Traffic Crash Modeling Considering Inconsistent Observations, Interaction Behavior, and Nonlinear Relationships

Yunteng Lao

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Yinhai Wang, Chair

Qiuzi (Cynthia) Chen

John Milton

Program Authorized to Offer Degree:
Department of Civil and Environmental Engineering

University of Washington

Abstract

Traffic Crash Modeling Considering Inconsistent Observations, Interaction Behavior, and Nonlinear Relationships

Yunteng Lao

Chairs of the Supervisory Committee:

Professor Yinhai Wang

Department of Civil and Environmental Engineering

Traffic collisions are a worldwide issue that can cause injury and death, which leads to billions of dollars in damages every year. Significant research efforts have been undertaken to develop and utilize statistical modeling techniques for analyzing the characteristics of crash count data. While these modeling techniques have been providing meaningful outputs, improvements on these modeling methods still need to better understand the crash risk and the contributing factors. Five important issues in crash data modeling are identified in this research. The first two issues are over or under dispersion with crash data and excess zeros within crash records. Considering that they have been well studied in the previous research, this study focuses on the remaining three major issues. The first one is relevant to the partial observations of multiple processes, i.e. crash data may be collected by different agencies that create multiple data sources and may be inconsistent. A modeling mechanism that takes advantage of all datasets for better estimation results is highly desirable. The second one is an interaction issue. Some collisions are single vehicle crashes, such as off-road crashes and rollover incidents, and some collisions involve interaction behavior, such as the Animal-Vehicle Collision (AVC) and the Vehicle-Vehicle

Collision. The characteristics of crashes with interaction behavior are different from those with only one vehicle involved. It is challenging to develop a crash modeling scheme that can capture the interaction behavior. The last one is the nonlinear relationship issue. Most previous collision models are Generalized Linear Model-based (GLM-based) approaches. Such GLM-based approaches are constrained by their linear model specifications because, in most situations, the relationship between the crash rate and its contributing factors are not linear or may not even be monotonic. Thus, finding a way to model the collision data with nonlinear and non-monotonic relationships is of utmost importance.

To address the issues of inconsistent observations, two techniques are developed. A fuzzy logic-based data mapping algorithm is proposed as the first technique to match data from two datasets so that duplicate crash records can be removed when combining these datasets. The membership functions of the fuzzy logic algorithm are established based on survey inputs collected from experts of the Washington State Department of Transportation (WSDOT). As verified by expert judgment collected through another survey, the accuracy of this algorithm was approximately 90%. Applying this algorithm to the two WSDOT datasets relevant to AVC, reported AVC data and the Carcass Removal (CR) data, the combined dataset has 15% –22% more records compared to the original CR dataset. The proposed algorithm is proven effective for merging the Reported AVC data and the CR data, with a combined dataset being more complete for wildlife safety studies and countermeasure evaluations.

The second technique is a diagonal inflated bivariate Poisson regression (DIBP) method. It is an inflated version of bivariate Poisson regression model adopted to directly fit two datasets together. The proposed model technique was also applied to the reported AVC and CR data sets collected in Washington State between 2002 and 2006. The diagonal inflated bivariate Poisson model not only can model paired data with correlation, but also handle under- or over- dispersed data sets. Compared with three other types of models; double Poisson, bivariate Poisson, and zero-inflated double Poisson; the diagonal inflated bivariate Poisson model demonstrates its capability of fitting two datasets with remarkable overlapping portions resulting from the same stochastic process. Therefore, the diagonal inflated bivariate Poisson model provides researchers another new approach to investigating paired data sources from a different perspective.

To address the issues with the interaction issue, a new occurrence mechanism-based probability model, an interaction-based model, which explicitly formulates the interactions between the objects, is introduced. The proposed method was applied to the AVC data and this method can explicitly formulate the interactions between animals and drivers to better capture the

relationships among drivers' and animals' attributes, roadway and environmental factors, and AVCs. Findings of this study show that the proposed occurrence mechanism-based probability model better capture the impact of drivers' and animals' attributes on the AVC. This method can be further developed to model other types of collisions with interaction behavior.

To address the nonlinear relationship issue, a Generalized Nonlinear Model (GNM)-based approach is put forward. The GNM-based approach is developed to utilize a nonlinear regression function to better elaborate non-monotonic relationships between the independent and dependent variables. Previous studies focused mainly on causal factor identification and crash risk modeling using Generalized Linear Models (GLMs), such as Poisson regression, and logistic regression among others. However, their basic assumption of a generalized linear relationship between the dependent variable (for example, crash rate) and independent variables (for example, contributing factors to crashes) established via a link function can often be violated in reality. Consequently, the GLM-based modeling results could provide biased findings and conclusions when the contributing factors have parabolic impact on the crashes. In this research, a GNM-based approach is applied with the rear end accident data and the AVC data collected from ten highway routes starting in 2002 and ending in 2006. For the rear-end collision application, the results show that truck percentage and grade have a parabolic impact: both items increase crash risks initially, but decrease risks after certain thresholds. Similarly, Annual Average Daily Traffic (AADT) and grade also have a parabolic impact on the AVC rate. Such non-monotonic relationships cannot be captured by regular GLM's, which further demonstrates the flexibility of GNM-based approaches in modeling the nonlinear relationship among data and providing more reasonable explanations. The superior GNM-based model interpretations better explain the parabolic impacts of some specific contributing factors and help in selecting and evaluating rear-end crash safety improvement plans.

In Summary, these solutions proposed to address the three major issues in crash modeling are important for crash studies. The fuzzy-logic based data mapping algorithm can combine partial observations from different processes to form up a more complete dataset for a thorough analysis. The diagonal inflated bivariate Poisson models can directly take two data observation processes into account. The occurrence mechanism based probability models and GNM based models are effective methods for handling the interaction issue and non-linear relationships between dependent and independent variables.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

First of all, I would like to thank Professor Yinhai Wang, my advisor and doctoral committee chair for his dedication in helping me with his professional and thoughtful guidance. My accomplishment should be attributed to his valuable mentoring advice, which includes teaching, research, problem solving, and networking. His enthusiasm in academia has fully encouraged me to pursue the same career goal.

