative to property damage), the month of the year, the number of vehicles involved in the crashes, the light condition, the driver drink, the driver aggressiveness, the cell-text, and the spatial autocorrelation index $Gi*$ are significant at the 95% confidence level. We can see that two risk factors (i.e. the number of vehicles involved in the crashes and using the cell phones or texts when driving) are significant at the three crash severity levels (i.e. minor injury, disabled, fatal), indicating the importance of these two risk factors in modeling the severity of crashes of the I-70 corridor in MO. Some other risk factors are significant at only two levels of crash severity, but not at the third level. These risk factors are, the accident type, the driver drink, the driver aggressiveness, and the spatial autocorrelation index $Gi*$. The speed, the light condition, the number of lanes, the grade of the road, the day of the week, and the month of the year are significant at only one level of crash severity. In term of the significant group of factors, we can see that the driver's behavior group is the most important one as it has been related to the three crash severity levels, whereas the accident type, the time, and the spatial autocorrelation are next in their importance.

5.4.3: Predicted Odd Ratios for Roads in Boone County

In the MNL model of crashes along roads in Boone County, incidents involving property damage are also considered the referent group (i.e. base level), because it is the most frequent outcome of crash severity, and the other outcome levels (i.e. minor injury, disabled, and fatal) are estimated relative to the property damage. The standard interpretation of the multinomial logistic regression is that for a unit change in the predictor variable, the odd ratio of outcome *m* relative to the referent group is expected to change by its respective parameter estimate given the other predictors in the model are held constant (Judge et al. 1985; Greene 2012). The predicted odd ratios for the Boone County road network (for both training and testing data) are obtained using Stata 14 and summarized in Table 5.20. The odd ratios are significant when their related *p*-values at the 95% confidence level are less than 0.05. Again, if the odd ratios are greater than 1.0, then the predictors are positively correlated with the dependent variable (i.e. crash severity), and if the odd ratios are smaller than 1.0, then the predictors are negatively correlated with the dependent variable. That is, if the odd ratios are greater than 1.0, then the predictors would increase the likelihood of the crash severity occurrence at the specified level, indicating positive contribution to the crash occurrence at that level, and if the odd ratios are smaller than 1.0, then the predictors would decrease the likelihood of the crash severity occurrence at the specified level, indicating negative contribution to the crash occurrence at that level. For example, when inspecting the MONTH predictor in the 1st case of crash severity (i.e. minor injury relative to property damage) in Table 5.20 for the training dataset, the odd ratio is greater than 1.0 (i.e.1.023512), which indicates that this predictor is positively contributing to the crash severity at this level.

Table 5.20: Predicted odd ratios for Boone County crashes

Variable	Boone County Training Data			Boone County Testing Data		
	Odd Ratio	Std. error	p-val.	Odd Ratio	Std. error	p-val.
Crash Severity-Case 1: Minor Injury relative to base level (property damage)						
MONTH	1.023512	.0178198	0.182	1.0391251	.0267104	0.433
DAY_WEEK	1.050252	.0326366	0.115	1.051563	.0491662	0.282
HOUR	1.008425	.0109369	0.339	1.009422	.0172743	0.384
NO_VEHICLE	1.109713	.1340897	0.289	1.097612	.2147287	0.534
DIRECTION	.9594881	.0528776	0.353	.9926082	.0877978	0.333
LIGHT_COND	.9285046	.0790094	0.383	.9857805	.1285591	0.513
ACC_TYPE	1.11028	.0708534	0.011	1.1992554	.1067796	0.045
DR_DRINK	.4641011	.1170542	0.002	.43295565	.5322886	0.029
SPEED	.6548458	.1151811	0.016	.6649197	.2915138	0.023
CZONE	.7388649	1.083669	0.474	.7378288	.1595233	0.515
DR_AGGRESSIVE	.2881921	.1379217	0.022	.1202204	.1085349	0.019
CELL_TEXT	.5927003	.2307324	0.179	.6558653	.1403237	0.149
DR_AGE	1.0584471	.2267390	0.276	1.038509	.2148788	0.229
VEH_TYPE	.89795529	.16829845	0.371	0.818856	.1693376	0.393
RURAL_URBAN	.7351476	.2060147	0.048	.7454935	1.164492	0.039
NUMBER_LANES	1.001044	.0954435	0.491	1.0915989	.0893968	0.063
AADT	.9875969	.0048482	0.011	.97341468	.3273556	0.029
GRADE_LEVEL	.8373625	.1652473	0.008	.83470132	.1375984	0.029
SIGHT_DIST	.3958206	.1643202	0.026	.3771198	.0071083	0.038
Gi*	.9784839	.1000868	0.432	.9728619	.1787239	0.481
CONSTANT	.4663202	.4573384	0.437	.8331453	.4887709	0.556
Crash Severity-Case 2: Disabled relative to base level (property damage)						
MONTH	1.027664	.0396199	0.279	1.027918	.060554	0.341
DAY_WEEK	1.0867191	.0674279	0.345	1.066993	.1081113	0.503
HOUR	1.026415	.0234237	0.043	1.104746	.0458959	0.016
NO_VEHICLE	1.051400	.3663203	0.086	1.048285	.4649553	0.415
DIRECTION	.9108936	.1373576	0.404	.91477786	.1940384	0.610
LIGHT_COND	.9803828	.1763913	0.312	.8947385	.2473296	0.387
ACC_TYPE	1.125627	.1160428	0.017	1.1729226	.1992348	0.042
DR_DRINK	.2234158	.0939533	0.000	.4175788	.2678347	0.033
SPEED	.9628479	.3460866	0.316	.8971088	.547267	0.359
CZONE	.73511179	.1070972	0.491	.7364136	.5387798	0.364
DR_AGGRESSIVE	.4383074	.1475095	0.014	.4634503	.1668073	0.000
CELL_TEXT	.4083762	.2752782	0.018	.7646755	.3778548	0.048
DR_AGE	1.068703	.1673985	0.226	1.018493	.2373981	0.369
VEH_TYPE	.7773818	.1539893	0.418	.8398558	.1604776	0.441
RURAL_URBAN	.8308998	.2637057	0.316	.8357229	.3880534	0.345
NUMBER_LANES	1.0812004	.2357547	0.043	1.0709151	.1879346	0.041
AADT	.9906553	.0114682	0.417	.96553461	.8555893	0.424
GRADE_LEVEL	.7912488	.2725091	0.036	.79494419	.2314235	0.013

SIGHT_DIST	.6969566	.5264204	0.033	.6825348	.0152198	0.042
Gi*	.9064754	.2065277	0.466	.908094	.2845001	0.378
CONSTANT	.427664	.0396199	0.479	.443272	1.175688	0.870
Crash Severity-Case 3: Fatal relative to base level (property damage)						
MONTH	1.05751	.0953748	0.043	1.0328668	.2303312	0.039
DAY_WEEK	1.129193	.1291938	0.186	1.1214809	.8643811	0.677
HOUR	1.010052	.052407	0.347	1.0110705	.2534203	0.584
NO_VEHICLE	1.991718	.8497575	0.016	1.9654836	.3418705	0.028
DIRECTION	.8367988	.4153565	0.302	.8232213	.3379216	0.776
LIGHT_COND	.9437547	.4170947	0.496	.9205592	.7692041	0.463
ACC_TYPE	1.1730617	.3303531	0.336	1.1560646	.1175807	0.390
DR_DRINK	.0720745	.0629678	0.003	.1535938	.1099242	0.037
SPEED	.4816572	.3670916	0.038	.4213397	.5380519	0.042
CZONE	.7131492	.757008	0.596	.7562388	.258676	0.486
DR_AGGRESSIVE	.3602092	.3184739	0.024	.4555812	.1928740	0.049
CELL_TEXT	.3223915	.4151208	0.038	.4194732	.2332836	0.032
DR_AGE	1.027598	.3879915	0.275	1.059332	.247812	0.217
VEH_TYPE	.7699251	.1593362	0.383	.7299471	.1588603	0.428
RURAL_URBAN	.8389658	.3068237	0.231	.82182219	.2541412	0.495
NUMBER_ LANES	1.0087394	.5154228	0.466	1.0041855	.2802945	0.263
AADT	.9712408	.0229827	0.387	.98425828	.2727403	0.461
GRADE_LEVEL	.74658076	.5513545	0.421	.74048891	.1052855	0.375
SIGHT_DIST	.2706806	.0024528	0.689	.4225826	.2741932	0.672
Gi*	.9430182	.710843	0.372	.9406916	.1014820	0.496
CONSTANT	.1873913	.607109	0.592	.1881586	.2258365	0.551