I wish to express sincere appreciation to my doctoral committee, Dr. Cynthia Chen, Dr. Alan Borning, and Dr. John Milton for providing comments, which have proved to be very helpful in improving this dissertation.

I would like to thank the many local agencies for their consistent support of my research. One of which is the Washington State Department of Transportation (WSDOT), a major contributor to this dissertation due to my research heavily relying on WSDOT data.

I am also profoundly grateful to all the fellows at University of Washington. I would like to thank to Guohui Zhang and Yao-jan Wu, who have provided me with useful comments on my research. In addition, Xiaolei Ma, Cathy Liu, Yegor Malinovskiy, Runze Yu and the many other members who have discussed and commented on my research. I also would like to express my special thanks to Jonathan Corey, Matthew Palzkill, Omar Abdelbadie, and Matthew K Dunlap, all of whom help me with my English. I enjoyed my time in STAR Lab and we have worked as a

great team.

Last, but not least, I would like to thank my family. I give my deepest gratitude to my parents and sisters, who stay in China, but try their best to support my study in the U.S. Of all people, I am most grateful to my wife, Yang Zhang, who has always been supporting me.

DEDICATION

To my family.

Chapter 1. Introduction

1.1 Problem Statement

Traffic crashes cause injury and death, costing billions of dollars every year. Based on traffic safety facts published by the National Highway Traffic Safety Administration (NHTSA) (2011), over 411,000 people in the USA died in motor vehicle traffic crashes between 2000 and 2009. In 2010, there were 32,885 people killed and 2,239,000 people were injured in the estimated 5,419,000 police-reported motor vehicle traffic crashes (NHTSA, 2012). Although fatalities in roadway traffic accidents have been decreasing since 2005, there were still 93 people killed in 2009 (NHTSA, 2011) and 90 people killed in 2010 (NHTSA, 2012) in motor vehicle crashes each day. This statistical data indicates the importance of improving the existing traffic system for reducing crash frequency and severity.

Transportation agencies have made a lot of effort to improve traffic safety. For example, in 2000, WSDOT began a new Strategic Highway Safety Plan that aims at ending traffic deaths and serious injuries by 2030. To accomplish such a goal, WSDOT has installed 1,237 miles of shoulder rumble strips since May 2003 and additional guardrail for a total cost of $50 million. For further reducing collisions and achieving the Target Zero goal, WSDOT executives adopted a ten year safety investment plan valued at $678 million (Hammond, 2012). To be identified in its

ten-year safety investment plan, WSDOT requires cost effective countermeasures selection and that project proposals are scoped within the limits of existing and likely future funding availability

No matter which type of countermeasures apply, it is essential to properly identify high crash risk locations and the key contributing factors from those locations in order to allocate the limited safety improvement resources to the highest priority sites. This requires having a better understanding of the relationships between crash frequency/severity and its contributing factors. Statistical models can describe these relationships and provide information about the impacts of these contributing factors on crash frequency/severity. Thus, it is critical to develop solid and reliable statistical models for analyzing collisions. The primary goal of this research is to better model the relationships between crash frequency/severity and its contributing factors by provide effective solutions to the key collision modeling issues.

1.2 Collision Modeling Process

The collision modeling process includes four parts: input data, collision analysis, modeling construction, and model outputs. Figure 1-1 details the whole modeling process. Input data for collision data modeling includes collision data, traffic data, weather, roadside, and geometric data. Collision analysis focuses on the analysis of collision mechanism. The characteristics of

multiple object collisions, such as Animal-Vehicle Collisions (AVCs). For a single vehicle collision, the characteristics of the vehicle, roadway and roadside environment, and the driver are the focuses; whereas in the multiple objects involved collisions, the attributes from the object struck, such as the sex and habitat of the animals in the AVC, also need to be considered. Modeling construction builds up the relationships between the crash frequency/severity and the contributing factors to those crashes. These relationships could be linear or nonlinear. Both collision analysis and modeling construction will provide feedback on the data collection process. After constructing the relationships, the model can output the significant coefficients. The contributing factors, collision hotspots, and countermeasures for preventing collisions can be further identified based on the model results.

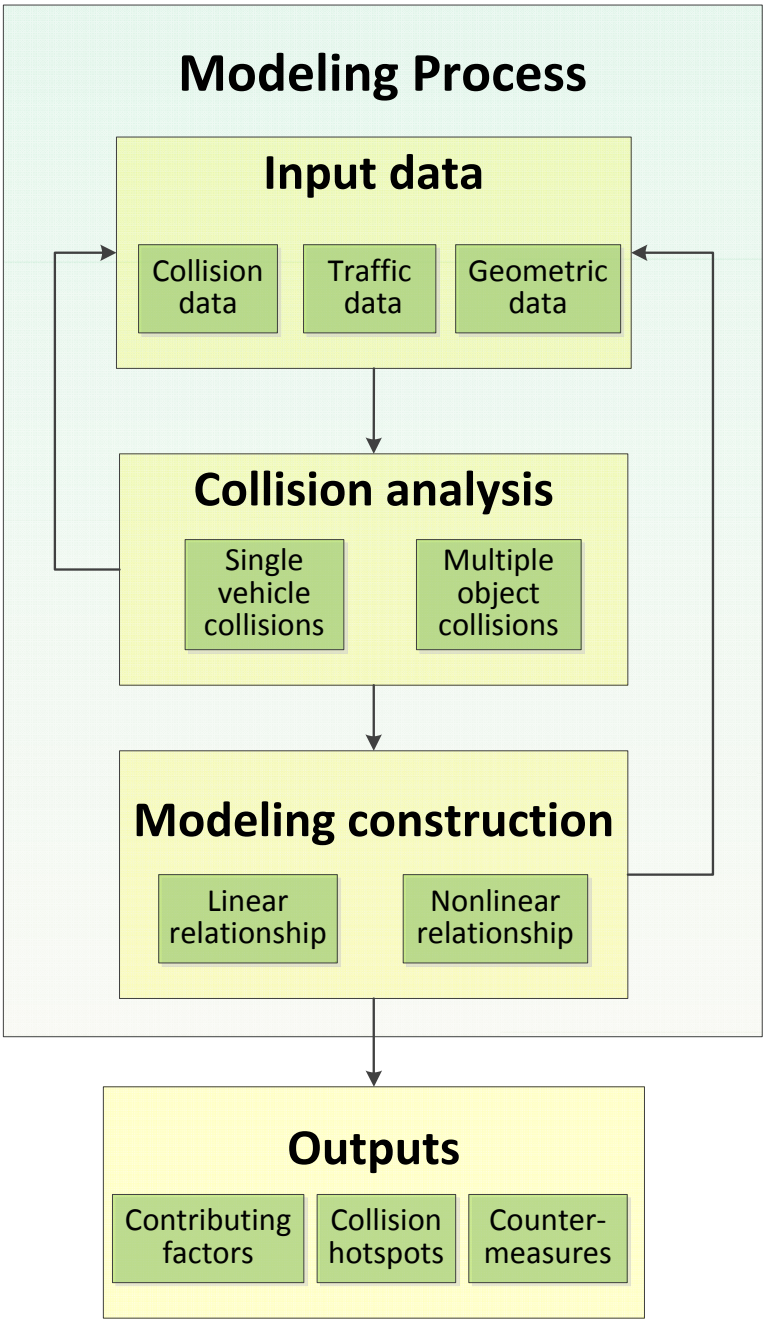


Figure 1-1 Traffic crash modeling process.

1.3 Key Modeling Issues

During the modeling process, several key modeling issues need to be considered. Considering the characteristics of crash data, traffic collisions can be classified as count data, and from a statistical perspective, traffic collisions are discrete, rare events and nonnegative integers. Thus, Poisson regression becomes one of the most suitable solutions for modeling traffic collision. A potential issue with the Poisson regression model is that its sample variance needs to be equal to the sample mean. When the traffic collision variance is significantly bigger (or smaller) than its mean, the collision data is called over-dispersed (or under-dispersed). The Poisson model is inadequate for over-dispersed (or under-dispersed) data whose variance is greater (or smaller) than the mean (Maycock and Hall, 1984; Wang et al., 2003; Lao et al., 2011a). Over-dispersion is a common phenomenon identified by many previous studies (Miaou, 1994; Shankar et al., 1995; Poch and Mannering, 1996; Milton and Mannering, 1998; Wang et. al., 2003; Lao et al., 2011a) and under-dispersion can also occasionally happen when datasets have a very low sample mean due to the many zeros in the data set (Oh et al., 2006). Over-dispersion (or under-dispersion) is one of the key issues in traffic collision data modeling.

The second key issue with the data is the phenomena of an apparent excess of zeros in the collision data. Figure 1-2 shows the single-vehicle truck accidents on 7427 segments of rural two lane highways in Washington State between 2002 and 2005. 6940 out of 7427 segments have 0 zero truck collision in these four years. Collision data with excess of zeros is also a common issue for traffic collision data modeling. Special modeling techniques are necessary for

this phenomenon.

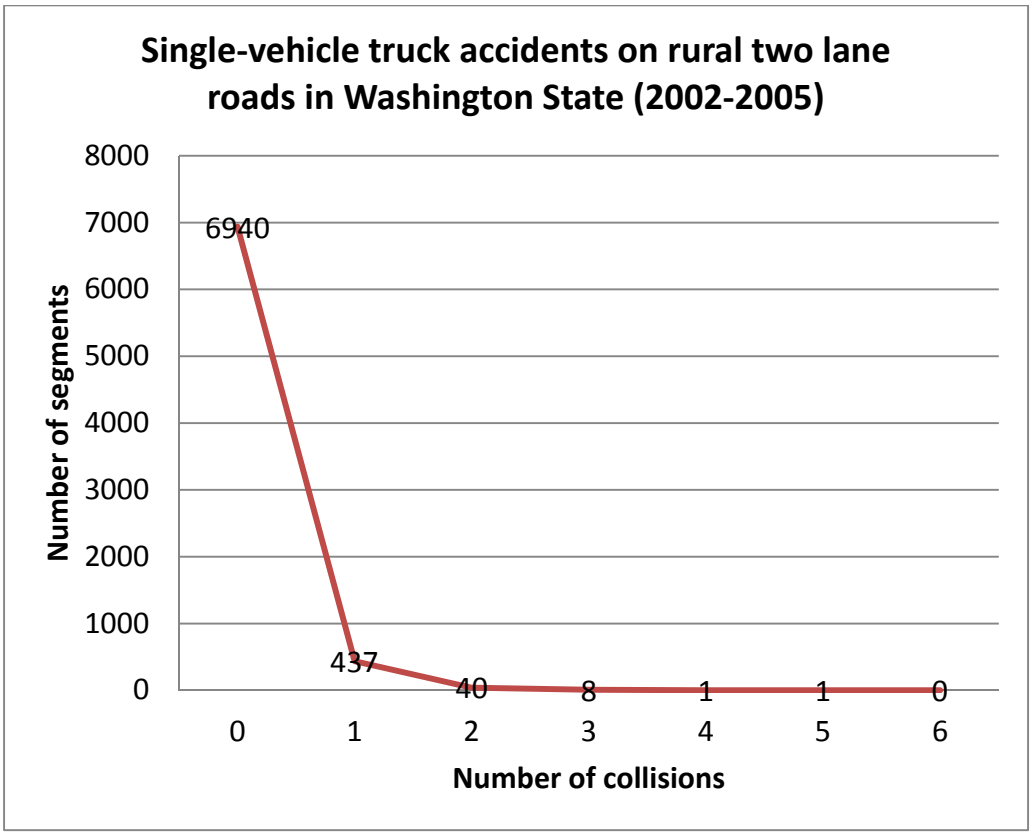


Figure 1-2 Single-vehicle truck accidents on 7427 segments of rural two lane highways in Washington State between 2002 and 2005.