In other words, the contribution of the predictor MONTH to the crash severity of the level of minor injury, would be expected to increase by a factor of 1.023512 given the other variables in the model are held constant. When inspecting the LIGHT_COND predictor in the 1st case of crash severity (i.e. minor injury relative to property damage) in Table 5.20 for the training dataset, the odd ratio is smaller than 1.0 (i.e.0.9285046), which indicates that this predictor is negatively contributing to the crash severity at this level (i.e. minor injury), and it is not significant at the 95% confidence as its p-value is greater than 0.05. When inspecting the ACC_TYPE predictor in the 1st case of crash severity (i.e. minor injury relative to property damage) in Table 5.20 for the training dataset, the odd ratio is greater than 1.0 (i.e.1.11028), which indicates that this predictor is positively contributing to the crash severity at this level (i.e. minor injury), and it is significant at the 95% confidence as its *p*-value is less than 0.05. So, the contribution of the predictor ACC_TYPE to the crash severity of the level of minor injury, would be expected to increase by a factor of 1.11028 given the other variables in the model are held constant. Likewise, when inspecting the MONTH predictor in the 2nd case of crash severity (i.e. disabled relative to property damage) in Table 5.20 for the training dataset, the odd ratio is greater than 1.0 (i.e. 1.027664), which indicates that this predictor is positively contributing to the crash severity at this level (i.e. disabled), however it is not significant at the 95% confidence as its p-value is greater than 0.05. In other words, the contribution of the predictor MONTH to the crash severity of the level of "disabled", would be expected to increase by a factor of 1.027664 given the other variables in the model are held constant. When inspecting the MONTH predictor in the 3rd case of crash severity (i.e. fatal relative to property damage) in Table 5.20 for the training dataset, the odd ratio is greater than 1.0 (i.e.1.05751), which indicates that this predictor is positively contributing to the crash severity at this level (i.e. fatal), and it is significant at the 95% confidence as its p-value is less than 0.05.

5.4.4: Significant Risk Factors for Boone County Crashes

The statistically significant risk factors (i.e. predictors or independent variables) of the Boone County road network in Missouri at the 95% confidence level are shown in Table 5.21. For the 1st case of crash severity level (i.e. minor injury relative to property damage), the accident type, the driver drink, the speed, the driver aggressiveness, the rural-urban, the AADT, the grade of the road, and the sight distance of the road are significant at the 95% confidence level. For the 2nd case of crash severity level (i.e. disabled relative to property damage), the hour of the day, the accident type, the driver drink, the driver aggressiveness, the cell-text, the number of lanes, the grade of the road, and the sight distance of the road are significant at the 95% confidence level.

Table 5.21: Significant risk factors for Boone County crashes

Crash Severity Level	BOONE COUNTY, MO Roads	
	Significant Risk Factors	Significant Group Factors
Minor Injury	1. ACC_TYPE 2. DR_DRINK 3. SPEED 4. DR_AGGRESSIVE 5. RURAL_URBAN 6. AADT 7. GRADE_LEVEL 8. SIGHT_DIST	1. Driver Behavior 2. Accident Type 3. Road Geometry 4. Traffic Operation 5. Sight Distance
Disabled	1. HOUR 2. ACC_TYPE 3. DR_DRINK 4. DR_AGGRESSIVE 5. CELL_TEXT 6. NUMBER_LANES 7. GRADE_LEVEL 8. SIGHT_DIST	1. Time 2. Driver Behavior 3. Accident Type 4. Road Geometry 5. Sight Distance
Fatal	1. MONTH 2. NO_VEHICLES 3. DR_DRINK 4. DR_AGGRESSIVE 5. SPEED 6. CELL_TEXT	1. Time 2. Driver Behavior

For the 3rd case of crash severity level (i.e. fatal relative to property damage), the month of the year, the number of vehicles involved in the crashes, the driver drink, the driver aggressiveness, the cell-text, and the speed are significant at the 95% confidence level. It can be seen that two risk factors (i.e. the driver drink, and the driver aggressiveness) are significant at the three crash severity levels (i.e. minor injury, disabled, fatal), indicating the importance of these two risk factors in modeling the severity of crashes along Boone County roads. Some other risk factors are significant at only two levels of crash severity, but not at the third level. These risk factors are, the accident type, cell-text, the speed, the grade of the road, and the sight distance of the road. The number of vehicles involved in the crash, the AADT, and the number of lanes are significant at only one level of crash severity. In term of the significant group of factors, we can see that the driver's behavior group is the most important one as it has been related to the three crash severity levels, whereas the accident type, the road geometry, the time, and the sight distance are next in their importance. The traffic operation (i.e. the AADT) is less important among the other groups as it contributes to only one crash severity level.

5.4.5: Marginal Effects for Crashes Along I-70 Corridor

The marginal effect reflects the impact of a one-unit change of an independent variable (predictor) on the event probability of the dependent variable (keeping all other independent variables constant at their mean values). In MNL, the marginal effect of an explanatory variable (predictor) is the partial derivative of the event probability with respect to the predictor of interest (i.e. the change in the event probability of the dependent variable for a unit change in the predictor), and they could be positive or negative values. Positive values indicate that the predictor would positively contribute to crash severity (i.e. would increase the degree severity of crashes), and negative values indicate that the predictor would negatively contribute to crash severity (i.e. would decrease the degree severity of crashes). The marginal effect for a dummy or discrete independent variable is the difference of the predicted probability values at their different levels (Long and Freese 2014). The marginal effects for the I-70 corridor (for both training and testing data) are obtained using Stata 14 and reported in Table 5.22. It can be seen from the table that some predictors have higher marginal effects than others. For instance, the driver drink predictor has a marginal effect of 15.56 % for training data, and 16.07% for testing data. These values present the difference of the event probability of the crash severity when drivers using the road being drunk and not drunk.

Table 5.22: Marginal effects for crashes along I-70

Variable Name	Variable Subgroup	% Marginal Effect	
		I-70 Training	I-70 Testing
GRADE_LEVEL	grade level	3.22 - 1.58	3.62 - 1.74
NUMBER_LANES	one lane two lanes three lanes four lanes five lanes six lanes or more	1.06 2.05 - 2.28 - 2.94 1.31 0.42	1.23 2.16 - 2.77 - 2.49 1.53 0.22
RURAL_URBAN	rural urban	1.97 - 1.56	2.31 - 1.81
CZONE	n/a	1.71	2.33
G_i*	high-high (HH) low-low (LL) random	5.67 4.12 - 2.19	5.72 4.38 - 1.96
SIGHT_DIST	n/a	2.88	2.41
AADT	n/a	1.92	1.72
HOUR	n/a	1.74	2.09
DAY_WEEK	Sun Mon Tues Wed Thurs Fri Sat	- 2.02 2.31 - 2.09 - 1.65 - 1.38 3.15 2.88	- 1.79 1.84 - 1.98 - 1.43 - 1.17 3.37 2.49
MONTH	n/a	1.67	1.89
DIRECTION	east west	1.47 1.31	1.52 1.36

LIGHT_COND	Daylight Dark, lighted Dark, unlighted	- 0.43 - 0.79 0.59	- 0.23 - 0.62 0.44
DR_AGE	Less than 21 years from (21- 64) years more than 64 years	2.58 - 1.87 2.49	2.87 - 1.63 2.61
VEH_TYPE	passenger car motorcycle truck	- 1.62 2.16 - 1.79	- 1.44 2.06 - 1.48
NO_VEHICLE	one vehicle two vehicles three vehicles four vehicles five vehicles six or more vehicles	9.58 14.54 13.17 14.39 13.33 15.17	10.62 15.87 13.16 15.04 13.94 14.81
ACC_TYPE	animal fixed object overturn pedestrian vehicle in transport	1.78 7.06 8.39 7.17 7.38	2.19 6.48 7.79 7.36 7.27
DR_DRINK	n/a	-15.56	-16.07
SPEED	n/a	-8.04	-10.12
DR_AGGRESSIVE	n/a	-8.84	-8.41
CELL_TEXT	n/a	-12.54	-14.17