The third issue identified by this research is the issues of inconsistent observations. Inconsistent

observations from different data sources are common phenomena due to human error or other unpredictable factors. The collision data collected from different data sources may not match each other. For example, reported Animal-Vehicle Collision (AVC) data can be extracted from the Highway Safety Information System (HSIS) data. HSIS is operated by the University of North Carolina Highway Safety Research Center and the LENDIS corporation under a contract with Federal Highway Administration (FHWA) (HSIS, 2009). The HSIS collision data of Washington were compiled from both the State Trooper and citizen filed reports. Meanwhile, the carcass removal (CR) data provided by the maintenance team of WSDOT can also provide the AVC information. Figure 1-3 shows the total numbers of records in each data set over a five-year period (2002-2006) on each of ten State Routes (SRs) (US-2, SR-8, US-12, SR-20, I-90, US-97, US-101, US-395, SR-525 and SR-970) with relatively high AVC rates in the past several years. It is obvious that the reported AVC and CR datasets are substantially different. The number of CR records is typically more than that of the Reported AVC data on each route except for US-101. The issue of how to model these data and get better estimation results is important to better utilize the limited resources.

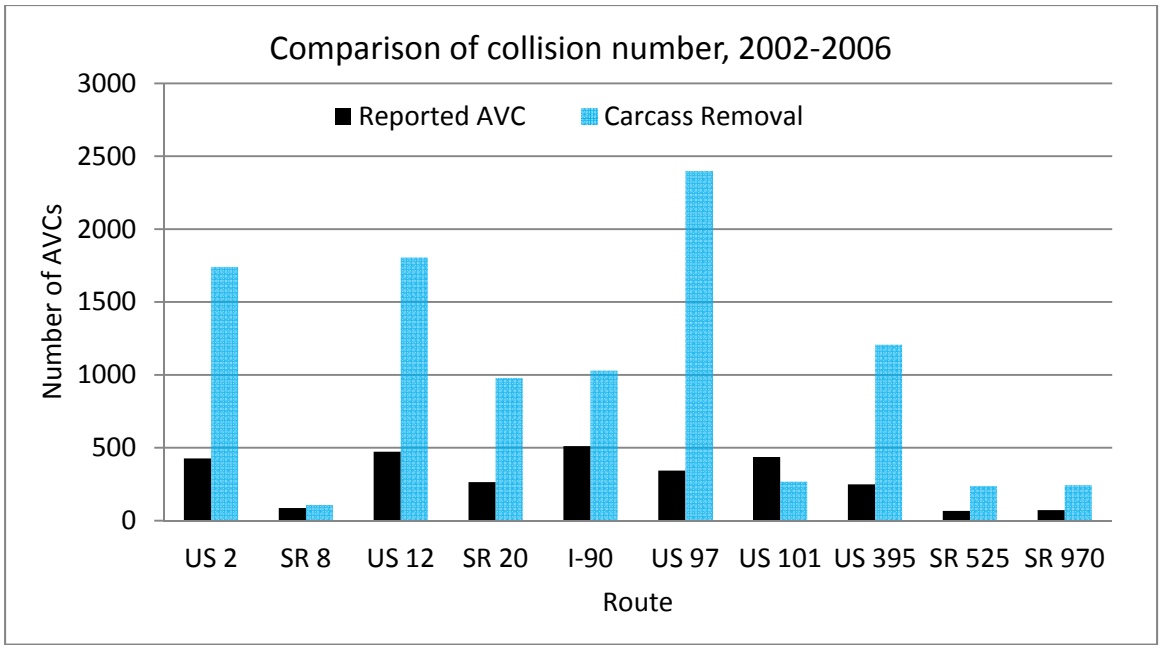


Figure 1-3 Comparisons of total number of records between two datasets for each study route during 2002-2006.

Considering the crash mechanism during the collision analysis process, interaction issue is the fourth issue identified by this research. Some crashes involve only one vehicle, such as the rollover incident, and some crashes involve interaction behavior, such as Pedestrian-Vehicle Collisions, Vehicle-Vehicle Collisions, Bicycle-Vehicle Collisions, and Animal-Vehicle Collisions (AVCs). In regards to the AVCs, when vehicles are approaching animals, the animals will try to run away to avoid the collision. While the animal is moving its position with relation to the car, the car is also attempting to avoid the animal's current location. The crash

characteristics with interaction behavior will differ from the crash with only one single vehicle. Collision models implemented to AVCs with interaction behavior should also consider the reactions between the drivers and the animals. Determining how to describe the collisions with interaction behavior using model technique is critical to better understand the crash characteristics.

Considering the model construction, the nonlinear relationship issue is identified as the last key modeling issue in this research. In most situations, the relationship between the crash rate and its contributing factors will not be linear or may not even be monotonic. Figure 1-4 shows the Animal-Vehicle crash rate (crash frequency per mile) in five years (2002-2006) from ten highways (US2, SR8, US12, SR20, I90, US97, US101, US395, SR525 and SR970) in Washington State. The x-axis is the AADT (in thousands) and the y-axis is the average crash rate. Previous studies using GLM assume the relationship to be in a generalized linear form and show that the expected crash rate should increase with increasing AADT (Chin and Quddus, 2003; Lao et al., 2011). This increasing linear relationship is shown as the dashed line $L(x)$ in Figure 1-4. The R^2 for the fitted line $L(x)$ is nearly equal to zero (R^2=4E-5). Accordingly, $L(x)$ could not reflect the true relationship between the expected crash rate and AADT, assuming all other contributing factors are approximately equal across AADT levels. It is necessary to identify the nonlinear relationship between the expected crash rate and some of its associated factors. Assume that each associated factor x corresponds with a contribution value $U(x)$ related with crash data. Here, the correspondent value $U(x)$ is defined as a nonlinear predictor. For the same associated factor AADT, as shown in Figure 1-4, the solid curve $U(x)$ is a second order

polynomial curve with an R^2=0.63. This curve $U(x)$ does a much better job than $L(x)$ for fitting the expected crash rate. In reality, with a very low AADT, the chance of an animal conflicting with a vehicle is low. Likewise, with a much higher AADT, the animals may be also driven away by the noise of the vehicles. In this situation, GLM-based models could not handle this issue and the predictor $U(x)$ can better represent this non-monotonic relationship between the crash rate and AADT. Therefore, a better modeling technique dealing with the collision data with nonlinear and non-monotonic relationship is essential for better understanding crashes and their contributing factors.

Among the five modeling issues discussed above, the first two issues (dispersion data and excess zeros data) have been well studied and solutions can be found in many previous studies. The remaining three issues have not been discussed in detail and still need further investigation. This research will focus on these last three issues.

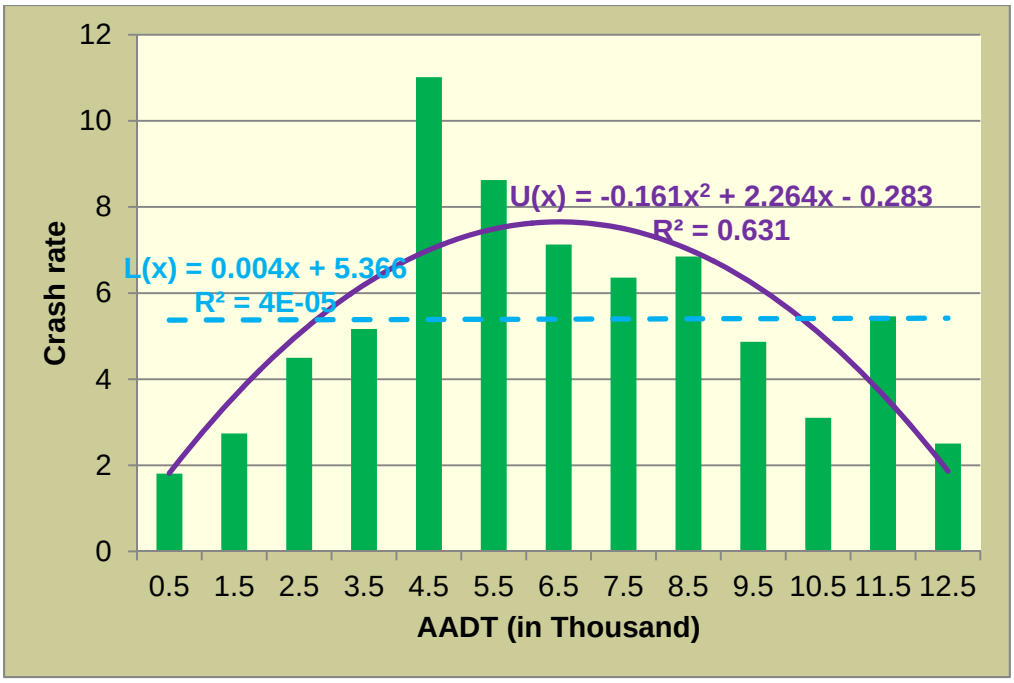


Figure 1-4 Animal-Vehicle crash rate (crash frequency per mile) in five years (2002-2006) from ten highways in Washington State, by AADT.

1.4 Research Objectives

The primary goal of this research is to provide modeling techniques to solve the key issues of the crash data modeling. There are already some well-studied issues, such as over-dispersion and apparent excess of zeros, but there are also three important issues (identified in the Key Modeling Issue section of Chapter 1) that still need to be addressed. To solve these three issues, the following objectives are planned to be addressed in this research:

- Improve data quality and the current crash modeling technique dealing with issues of inconsistent observations;

- Improve the current crash modeling technique dealing with collisions with interaction behavior; and

- Improve current crash modeling methods by introducing a non-linear prediction function in the Generalized Nonlinear Models (GNMs) to describe the relationship between crash and its contributing factors.

1.5 Research Organization

To address the key modeling issues identified in section 1.3, this research will cover three major components: 1) methods dealing with inconsistent observations, 2) models dealing with interaction behavior, and 3) models dealing with nonlinear or even non-monotonic relationships. The remainder of the dissertation is organized in the following manner: Chapter 2 reviews previous studies on the five key issues of modeling collision data. Chapter 3 focuses on data collection for the preformed research and data quality control with the issues of inconsistent observations. Chapter 4 puts forward several modeling technologies for dealing with three key modeling issues: inconsistent observations, interaction issues, and nonlinear relationships.

Chapter 5 to Chapter 7 present the application examples for the modeling technologies described in Chapter 4. Lastly, Chapter 8 summarizes the research effort and provides recommendations for future research.

Chapter 2. Background

Throughout the past, significant research efforts have been made toward roadway traffic accident reduction. A number of studies have been performed to understand the relationships between crashes and potential contributing factors, which includes roadway geometric, environmental, traffic, and human factors (Lao et al. 2011b). For instance, various statistical modeling techniques have been developed to analyze collision characteristics under certain circumstances (Hubbard et al., 2000; Elvik, 2011; Hauer, 2004; Knapp and Yi, 2004; Lord and Mannering 2010). The majority of these models are Generalized Linear Models (GLMs), such as Poisson regression (e.g., Jovanis and Chang, 1986; Miaou and Lum, 1993; Miaou, 1994; Chiou and Fu, 2013), Gamma regression model (Winkelmann and Zimmermann, 1995; Oh et al., 2006), or GLM oriented models, such as negative binomial (NB) regression (or Poisson-gamma regression) (Miaou, 1994; Maher and Summersgill, 1996; Milton and Mannering, 1998; Chin and Quddus, 2003; Wang et. al., 2003; Wang and Nihan, 2004; Donnell and Mason, 2006; Kim et al., 2007; Malyshkina and Mannering, 2010; Daniels et al., 2010; Wei and Lovegrove, 2012; Geedipally et al., 2012; Zou et al, 2013), Random-parameters models (Anastasopoulos and Mannering, 2009; El-Basyouny and Sayed, 2009; Anastasopoulos and Mannering, 2011; El-Basyouny and Sayed, 2011; Anastasopoulos at al., 2012), Bayesian approaches (Deublein et al., 2013) and Bivariate/multivariate models (Park and Lord, 2007; Lao et al., 2011a).