In other words, if all the drivers that use the I-70 corridor in MO were not in intoxicated conditions, then the probability of crash severity at the I-70 corridor would decrease by 15.56% using training data and 16.07% using testing data. The speed predictor has a marginal effect of 8.04 % for training data, and 10.12% for testing data. These values present the difference of the event probability of the crash severity when drivers using the road are speeding and not speeding so that the crash severity would decrease by (8.04% using training data and 10.12% using testing data) if all drivers were not speeding. The cell-text predictor has a marginal effect of 12.54% for training data, and 14.17% for testing data. These values present the difference of the event probability of the crash severity when drivers are using the cell phones and/or texting during the driving and not using them so that the crash severity would decrease by 12.54% using training data and 14.17% using testing data if all drivers were not using cell-text when driving. The *Gi** predictor relative to high spatial autocorrelation (HH) crashes has a marginal effect of 5.67% for training data, and 5.72% for testing data. Meaning that crashes with HH spatial autocorrelation would increase the severity by 5.67% using training data and 5.72% using testing data. The *Gi** predictor relative to low spatial autocorrelation (LL) crashes has a marginal effect of 4.12% for training data, and 4.38% for testing data. Meaning that crashes with LL spatial autocorrelation would increase the severity by 4.12% using training data and 4.38% using testing data. The *Gi** predictor relative to insignificant random crashes has a marginal effect of -2.19% for training data, and -1.96% for testing data. Meaning that random spatial autocorrelation crashes would decrease the severity by 2.19% using training data and 1.96% using testing data. The number of vehicles involved (assuming one vehicle) in the crash has a marginal effect of 9.58% for training data, and 10.62% for testing data. Meaning that if only one vehicle is involved in the crash, then it would increase the severity by 9.58% using training data and 10.62% using testing data. However, if the number of vehicles involved were increased to two vehicles, then this would increase the severity by 14.54% using training data and 15.87% using testing data. If the number of vehicles increased to three vehicles, then this would increase the severity by 13.17% using training data and 13.16% using testing data. If the number of vehicles further increased to four vehicles, then this would increase the severity by 14.39% using training data and 15.04% using testing data. The sight distance predictor has a marginal effect of 2.88 % for training data, and 2.41% for testing data. These values present the difference of the event probability of crash severity if all road segments were adequate in their decision sight distance relative to AASHTO (2011) standards. Meaning that segments that may have visibility issues would increase the severity by 2.88% using training data and 2.41% using testing data. The accident type predictor (ACC_TYPE) relative to an animal has a marginal effect of 1.78% for training data and 2.19% for testing data. Meaning if an animal would have caused the accident, then this would increase the severity by 1.78% using training data and 2.19% using testing data. However, the accident type predictor relative to a fixed object has a marginal effect of 7.06% for training data and 6.48% for testing data. Meaning if a fixed object (such as a tree or a traffic sign) would have caused the accident, then this would increase the severity by 7.06% using training data and 6.48% using testing data. However, the accident type predictor relative to an overturn has a marginal effect of 8.39% for training data and 7.79% for testing data. Meaning if an overturn was the accident type, then this would increase the severity by 8.39% using training data and 7.79% using testing data. Similarly, the accident type predictor relative to a pedestrian has a marginal effect of 7.17% for training data and 7.36% for testing data. Meaning if a pedestrian would have caused the accident, then this would increase the severity by 7.17% using training data and 7.36% using testing data. In similar manner, the accident type predictor relative to a vehicle in transport has a marginal effect of 7.38% for training data and 7.27% for testing data. Meaning if a vehicle in transport would have caused the accident, then this would increase the severity by 7.38% using training data and 7.27% using testing data.

5.4.6: Marginal Effects for Boone County Crashes

The marginal effects for crashes along Boone County roads (for both training and testing data) are obtained using the Stata 14 and are reported in Table 5.23. It can be seen from the table that some predictors have higher marginal effects than others. For instance, the driver drink predictor has a marginal effect of 16.67% for training data, and 16.96% for testing data. These values present the difference of the probability of the crash severity when drivers using the road being drunk and not drunk.

Table 5.23: Marginal effects for Boone County crashes

Variable name	Variable Subgroup	% Marginal Effect	
		Boone Training	Boone Testing
GRADE_LEVEL	grade	6.17	6.22
	level	- 0.73	- 1.38
NUMBER_LANES	one lane	3.27	2.79
	two lanes	2.14	1.87
	three lanes	2.49	1.81
	four lanes	- 1.72	- 2.33
	five lanes	- 2.51	- 1.83
	six lanes or more	- 2.65	- 3.31
RURAL_URBAN	rural	3.86	3.27
	urban	- 0.69	- 1.36
CZONE	n/a	2.18	2.27
G_i*	high-high (HH)	2.72	1.85
	low-low (LL)	1.82	2.53
	random	- 2.11	- 2.54
SIGHT_DIST	n/a	4.27	4.13
AADT	n/a	2.48	2.32
HOUR	n/a	2.17	2.52
DAY_WEEK	1 - Sun	1.13	1.64
	2 - Mon	2.21	1.69
	3 - Tues	- 1.59	- 1.17
	4 - Wed	- 1.83	- 1.48
	5 - Thurs	- 1.61	- 1.29
	6 - Fri	2.05	2.36
	7 - Sat	1.74	2.17
MONTH	n/a	1.41	1.52

DIRECTION	east west north south	- 0.79 - 0.94 1.16 1.44	- 1.15 - 0.48 1.53 1.32
LIGHT_COND	Daylight Dark, lighted Dark, unlighted	- 0.88 - 0.34 2.08	- 1.37 - 0.61 1.58
DR_AGE	Less than 21 years from (21- 64) years more than 64 years	3.24 - 0.64 3.44	2.69 - 1.83 3.72
VEH_TYPE	passenger car motorcycle truck	- 0.95 3.12 - 1.66	- 1.22 2.58 - 1.29
NO_VEHICLE	one vehicle two vehicles three vehicles four vehicles five vehicles six or more vehicles	4.37 6.61 5.54 7.37 7.62 7.39	5.75 7.18 6.04 8.16 6.88 5.47
ACC_TYPE	animal fixed object overturn pedestrian vehicle in transport	2.22 4.62 6.04 5.39 6.69	1.52 5.19 6.79 6.47 7.45
DR_DRINK	n/a	-16.67	-16.96
SPEED	n/a	-7.31	-8.53
DR_AGGRESSIVE	n/a	-14.66	-13.68
CELL_TEXT	n/a	-11.78	-13.04

That is, if all the drivers that use the Boone County roads were not in intoxicated conditions, then the probability of crash severity at Boone roads would decrease by 16.67% using training data and 16.96% using testing data. The speed predictor has a marginal effect of 7.31% for training data, and 8.53% for testing data. These values present the difference of the event probability of the crash severity when drivers using the road are speeding and not speeding so that the crash severity would decrease by (7.31% using training data and 8.53% using testing data) if all drivers were not speeding. The cell-text predictor has a marginal effect of 11.78% for training data, and 13.04% for testing data. These values present the difference of the event probability of the crash severity when drivers are using the cell phones and/or texting during the driving and not using them so that the crash severity would decrease by 11.78% using training data and 13.04% using testing data if all drivers were not using cell-text when driving. The *Gi** predictor relative to high spatial autocorrelation (HH) crashes has a marginal effect of 2.72% for training data, and 1.85% for testing data. Meaning that crashes with HH spatial autocorrelation would increase the severity by 2.72% using training data and 1.85% using testing data. The *Gi** predictor relative to low spatial autocorrelation (LL) crashes has a marginal effect of 1.82% for training data, and 2.53% for testing data. Meaning that crashes with LL spatial autocorrelation would increase the severity by 1.82% using training data and 2.53% using testing data. The *Gi** predictor relative to insignificant random crashes has a marginal effect of -2.11% for training data, and -2.54% for testing data. Meaning that random spatial autocorrelation crashes would decrease the severity by 2.11% using training data and 2.54% using testing data. The number of vehicles involved (assuming one vehicle) in the crash has a marginal effect of 4.37% for training data, and 5.75% for testing data. Meaning that if only one vehicle is involved in the crash, then it would increase the severity by 4.37% using training data and 5.75% using testing data. However, if the number of vehicles involved were increased to two vehicles, then this would increase the severity by 6.61% using training data and 7.18% using testing data. If the number of vehicles increased to three vehicles, then this would increase the severity by 5.54% using training data and 6.04% using testing data. If the number of vehicles further increased to four vehicles, then this would increase the severity by 7.37% using training data and 8.16% using testing data. The sight distance predictor has a marginal effect of 4.27% for training data, and 4.13% for testing data. Meaning that segments that may have visibility issues would increase the severity by 4.27% using training data and 4.13% using testing data. The accident type predictor (ACC_TYPE) relative to an animal has a marginal effect of 2.22% for training data and 1.52% for testing data. Meaning if an animal would have caused the accident, then this would increase the severity by 2.22% using training data and 1.52% using testing data. However, the accident type predictor relative to a fixed object has a marginal effect of 4.62% for training data and 5.19% for testing data. Meaning if a fixed object (such as a tree or a traffic sign or wall fence) would have caused the accident, then this would increase the severity by 4.62% using training data and 5.19% using testing data. However, the accident type predictor relative to an overturn has a marginal effect of 6.04% for training data and 6.79% for testing data. Meaning if an overturn was the accident type, then this would increase the severity by 6.04% using training data and 6.79% using testing data. Similarly, the accident type predictor relative to a pedestrian has a marginal effect of 5.39% for training data and 6.47% for testing data. Meaning if a pedestrian would have caused the accident, then this would increase the severity by 5.39% using training data and 6.47% using testing data. In similar manner, the accident type predictor relative to a vehicle in transport has a marginal effect of 6.69% for training data and 7.45% for testing data. Meaning if a vehicle in transport would have caused the accident, then this would increase the severity by 6.69% using training data and 7.45% using testing data.

CHAPTER 6: CONCLUSIONS

6.1: Conclusion of Results

Modeling crash severity is very important in highway safety, as it can help to establish linkages between crash severity levels and associated risk factors such as driver behavior, vehicle characteristics, roadway geometry, and road-environment conditions. This dissertation examined three transportation systems within the State of Missouri: 1) Interstate I-70 corridor; 2) Boone County roads, and 3) MO Route 5. The study sites of I-70 corridor and Boone County roads were used to model crash severity along both of them, while MO Route 5 was used to locate passing and no-passing zones along it. Missouri crash data as recorded in the Missouri Statewide Traffic Accident Records System (STARS) were analyzed using three years' crashes (2013-2015). The total number of the observed crashes within the three years' period was 5869.0 for I-70 corridor and 2348.0 for Boone County roads. The response variable (i.e. crash severity) was modeled with four outcome categories: 1) property-damage-only; 2) minor-injury; 3) disabling-injury; and 4) fatal injury. In order to explore potential ways in which crash severity models can be improved to better overcome limitations of traditional modeling approaches that assume all observations are independent of each other, this dissertation developed a detailed framework for detecting temporal and spatial autocorrelation in crash data. In addition, an approach for evaluating the sight distance available to drivers along roadways was also proposed. Finally, a crash severity model was utilized using a multinomial logistic regression approach that incorporates the available sight distance and spatial autocorrelation as potential risk factors, in addition to a wide range of other factors related to road geometry, traffic volume, driver's behavior, environment, and vehicles. The temporal autocorrelation was thoroughly investigated among the time independent variables in crash data using several test statistics to detect the amount of temporal autocorrelation and whether it's significant in crash data. The tests employed were: a) Durbin-Watson (*DW*) test; b) Breusch-Godfrey (*LM*) test; and c) Ljung-Box Q (*LBQ*) test. The removal of any significant temporal autocorrelation in crash data was presented using a) the differencing procedure; and b) Cochrane-Orcutt method. The analysis of the road sight distance was performed by a GIS-based approach using viewshed tools using the AASHTO (2011) sight distance criteria. In addition, it was also used to identify both passing zones and no-passing zones along MO Route 5. The viewshed analysis showed that the available stopping sight distance at both I-70 corridor and Boone roads was adequate, however, eleven segments along I-70 corridor were identified with potential visibility issues as they were not conforming to AASHTO (2011) decision sight distance, and thirteen segments along Boone roads were also identified with potential visibility issues relative to AASHTO (2011) decision sight distance. The decision sight distance at these segments for both I-70 corridor and Boone County roads was used as a potential risk factor in modeling crash severity. In exploring the spatial autocorrelation of crashes, two indices of spatial autocorrelation were utilized: Moran's *I* and Getis-Ord *Gi** statistic. Then the integration of *Gi** statistic as a potential risk factor in crash modeling was explored for the first time in the literature. This dissertation also introduced a new hybrid method for assessing the spatial autocorrelation by combining both Moran's *I* and *Gi** statistic to examine the spatial clustering patterns of crashes. the Global Moran's *I* and the General *Gi** for MO I-70 and Boone roads indicated the existence of significant spatial autocorrelation of the overall crashes at these sites. To further identify the type and extension of spatial autocorrelation among crashes within I-70 and Boone roads, the Anselin local Moran's *I* and the local *Gi** indices were employed. These indices identified crashes with significant high spatial autocorrelation, crashes with significant low spatial autocorrelation, significant outliers, and insignificant random crashes. The multinomial logistic regression (MNL) approach was used to model the relationships between the dependent variable (i.e. crash severity) and risk factors that were included in Missouri crash data for both MO I-70 and Boone roads. The odd ratios of the MNL were used to interpret the results of crash severity. For I-70 corridor, the results showed that the significant risk factors that contributed to crash severity were: the driver drink; the number of vehicles involved in the crash; the accident type; the speed; the driver aggressiveness; the use of cell-text; and the spatial autocorrelation index *Gi**. For Boone roads, the results showed that the significant risk factors that contributed to crash severity were: the driver drink; the accident type; the speed; the driver aggressiveness; the use of cell-text; and the sight distance.

6.2: Potential Impacts of Other Risk Factors

There are other potential risk factors that could be introduced into this modeling framework in the future, such as the weather conditions, and the road surface conditions. Both were not existed in the MO crash data used in this dissertation. Adverse weather conditions (i.e. rain, sleet, snow, fog, severe crosswinds, or blowing snow/sand/debris) and/or slick pavement (i.e. wet pavement, snowy/slushy pavement, or icy pavement) can increase the number of crashes and the degree of the crash severity injury (Andrey et al. 2003). Weather conditions can affect driver capabilities, vehicle performance (i.e.

traction, stability and maneuverability), pavement friction, roadway infrastructure, and traffic flow. In the absence of real-time weather data, their impact can be predicted using historic weather data from nearby weather stations to develop information that could help in crash modeling (Han et al. 2003).

6.3: Modeling Crash Points vs. Crash Segments

Modeling road crash data may include crashes presented either as crash points or crash segments along the road. Crash point events are usually presented by their latitude, longitude values (or x, y coordinates), while crash segment events are presented by specified lengths of road segments (also called crash exposure length). Since there are no specific criteria for choosing the crash exposure length, the use of crash points is preferred in crash prediction models as it can eliminate potential errors that could arise from specifying the crash exposure length. The length of crash segment can affect the crash frequency and/or the crash severity at that segment, and researchers have reported that the probability and severity of crashes tend to be smaller on shorter road segments and higher at longer segments (Lord and Bonneson 2007; Milton et al. 2008). In addition, some researchers have indicated that a non-linear relationship may exist between the probability and severity of crashes and the length of a segment that could be pointing to unobserved heterogeneity (Lord and Mannering 2010). In order to avoid any potential crash segment errors, this dissertation used crash points by utilizing the latitude and longitude values of each crash as reported in the STARS data, rather than specifying road segments in the modeling process.