During the modeling procedure, previous research has identified several key modeling issues

including data with dispersion, data with excess zero, inconsistent observations, interaction issues, and the nonlinear relationships. In order to have an overview of previous effects, this chapter is divided into five sections to review the research background related with these modeling issues.

2.1 Modeling with Dispersed Data

The research efforts focused on modeling crash data have resulted in several widely accepted models. The Poisson regression model, one of the most classical and basic methods, has been frequently used to model collision count data (e.g., Jovanis and Chang, 1986, Miaou et al., 1992; Miaou and Lum, 1993; Miaou, 1994; Chiou and Fu, 2013). A well-recognized problem with the Poisson regression model is that the sample variance needs to be equal to the sample mean. In reality, however, most accident data sets do not meet this requirement. Therefore, the Poisson model is inadequate for over-dispersed (or under-dispersed) data whose variance is greater (or smaller) than the mean (Maycock and Hall, 1984; Wang et al., 2003).

2.1.1 Traffic Crash Models for Over-dispersion

In most cases, count data in ecology are "over-dispersed" (Ver Hoef and Boveng, 2007), meaning that the variance is greater than its mean. To address this over-dispersed issue in the collision data, several modeling approaches including NB regression (or Poisson-gamma) (Miaou, 1994; Shankar et al., 1995; Poch and Mannering, 1996; Maher and Summersgill, 1996; Milton and Mannering, 1998; Chin and Quddus, 2003; Wang et. al., 2003; Wang and Nihan, 2004; Lord, 2006; El-Basyouny and Sayed, 2006; Donnell and Mason, 2006; Kim et al., 2007; Malyshkina and Mannering, 2010; Daniels et al., 2010; Wei and Lovegrove, 2012; Geedipally et al., 2012; Zou et al, 2013), Quasi-Poisson regression (Cameron and Trivedi, 1998; Lao et al., 2012b), and Poisson-lognormal regression models (Miaou et al., 2005; Lord and Miranda-Moreno, 2008; Aguero-Valverde and Jovanis, 2008) have been developed and widely applied. Ver Hoef and Boveng (2007) provided comparison between Quasi-Poisson regression and NB regression. The difference between these two methods is that the variance of a NB model is a quadratic function of its mean whereas the variance of a quasi-Poisson model is a linear function of the mean. Gonzales-Barron and Butler (2011) compared NB regression and Poisson-lognormal regression by using six different data sets. Based on their findings, it was determined that Poisson-lognormal regression fitted better with the data sets having high counts, whereas the NB regression performs much better with the low count data sets (13–81% zero counts).

2.1.2 Traffic Crash Models for Under-dispersion

Collision data can also occasionally be under-dispersed. Under-dispersion can be caused by data sets having a very low sample mean due to many zeros in the data set (Oh et al., 2006). In this situation, a Gamma regression model (Winkelmann and Zimmermann, 1995; Oh et al., 2006) can be used to deal with this issue. Subsequently, the Conway-Maxwell-Poisson (COM-Poisson) based models (Shmueli et al., 2005; Kadane et al., 2006; Lord et al., 2008) and Diagonal Inflated Bivariate Poisson (DIBP) regression models (Lao et al., 2011a) are introduced for handling either over or under-dispersed data counts. DIBP models will be described in detail in Chapter 4 in this dissertation.

2.2 Modeling with Excess Zero Data

Another issue in modeling collision data is the phenomena of an apparent excess of zeros. In this situation, zero-inflated models, such as zero-inflated Poisson and zero-inflated NB, have been used for modeling such collision data sets (Shankar et al., 1997; Garber and Wu, 2001; Lee and Mannering, 2002; Kumara and Chin, 2003; Miaou and Lord, 2003; Rodriguez et al., 2003; Shankar et al., 2003; Noland and Quddus, 2004; Qin et al., 2004; and Lord et al., 2005). However, Lord et al. (2005) and Warton (2005) have argued that considerable caution should be exercised when applying zero-inflated models to crash data because a true two-state process may

not exist.

Recently, other new models have been introduced to deal with large zero crash data. For example, the DIBP regression models (Lao et al., 2011a) detailed in Chapter 4 can handle the issue with excess zero. The NB generalized linear model with Lindley mixed effects (NB-L GLM) (Geedipally et al., 2012), was also used to analyze crash data with many zeros. NB-L GLM is based on a recently introduced NB-Lindley (NB-L) distribution (Zamani and Ismail, 2010; Lord and Geedipally, 2011). Based on Geedipally et al.'s research (2012), both NB-L GLM and NB-L distribution performed better than traditional NB for the dataset with large zero crash data.

2.3 Methods Dealing with Inconsistent Data Sources

In some situations, collision data is attained from different data sources. For example, the Animal-Vehicle Collision (AVC) data can be extracted from the Highway Safety Information System (HSIS) (HSIS, 2009). Meanwhile, the carcass removal (CR) data provided by the maintenance team of WSDOT can also provide the AVC information. In most cases, data from different sources do not match with each other well.

Based on the findings of a survey conducted by this study, carcass removal professionals at the

Washington State Department of Transportation (WSDOT) basically agree that over 90% of the carcasses removed from the road are likely struck by a vehicle. Thus, these two datasets should overlap to a large extent. However, previous studies (Romin and Bissonette, 1996; Knapp et al., 2007; Wang et al., 2010) found that they are significantly different. This implies that the two sets of data complement each other and should be combined to improve the quality of AVC data.