6.4: Effects of Changing DEM's Resolution

In this dissertation, the digital elevation models (DEMs) were used to assess the stopping and decision sight distances along I-70 corridor and Boone roads in Missouri. They were also used to assess the passing sight distance and locating the passing and no-passing zones along MO Route 5. The accuracy of the DEMs depends on the data source and the spatial resolution of DEMs. Research has shown that using higher resolution DEMs can derive more details, but not necessarily offer more accurate results (Claessens et al. 2005; Zhang et al. 2008). The resolution of all DEMs used in this dissertation was 30 meters. However, even if higher resolution DEMs were used in the sight distance analysis, such as 10 meters, this would have been resulted in more cells per unit area, and more surface details, but not necessarily offered more accurate results.

6.5: The Amount of Violation of AASHTO Standards

Although the viewshed analysis used in this dissertation can effectively identify the segments of the roads with potential visibility issues relative to AASHTO standards, however, there is no easy way to determine how much they might be in violation of AASHTO standards. One practical method could be to use the GIS functionality to manually measure the distance spacing between the observer points along a segment, then obtaining an average value of all spacing that would represent the available sight distance at that segment, and then determining the difference between the AASHTO standards and the available sight distance at that segment. For example, if the required AASHTO sight distance is 330 m, and the average spacing between the observer points at a segment that may have visibility issues is found to be 290 m, then the amount of violation would be (330 – 290 = 40 m) of ASSHTO standards.

REFERENCES

AASHTO, (2011). *A policy on geometric design of highways and streets*. The American Association of State Highway and Transportation Officials, Washington, D.C.

AASHTO, (2004). *A policy on geometric design of highways and streets*. The American Association of State Highway and Transportation Officials, Washington, D.C.

Abdel-Aty, M. (2003). Analysis of Driver Injury Severity Levels at Multiple Locations Using Ordered Probit Models. *Journal of Safety Research, 34*(5), 597-603.

Abdelwahab, H.T., & Abdel-Aty, M.A. (2002). Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas. *Accident Analysis and Prevention*, *1784*, 115-125.

Aguero-Valverde, J. (2013). Multivariate spatial models of excess crash frequency at area level: case of Costa Rica. *Accident Analysis and Prevention*, *59*, 365–373.

Aguero-Valverde, J., & Jovanis, P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis and Prevention*, *38*(3), 618–625.

Aguero-Valverde, J., & Jovanis, P. (2008). Analysis of road crash frequency with spatial models. *Transportation Research Record*, *2061*, 55–63.

Amoros, E., Martin, J.L., & Laumon, B. (2003). Comparison of road crash incidents and severity between some French counties. *Accident Analysis and Prevention*, *35*, 537–547.

Anastasopoulos, P.C., & Mannering, F. (2009). A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis and Prevention*, *41*(1), 153–159.

Anderson, T.K. (1984). On the theory of testing serial correlation. *Accident Analysis and Prevention*, *31*, 88–116.

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hot spots. *Accident Analysis and Prevention*, *41*, 359–364.

Andrey, J., Mills, B., & Vandermolen, J. (2003). A Temporal Analysis of Weather-Related Collision Risk for Ottawa, Canada: 1990-1998. *Transportation Research Board, 2003*-3488, Washington, D.C.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic.

Anselin, L. (1992). *Space Stat: A Program for the Statistical Analysis of Spatial Data*. Santa Barbara, CA: National Center for Geographic Information and Analysis, University of California.

Anselin, L. (1995). Local indicators of spatial association- LISA. *Geographic Analysis, 27*(2), 93–115.

Arminger, G., Clogg, C., & Sobel, M. (1995). *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum Press.

Azimi, M., & Hawkins, H. (2013). Algorithm for analyzing horizontal sight distance from lane centerline coordinates. *Transportation Research Record,* (2358), 12-9.

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis*. Harlow, England.: Addison Wesley Longman Limited.

Baker, P., O'Neill, B., Haddon, W., & Long, W, B. (1974). The Injury Severity Score: a method for describing patients with multiple injuries and evaluating emergency care. *The Journal of Trauma, 14* (3), 187–196

Baltagi, B. H. (2011). *Econometrics*. 5th ed. Berlin: Springer.

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C, 15*, 380–391.

Beirness, D.J., & Beasley, E. (2011). *A Comparison of Drug- and Alcohol-involved Motor Vehicle Driver Fatalities*. Canadian Centre on Substance Abuse.

Ben-Akiva, M., & Lerman, R. (1985). *Discrete Choice Analysis: Thoery and application to travel demand*. Cambridge, MA: MIT Press.

Ben-Arieh, D., Chang, S., Rys, M., & Zhang, G. (2004), Geometric modeling of highways using global positioning system data and B-spline approximation, *Journal of Transportation Engineering, 130* (5), 632–636.

Berglund, S., & Karlstrom, A. (1999). Identifying Local Spatial association in Flow Data Geographical Systems. *Transportation Geography, 1*, 219-236.

Bham, G., Javvadi, B., & Manepalli, U. (2012). Multinomial Logistic Regression Model for Single-Vehicle and Multivehicle Collisions on Urban U.S. Highways in Arkansas. *Journal of Transportation Engineering, 138*(6), 786-797

Black, W. (1992). Network autocorrelation in transport network and flow systems. *Geographical Analysis, 24*(3), 207–222.

Black, W.R., & Thomas, I. (1998). Accidents on Belgium’s motorways: a network autocorrelation analysis. *Transport Geography*, *6*(1), 23–31.

Blincoe, J., Miller, R., Zaloshnja, E., & Lawrence, A. (2015). *The economic and societal impact of motor vehicle crashes, 2010.* Washington, DC.

Box, G.E.P., Jenkins, G.M., & Reinsel, G.C. (1994). *Time series analysis – Forecasting and control*. Englewood Cliffs, NJ, USA: Prentice Hall.

Cai, H., & Rasdorf, W. (2008), Modeling Road Centerlines and Predicting Lengths in 3-D Using LIDAR Point Cloud and Planimetric Road Centerline Data. *Computer-Aided Civil and Infrastructure Engineering, 23* (8), 157-173.

Caliendo, C., Guida, M., & Parisi, A. (2007). A crash-prediction model for multilane roads. *Accident Analysis and Prevention, 39* (4), 657-670.

Cameron, A.C., & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

CAPS (2016). The Center for Advanced Public Safety. Retrieved November 27, 2016, from <http://www.caps.ua.edu/analytics/downloads/datasets/public/>.

CDC (2016). Centers for Disease Control and Prevention-Global Road Safety. Retrieved November 27, 2016, from <http://www.cdc.gov/motorvehiclesafety/global/index.html>.

Chang, L.Y. (2005). Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Accident Analysis and Prevention, 43*(8), 541–557.

Chang, L.Y., & Wang, H. (2006). Analysis of traffic injury severity: an application of nonparametric classification tree techniques. Accident Analysis and Prevention 38 (5), 1019–1027.

Chatfield, C. (1996). *The analysis of time series – an introduction* (5th ed.). London, UK: Chapman and Hall, CRC.

Chiou, Y.C., Fu, C., & Chih-Wei, H. (2014). Incorporating spatial dependence in simultaneously modeling crash frequency and severity. *Analytic Methods in Accident Research, 2*, 1-11.

Claessens, L., Heuvelink, M., Schoorl, M., & Veldkamp, A. (2005). DEM resolution effects on shallow landslide hazard and soil redistribution modeling. *Earth Surface Processes and Landforms, 30*, 461-477

Cliff, A, D., & Ord, K. (1975). *The choice of a test for spatial autocorrelation*. London, UK: Pion.

Cliff, A, D., & Ord, K. (1981). *Spatial processes: models and applications*. London, UK: Pion.

Cochrane, D. & Orcutt, G.H. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, 44, 32-61.

Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*. Second Edition. Chapman & Hall.

Daniels, S., Brijs, T., Nuyts, E., & Wets, G. (2010). Explaining variation in safety performance of roundabouts. *Accident Analysis and Prevention*, *42*(2), 292–402.

Data.gov. (2016). US Government Data. Retrieved November 27, 2016, from < http://catalog.data.gov/dataset/preliminary-accidentincident-data-daily-data-file>.

Delen, D., Sharada, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis and Prevention, 38*, 434–444.

Ehlert, A., Bell, M.G.H., & Grosso, S. (2006). The optimization of traffic count locations in road networks. *Transportation Research Part B, 40*, 460–479.

El-Basyouny, K., & Sayed, T. (2006). Comparison of two negative binomial regression techniques in developing accident prediction models. *Transportation Research Record, 1950*, 9–16.

El-Basyouny, K., & Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis and Prevention 41*(4), 820–828.

Elvik, R. (2006). Laws of accident causation. *Accident Analysis and Prevention*, *38* (4), 742–747.

Erdogan, S., I. Yilmaz, T. Baybura, & M. Gullu. (2008). Geographical information systems aided traffic accident analysis system case study: City of Afyonkarahisar. *Accident Analysis & Prevention 40(1)*: 174-181.

Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Accident Analysis and Prevention*, *40*, 341–351.

ESRI (2016). ArcGIS Resources Center. Retrieved November 27, 2016, from < http//resources.arcgis.com/en/help/main/10.1/index>.

ESRI (2016 a). ArcGIS Resources Center. Retrieved November 27, 2016, from < http://resources.esri.com/help/10.1/ArcGISEngine/java/Gp_ToolRef/Spatial_Statistics_tools/how_hot_spot_analysis_colon_getis_ord_gi_star_spatial_statistics_works.htm>.

Fagerland, M. W., Hosmer, D. W., & Bofin, A. M. (2008). Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine 27*: 4238–4253.

Fagerland, M. W., & Hosmer, D.W. Jr. (2012). A generalized Hosmer Lemeshow goodness-of-fit test for multinomial logistic regression models. *Stata Journal 12*: 447–453.

Fischer, M. M., & Wang, J. (2011). *Spatial Data Analysis: Models, Methods, and Techniques*. New York: Springer.

Flahaut, B. (2004). Impact of infrastructure and local environment on road unsafety: Logistic modeling with spatial autocorrelation. *Accident Analysis & Prevention, 36(6)*, 1055-1066.

Fotheringham, A.S., Brunsdon, C., & Charlton, M.E. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationship*. Chichester: Wiley.

Fraser, S. (2007). *The use of floating cellular telephone data for real-time transportation incident management.* Canada: McMaster University.

Freese, J., & Long, J. S. (2000). Tests for the multinomial logit model. *Stata Technical Bulletin, 10*, 247–255. College Station, TX: Stata Press.

Fricker, J., & Kumapley, R. (2002). *Updating Procedures to Estimate and Forecast Vehicle-Miles Traveled.* USA: Purdue University.

Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician 5*, 115-45

Geedipally, R., Lord, D., & Dhavala, S. (2012). The negative-binomial Lindley generalized linear model: characteristics and application using crash data. *Accident Analysis and Prevention*, *45*(2), 258–265.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel-Hierarchical Models.* UK: Cambridge University Press.

Getis, A. & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*, 189-206.

Getis, A. & Ord, K. (1996). *Local spatial statistics: an overview.* Geo Information International: Cambridge, England.

Glenberg, A. (1996). *Learning from Data: An Introduction to Statistical Reasoning (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Glennon, J. C. (1998). New and Improved Model of Passing Sight Distance on Two-Lane Highways. *Transportation Research Record 1195*. TRB, National Research Council, Washington, DC.

Goldstein, H. (1995). *Multilevel Statistical Models (2nd ed.)*. New York, Edward Arnold.

Goodchild, M, F. (1987). Spatial autocorrelation. *Concepts and Techniques in Modern Geography*, *48*, 56-63.

Goovaerts, P., & Jacquez, M. (2005). Detection of temporal changes in the spatial distribution of cancer rates using local Moran's I and geostatistically simulated spatial neutral models. *Journal of Geographical Systems*, *7*, 137–159.

Greene, W. (2008). *Econometric Analysis*. *6th ed*. Upper Saddle River, NJ: Prentice-Hall.

Greene, W. (2012). *Econometric Analysis*. *7th ed.* Upper Saddle River, NJ: Prentice Hall.

Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis and Prevention*, *35*, 273–285.

Greiling, D. A., & Jacquez, G. M. (2005). Spacetime visualization and analysis in the Cancer Atlas Viewer. *Geographical System, 7*, 67–84.

Griffith, D. A. (1987). *Spatial Autocorrelation: A Primer*. Resource Publications in Geography, the Association of American Geographers: Washington, DC.

Gujarati, D. (1992). *Essentials of Econometrics*. NY: McGraw-Hill.

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

Gundogdu, I. B. (2010). Applying Linear Analysis Methods to GIS-supported Procedures for Preventing Traffic Accidents: Case Study of Konya. *Safety Science, 48(6),* 763-769.

Hadayeghi, A., Shalaby, A.S., & Persaud, B.N. (2010). Development of planning level transportation safety tools using geographically weighted Poisson regression. *Accident Analysis and Prevention*, *42*(2), 676–688.

Han, D., Chin, S., & Hwang, H. (2003). Estimating Adverse Weather Impacts on Major U.S. Highway Network. *Transportation Research Board,* Washington, D.C.

Hassan, Y., Easa, S. M., & Abd El Halim, A. O. (1996). Passing Sight Distance on Two-Lane Highways: Review and Revision. *Transportation Research Part A,* 30 (6).

Hauer, E. (1992). Traffic conflicts and exposure. *Accident Analysis and Prevention, 14,* 359-364.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica 46*: 1251–1271.

Hausman, J. A., & McFadden, D. (1984). Specification Tests for the Multinomial Logit Model. *Econometrica, 52*, 1219-1240.

Hilbe, J. (2007). *Negative Binomial Regression*. UK: Cambridge University Press.

Hilbe, J. (2014). *Modeling Count Data*. Cambridge University Press.

Hosmer, D. W., Lemeshow, S. A., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Hox, J. (2002). *Multilevel Analysis, Techniques and Applications*. London, UK: Lawrence Erlbaum Associates.

Hoyle, R. (1995). *Structural Equation Modeling: Concepts, Issues, and Applications*. Thousand Oaks, CA., Sage Publications.

HSIS (2016). The Highway Safety Information System. Retrieved November 27, 2016, from <http://www.hsisinfo.org/index.cfm>.

Judge, G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., & Lee., T. C. (1985). *The Theory and Practice of Econometrics*. *2nd ed*. New York: Wiley.

Khan, X. Q. & D. A. Noyce. (2008). Spatial Analysis of Weather Crash Patterns. *ASCE Journal of Transportation Engineering 134(5)*, 191-202.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression: A Self-Learning Text*. *3rd ed.* New York: Springer.

Kim, Y., Rana, S., & Wise, S. (2004). Exploring Multiple Viewshed Analysis Using Terrain Features and Optimization Techniques. *Computers and Geosciences, 30(9)*, 1019.

Kim, D. G., Lee, Y., Washington, S., & Choi, K. (2007). Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. *Accident Analysis and Prevention, 39*(1), 125–134.

King, M.L. (1981). The alternative Durbin-Watson test: An assessment of Durbin and Watson’s choice of statistic. *Journal of Econometrics, 17,* 51–66.

King, M.L. (1983). The Durbin-Watson test for serial correlation: Bounds for regressions using monthly data. *Journal of Econometrics, 21*, 357–366.

Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics, 34*, 1–14.

Lee, J. & Wong, D. W. S. (2005). *Statistical Analysis with ArcView GIS and ArcGIS*. J. Wiley & Sons, Inc.: New York.

Lee, J., & Abdel-Aty, M. (2008). Presence of passengers: does it increase or reduce driver's crash potential? *Accident Analysis and Prevention 40 (5)*, 1703–1712.

Lemeshow, S. A., & Hosmer, Jr. D. W. (1982). A review of goodness of fit statistics for the use in the development of logistic regression models. *American Journal of Epidemiology 115*: 92–106.

LeSage, J. P., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. New York: Chapman and Hall, CRC.