One way to deal with the inconsistent observation issue is to merge them together before the collision model construction. In the case of AVC data, the same AVC captured by both datasets may have different values for date and milepost. This variability may not be solved by a precise quantitative matching technique. Rather, it requires qualitative inferences in addition to quantitative analyses to determine matching data. The fuzzy logic-based data mapping algorithm has proven to be an effective way to deal with such problems related to linguistic vagueness and human factors (Zhao, 1997). Fuzzy logic mapping algorithms have been widely used in various fields of transportation engineering, such as ramp metering (Taylor and Meldrum, 1998), speed control systems (Rao and Saraf, 1995), and map matching issues (Syed and Cannon, 2004; Mohammed et al., 2006). Generally, the fuzzy logic mapping algorithm involves three major steps (Chen and Pham, 2001): (1) fuzzification: converting the quantitative inputs into natural language variables, (2) rule evaluation: implementing the mapping logic; and (3) defuzzification: converting the qualitative rule outcomes into a numerical output. The fuzzy logic based mapping algorithm will be explained in Chapter 3.

Another way to deal with the inconsistent observation issue is to model these data together. Most

of the regression models described in literature are univariate Poisson- (or Gamma-) based models designed for modeling general count problems. These univariate models are capable of estimating only one distribution parameter and would be limited in modeling multivariate issues. Recently, multivariate Poisson regression models (Miaou and Song, 2005; Ma and Kockelman, 2006; Park and Lord, 2007), multivariate zero-inflated Poisson regression models (Li et al., 1999), or multivariate Poisson-lognormal regression models (Karim and Tarek, 2009) have been used for modeling different but correlated count data sets. As a special case of multivariate Poisson regression models, a bivariate Poisson regression model can be used for paired count data sets (Karlis and Ntzoufras, 2003). However, bivariate Poisson and other multivariate Poisson regression models cannot handle over- or under-dispersed count data. In order to concurrently utilize the reported AVC data and carcass removal data even when they are dispersed, DIBP regression models (Karlis and Ntzoufras, 2005) are developed in Chapter 4 and applied to AVC modeling in Chapter 5.

2.4 Modeling Crashes with Interaction Behavior

Crashes with interaction behavior include Vehicle-Vehicle collisions, Animal-Vehicle collisions, Bicycle-Vehicle collisions, and Pedestrian-Vehicle collisions. Significant research efforts have been undertaken to utilize statistical modeling techniques for Vehicle-Vehicle collisions. Abdel-Aty and Abdelwahad (2004) estimated the probability of a car-truck rear-end crash by using a

nested logit model. Wang and Abdel-Aty (2006) proposed generalized estimation equations (GEM) to model rear-end crash frequencies at signalized intersections. Kim et al. (2007) estimated rear-end crash rates using a modified NB regression. Harb et al. (2008) proposed a conditional logistic regression model to estimate rear-end crash risk in work zone. Oh and Kim (2010) developed a method for estimating rear-end crash potential using individual vehicle trajectory data. Meng and Weng (2011) developed rear-end crash risk models to evaluate the rear-end crash risk in work zone activity area based on the available work zone traffic data.

Research effort also tried to reduce the Animal-Vehicle collisions (Huijser et al., 2007a; Huijser et al., 2007b; Lao et al., 2011a). Recently, many studies also have focused on dealing with Bicycle-Vehicle collisions. Kim et al., (2007) applied a multinomial logit model to explore the contributing factors on the injury severity of Bicycle-Vehicle collisions. Pai (2011) employed mixed logit models on three types of bicycle collisions: overtaking, rear-end, and door crashes. For the Pedestrian-Vehicle collisions, researchers have investigated the characteristics of pedestrians in the crashes. Examples include the influence of alcohol, demographic and economic characteristics, roadway characteristics, environmental factors, and collision types (Anderson et al., 1997; Davis, 2001; Öström and Eriksson, 2001; Matsui, 2005; Kim et al., 2008; Ulfarsson et al., 2010).

However, most previous accident modeling studies dealing with collisions with interaction behavior did not reflect human factors, despite their critical roles in the crash mechanism (Wang et al., 2003). Wang (1998) put forwarded a microscopic probability (MP) model for rear-end

collisions to include drivers' responses as part of the collision model. Later on, Wang and Nihan (2004) implemented this MP model to estimate the Bicycle-Vehicle collisions at signalized intersections.

Most of the previous studies focused on finding the contributing factors of collisions by analyzing the relationship between the collisions and their explanatory variables using modeling techniques. This interaction behavior within collisions was not fully considered in the previous modeling techniques. The MP model develop from Wang (1998) can be improved and used to describe the probability of this reaction. Further investigation on crash models dealing with the interaction behavior will be detailed in the methodology part of chapter 4.

2.5 Generalized (Non) linear Models

To identify factors contributing to traffic accidents, researchers have tried various statistical modeling techniques (e.g., Hubbard et al., 2000; Knapp and Yi, 2004; Lord and Mannering 2010). These research efforts have resulted in several widely accepted models. As mentioned earlier, the majority of these models are Generalized Linear Models (GLMs), such as Poisson regression, Gamma regression model, or GLM oriented models, such as negative binomial (NB) regression (or Poisson-gamma regression).

GLM-based approaches have provided valuable insights in investigating and examining collision occurrence. In general, GLM-based approaches utilize a linear regression to aggregate a series of independent variables, such as roadway curvature, shoulder width, traffic speed limit, etc. and establish a mapping relationship between independent variables and the dependent variable (which is typically the expected value of crash rates/severity) through a specific link function. However, such a GLM-based approach is constrained by its linear regression and may lead to biased model estimation and interpretation when the independent variable data demonstrates strong nonlinear features. For example, this research shows the parabolic impacts of truck percentage on rear-end crash risks: higher truck percentages increase the likelihood of crash occurrence before a certain threshold is reached, and then continuously increased truck percentages have negative impacts on rear-end crash occurrence. GLM-based approaches are not designed to model such a relationship and further developments are needed. An incorrect relationship built from the models will significantly affect the elasticity analyses of specific factors and locations. The elasticity analyses are important because transportation agencies heavily rely on estimated elasticity values to quantify marginal costs of various countermeasures for potential traffic safety improvements. Therefore, it is important to develop a more flexible modeling approach to enable nonlinear model specifications to better characterize and analyze a rear-end crash occurrence and its associated contributing factors.