Leung, Y., & Mei, C. L. (2003). Statistical test for local patterns of spatial association. *Environment and Planning A*, *35*, 725–744.

Levine, N., Kim, K.E., & Nitz, L.H. (1995). Spatial analysis of Honolulu motor vehicle crashes: A Spatial pattern. *Accident Analysis and Prevention*, *27*(5), 663–674.

Li, X., Lord, D., Zhang, Y., & Xie, Y. (2009). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis and Prevention*, *40*(4), 1611–1618.

Li, Z., Wang, W., Bigham, J.M., & Ragland, D.R. (2013). Using geographically weighted Poisson regression for county-level crash modeling in California. *Accident Analysis and Prevention*, *58*, 89–97.

Long, S., (1996). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Long, S., & Freese, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata*. *3rd ed*. College Station, TX: Stata Press.

Lord, D. (2006). Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the Estimation of the fixed dispersion parameter. *Accident Analysis and Prevention*, *46*(5), 751–766.

Lord, D., & Bonneson, A. (2007). Development of accident modification factors for rural frontage road segments in Texas. *Transportation Research Record*, *2023*, 20–27.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash frequency data: a review and assessment of methodological alternatives. *Accident Analysis and Prevention*, *44*(5), 291–305.

Lord, D., & Miranda-Moreno, F. (2008). Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective. *Accident Analysis and Prevention*, *46*(5), 751–770.

Lord, D., & Persaud, N. (2000). Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Transportation Research Record, 1717,* 102–108.

MacNab, Y.C. (2004). Bayesian spatial and ecological models for small-area accident and injury analysis. *Accident Analysis and Prevention, 36*(6), 1028–1091.

Ma, J., Kockelman, K.M., & Damien, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention*, *40*(3), 964–975.

Malyshkina, N., & Mannering, F. (2010). Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. *Accident Analysis and Prevention*, *42*(1), 131–139.

Malyshkina, V., Mannering, F., & Tarko, P. (2009). Markov switching negative binomial models: an application to vehicle accident frequencies. *Accident Analysis and Prevention*, *41*(2), 217–226.

Malyshkina, N., & Mannering, F. (2009). Markov switching multinomial logit model: an application to accident-injury severities. A*ccident Analysis and Prevention 41 (4),* 829–838.

Ma, M., Yan, X., Abdel-Aty, M., Huang, H., & Wang, X. (2010). Safety analysis of urban arterials under mixed-traffic patterns in Beijing. *Transportation Research Record*, *2193*, 105–115.

Manepalli, U., Bham, G., & Kandada, S. (2011). Evaluation of Hotspots Identification Using Kernel Density Estimation (K) And Getis-Ord (Gi*) On I-630. *International Conference on Road Safety and Simulation, TRB,* Indianapolis, US.

Mannering, F., & Grosdsky, L. (1995). Statistical analysis of motorcyclist: perceived accident risk. *Accident Analysis and Prevention*, *27*, 21–31.

Mantel, N., & Bailar, C. (1970). A class of permutational and multinomial test arising in epidemiological research. *Biometrika*, *26*, 687–700.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, *105-142*. Academic Press.

McFadden, D., Tye, W. & Train, K. (1976). An Application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model. *Transportation Research Record, 637*, 39-45.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*: 17-24.

Menard, S. (2002). *Applied Logistic Regression Analysis*. Sage Publications, Inc., Thousand Oaks, CA.

Meng, H., Zheng, L., & Qing, M. (2009). Traffic accidents prediction and prominent influencing factors analysis based on fuzzy logic. *Accident Analysis and Prevention*, *9*(2), 87–92.

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

Miaou, S. P., Song, J.J., & Mallick, B.K. (2003). Roadway traffic crash mapping: A space-time modeling approach. *Accident Analysis and Prevention*, *6*(1), 33– 57.

Milton, J.C., Shankar, V., & Mannering, F. (2008). Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis and Prevention*, *40*(1), 260–266.

Mitra, S., & Washington, S. (2007). On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis and Prevention*, *39*(3), 459– 46.

Mobley, L. R., T. M. Kuo, D. Driscoll, L. Clayton & L. Anselin. (2008). Heterogeneity in mammography use across the nation: separating evidence of disparities from the disproportionate effects of geography. *International Journal of Health Geographics*, *17*(32), 16-26.

Mohamed Abdel-Aty. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. *Accident Analysis and Prevention*, *36*, 54–63.

Mohammadi, M., Samaranayake, V., & Bham, G. (2014). Crash frequency modeling using negative binomial models: An application of generalized estimating equation to longitudinal data. *Accident Analysis and Prevention*, *2*, 52–69.

Mohan, D. (2002). Road Safety in Less-Motorized Environments: Future Concerns. *International Journal of Epidemiology*, *31*, 527–532.

MSDIS (2016). Missouri Spatial Data Information Service, Data by theme. Retrieved November 27, 2016 from < http://msdis.missouri.edu/data/themelist.html>.

MSHP (2016). The Missouri State Highway Patrol, Accident Investigation Reports. Retrieved November 27, 2016 from <http://www.mshp.dps.mo.gov/HP68/static/Official.html>.

Myers, R. (1990). *Classical and Modern Regression with Applications (2nd ed.)*. Belmont, CA, Duxbury Press.

Myers, R., Branas, C., French, C., Nance, L., Kallan, J., Wiebe, J., & Carr, G. (2013). Safety in numbers: are major cities the safest places in the United States? *Annals of Emergency Medicine*, 62(4):408-418.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika 78*: 691-692.

NASS-CDS (2016). The Crash Worthiness Data System. Retrieved November 27, 2016 from < http://www.nhtsa.gov/Data/National+Automotive+Sampling+System+(NASS)/NASS+Crashworthiness+Data+System >.

NASS-GES (2016). The General Estimate System. Retrieved November 27, 2016 from <http://www.nhtsa.gov/Data/National+Automotive+Sampling+System+(NASS)/NASS+General+Estimates+System>.

NCSA (2015). NHTSA- National Center for Statistics and Analysis (NCSA). Retrieved November 27, 2016 from <http://www.nhtsa.gov/NCSA>.

NCSA (2016). NHTSA- National Center for Statistics and Analysis (NCSA). Retrieved November 27, 2016 from <http://www.nhtsa.gov/NCSA>.

Nehate, G. & Rys, M. (2006), 3D Calculation of Stopping-Sight Distance from GPS Data. J*ournal of Transportation Engineering*, *132*(9), 691–698.

NHTSA (2016). U.S. DOT announces steep increase in roadway deaths based on 2015 early estimates. Retrieved November 27, 2016 from< http://www.nhtsa.gov/About+NHTSA/Press+Releases/nhtsa-sees-roadway-deaths-increasing-02052016>.

NHTSA – CODES. (2011). The Crash Data Outcome Evaluation System. Retrieved November 27, 2016 from < http://www nrd.nhtsa.dot.gov/cats/listpublications.aspx?Id=219&ShowBy=Category >.

NHTSA-FARS (2016). The Fatality Analysis Reporting System. Retrieved November 27, 2016 from <http://www.nhtsa.gov/FARS>.

NHTSA – Ruling (2010). The Crash Data Services. Retrieved November 27, 2016 from <http://www.crashdataservices.net/NHTSAruling.html>.

NHTSA-SDS (2016). NHTSA-The State Data System. Retrieved November 27, 2016 from <http://www.nhtsa.gov/Data/State+Data+Programs/SDS+Overview>.

Noland, B., & Oh, L. (2003). The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data. *Accident Analysis and Prevention, 36*, 525–532.

Noland, R.B., & Quddus, M.A. (2004). A spatially disaggregate analysis of road casualties in England. *Accident Analysis and Prevention*, *36*(6), 973–984.

Oh, J., Lyon, C., Washington, S., Persaud, B.N., & Bared, J. (2003). Validation of FHWA crash models for rural intersections: lessons learned. *Transportation Research Record*, *1840*, 41–49.

Oh, J., Washington, S., & Nam, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis and Prevention*, *38*(2), 346–356.

Ord, J. K. & Getis, A., (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis, 27* (4), 286 – 305.

Polus A., Livneh, M. & Frischer B. (2000), Evaluation of the Passing Process on Two-Lane Rural Highways. *Transportation Research Record, 1701*, 53–60.

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

Park, S., & Lord, D. (2007). Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record, 2019*, 1-6.

Persaud, B.N., & Dzbik, L. (1993). Accident prediction models for freeways. *Transportation Research Record*, *1401*, 55–60.

Persaud, B.N., Retting, R.A., & Lyon, C. (2000). Guidelines for the identification of hazardous highway curves. *Transportation Research Record*, *1717*, 14–18.

Pindyck, R. S., & Rubinfeld, D.L. (1981). *Econometric Models and Economic Forecasts*, McGraw-Hill.

Pulugurtha, S. S., Krishnakumar, V.K., & Nambisan, S.S. (2007). New methods to identify and rank high pedestrian crash zones: An illustration. *Accident Analysis & Prevention 39(4)*, 800-811.

Raftery, A. E., & Lewis, S. M. (1992). One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *Statistical Science, 7*, 493–497.

Riviere, C., Lauret, P., Ramsamy, M., & Page, Y. (2006). A Bayesian neural network approach to estimating the energy equivalent speed. *Accident Analysis and Prevention*, *38*(2), 248–259.

Robichaud, K., & Gordon, M. (2003). Assessment of data-collection techniques for highway agencies. *Transportation Research Record*, *1855*, 129–135.

Rosenkrantz, W. (1997). *Introduction to Probability and Statistics for Scientists and Engineers*. NY: McGraw-Hill.

Savolainen, P., Mannering, F., Lord, D., & Quddus, M. (2011). The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident Analysis and Prevention 43(5)*, 1666-1676.

Schweitzer, L. (2006). Environmental justice and hazmat transport: A spatial analysis in southern California, Transportation Research Part D: *Transport and Environment, 11 (6)*, 408-421.

Shaker, A., Yan, W. Y., & Easa, S. (2011), Construction of digital 3D highway model using stereo IKONOS satellite imagery. *Geocartography International*, *26*(1), 49-67.

Shankar, N., Milton, J.C., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability process: an empirical enquiry. *Accident Analysis and Prevention*, *29*, 829–837.

SHRP2-NDS (2016). The Strategic Highway Research Program 2- Naturalistic Driving Study. Retrieved November 27, 2016 from <https://insight.shrp2nds.us/>.

Sliupas, T. (2006). Annual Average Daily Traffic Forecasting Using Different Techniques. *Transport*, *21*, 38–43.

Steenberghen, T., Aerts, K., & Thomas, I. (2010). Spatial clustering of events on a network. *Journal of Transport Geography, 18(3)*, 411-418.

Studenmund, A.H. (2001). *Using Econometrics - A Practical Guide*. Addison-Wesley-Longman.

Tanner, J.C. (1953). Accidents at rural three-way junctions. *Journal of Institution of Highway Engineers*, *11*(2), 56–67.

The US Census Bureau (2016). The Data of the US Census Bureau. Retrieved November 27, 2016 from < http://www.census.gov/en.html>.

Thomas, R.L. (1993). *Introductory Econometrics: Theory and Applications (2nd ed.).* UK, Longman.

Tjur, T. (2009). Coefficients of determination in logistic regression Models-A new proposal: The coefficient of discrimination. *The American Statistician 63*: 366-372.

Tobler, W. R. (1970). A Computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(2), 234–240.

Transport Canada (2016). Road Safety in Canada. Retrieved November 27, 2016 from <http://www.tc.gc.ca/eng/motorvehiclesafety/tp-tp15145-1201.htm>.

Truong, L., & Somenahalli, C. (2011). Using GIS to Identify Pedestrian Vehicle Crash Hot Spots and Unsafe Bus Stops. *Journal of Public Transportation*, *14*: 1-46

Tsai, Y., Hu, Z., & Wang, Z. (2010). Vision-Based Roadway Geometry Computation. *Journal of Transportation Engineering*, *136*(3), 223-233

Tsay, R, S. (2010). *Analysis of Financial Time Series*. Hoboken, NJ, USA: Wiley & Sons.

USDOT (2016). USDOT History of the Transportation in the United States. Retrieved November 27, 2016 from <http://ntl.bts.gov/usdothistory/usdothistory.html>.

Wang, Hao., Zheng Lai., & Meng, Xianghai. (2011). Traffic accidents Prediction Model based on Fuzzy Logic. *Communications in Computer and Information Science, 201*, 101–108.

Wang, X., & Abdel-Aty, M. (2006). Temporal and Spatial Analyses of rear-end crashes at signalized intersections. *Accident Analysis and Prevention*, *38*(6), 1137–1150.

Wang, X., & Kockelman, K. L. (2013). A Poisson-lognormal conditional-auto regressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis and Prevention*, *60*, 71–84.

Warner, R. M. (1998). *Spectral analysis of time-series data*. New York, NY, USA: Guilford Press.

Washington, P., Karlaftis, G., & Mannering, F. (2010). *Statistical and Econometric Methods for Transportation Data Analysis (2nd ed.)*. Boca Raton, FL: Chapman Hall, CRC.

WHO (2015). WHO Global Report on Road Safety 2015. Retrieved November 27, 2016 from<http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/>.

Winston, C., Maheshri, V., & Mannering, F. (2006). An exploration of the offset hypothesis using disaggregate data: the case of airbags and antilock brakes. *Journal of Risk and Uncertainty 32 (2)*, 83–99.

Wood, G.R. (2002). Generalized Linear Accident Models and Goodness of Fit Testing. *Accident Analysis and Prevention*, *34*(4), 417–427.

Wood, N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, Florida: Chapman and Hall, CRC.

Wooldridge, M. (2013). *Introductory Econometrics: A Modern Approach (Fifth ed.)*. Mason, OH: South-Western.

Wong, D. W. S., & Lee, J. (2005). *Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS*. Wiley, Hoboken.

World Bank (2015). The World Bank-Transport for Development. Retrieved November 27, 2016 from <http://blogs.worldbank.org/transport/why-vehicle-safety-matters-crash-related-deaths?cid=EXT_WBBlogSocialShare_D_EXT>.

Xie, Y., & Zhang, Y. (2008). Crash frequency analysis with generalized additive models. *Transportation Research Record*, *2061*, 39–45.

Yamada, I. & Thill, J.C. (2007), Local Indicators of Network-Constrained Clusters in Spatial Point Patterns. *Geographical Analysis, 39*, 268–292.

Yamada, I. & Thill, J.C. (2010). Local Indicators of Network-Constrained Clusters in Spatial Patterns Represented by a Link Attribute. *Annals of the Association of American Geographers. 100(2)*, 269-285.

Yamamoto, T., Hashiji, J., & Shankar, V. (2008). Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accident Analysis and Prevention 40 (4)*, 1320–1329.

Zhang, J. X., Chang, K. T., & Wu, J. Q. (2008). Effects of DEM resolution and source on soil erosion modeling: a case study using the WEPP model. *International Journal of Geographical Information Science*, *22*, 925-942.

VITA

Azad Salim Abdulhafedh holds a Bachelor of Science degree in Civil Engineering from the University of Mosul, Iraq, Department of Civil Engineering.

From 1999-2004, he worked as a Civil Engineer at the United Nations Development Program (UNDP) in Northern Iraq. From 2005-2008, he worked as a senior Project Engineer at the US Army Corps of Engineers (USACE) in Northern, Iraq.

In Spring 2011, he began work toward a Master degree in Civil Engineering at the University of Idaho, Moscow, ID, in the Department of Civil Engineering, and received his Master degree in August 2012.

In Fall 2013, he started his Ph.D. study program at the University of Missouri-Columbia (MU) in the Department of Civil and Environmental Engineering. He completed the requirements for the degree of Doctor of Philosophy in Civil and Environmental Engineering at the MU in December 2016.

PUPLICATIONS:

Abdulhafedh, A. (2016). Prototype Road Surface Management System. *World Journal of Engineering and Technology*, 04,325-334.

Abdulhafedh, A. (2016). Crash Frequency Analysis. *Journal of Transportation Technologies*, 6,169-180.