Considerable research efforts have been undertaken to study the nonlinear models and extract more complex relationships. For example, Lindsey et al., (2000) applied a Generalized Nonlinear Model (GNM) to analyze pharmacokinetic data. Some previous studies (e.g. Wong et al., 2007;

Abdel-Aty and Haleem, 2011) used the logarithm of AADT instead of simply the AADT to deal with the nonlinear relationship between the crash rate and AADT. Recently, Turner and Firth (2012) developed an R package to help estimate the parameters in GNMs. Based on previous research efforts, a Generalized Nonlinear Model (GNM)-based approach is proposed to address the GLM inherent linear predictor constraint for collision data modeling. A nonlinear predictor can be established to aggregate significant independent variables to quantify their impacts on crash risks through a specific link function. (GNM)-based approach provides more flexible and greater explanatory power than GLMs. Detailed methodology of GNMs will be introduced in Chapter 4.

Chapter 3. Data Collection and Quality Control

This chapter's focus is data collection for the preformed research and data quality control with issues of inconsistent data sources. The chapter is organized as follows: First, information about collection of different types of data is provided. The inconsistent data source issue (including an example) is introduced in the following section, followed by an introduction to the fuzzy logic matching method dealing with this issue. Next, an application case study and its results are conducted to illustrate the decision making process using the fuzzy logic based approach (Lao et al., 2012a). Then, the proposed methodology will be verified using the expert judgment data collected from a survey at WSDOT, followed by a summary.

3.1 Collected Data

Data collected from different sources and used in this research are listed as follows.

(1) Collision Report data

Reported collisions between vehicles and non-domestic animals were extracted from the traffic accident records maintained by WSDOT. This dataset was also extracted from the Washington

State accident files provided by the HSIS (HSIS, 2009). However, since a significant portion of accidents is not reported, this dataset is only a subset of collisions. Collision reports are only required for incidents that cause damage values greater than a particular threshold. For AVC data, the threshold value is high enough that only large animal collisions are likely to be reported.

(2) Carcass Removal Data

WSDOT maintenance employees record the location— by milepost, date, weather, animal type, sex, and age— of every deer and elk carcass removed from state highways (Myers et al., 2007). Given that carcasses may also be removed by un-authorized parties and that some animals leave the right-of-way after a collision, this dataset is also a subset of all AVCs and may complement the Reported AVC dataset to some extent.

(3) Highway Geographic Information System (GIS) map

This dataset contains locations and curvatures of state highways in the GIS format.

(4) Deer Distribution Data

Deer distribution data were supplied by the Washington Department of Fish and Wildlife (WDFW) through WSDOT. This data contain GIS-based species distribution data for Mule Deer (Mule Deer Foundation, unpublished Data), Elk (Rocky Mountain Elk Foundation, unpublished

Data), and White-tailed Deer (Washington Gap Analysis Project, 1997).

(5) Survey Data

The research team conducted two surveys to collect input from WSDOT maintenance employees. The first survey was used to determine threshold values for the reported AVC and CR data. The other survey was used to verify the quality of the data recovery algorithm.

(6) Priority habitat and species database

This database contains location data for deer and elk habitats in Washington State. These data were provided by the WDFW.

(7) WeyWild: a compilation of wildlife habitat information for the Pacific Northwest

This dataset is a compiled database derived from 20 sources of species habitat information for southwestern Washington.

(8) Wildlife Habitat Matrices

This tool, derived from the Johnson and O'Neil (2001) assessment of wildlife habitat relationships for Washington and Oregon, provides tabular data on the vegetation types, vegetation structures, important habitat elements, population structures, and historical trends of

all terrestrial vertebrates in the state.

Data sources (1) through (5) were mainly used for the analysis, whereas data sources (6) through (8) were used for reference. Table 3-1 shows the years of data covered by each of the five major data types used in this research.

Table 3-1: Data collection information

Data	Data Time Covered	Date Received	Providing Agency
Collision Report Data	2000-2006	Apr. 2008 (Jan. 2009 update)	HSIS
Roadlog Data	2002-2006	Apr. 2008 (Jan. 2009 update)	HSIS
Carcass Removal Data	1999-2007	Jul. 2008	WSDOT
Survey Data		Feb. 2008-Mar. 2009	WSDOT
Deer Distribution Data		Jul. 2009	WSDOT & WDFW

3.2 Issues of Inconsistent Observations

As mentioned earlier, in some situations collision datasets attained from different data sources

may be inconsistent. Analyses based solely on a single dataset may result in biased conclusions, so, methods with some data merging techniques are desired for data quality improvement. The AVC data is used as an example to demonstrate the merging process.

Two types of AVC data are commonly used in AVC modeling and analysis: reported AVC data and CR data. This study will use the two datasets collected in Washington State to demonstrate the fuzzy logic-based data mapping algorithm. Note that the AVC records in the HSIS database have no detailed animal type information other than "domestic" or "non-domestic." However, they do have other detailed information, such as collision time and weather. The CR data used in this study were provided by the maintenance team of WSDOT. This dataset contains detailed information about animal species, such as mule deer, white-tail deer, and elk.

Ten State Routes (SRs) (US-2, SR-8, US-12, SR-20, I-90, US-97, US-101, US-395, SR-525 and SR-970) with relatively high AVC rates in the past several years were chosen as the study routes following the recommendation from WSDOT. As shown in the Figure 1-3 in Chapter 1, the Reported AVC data and CR data are substantially different. The number of CR records is typically more than that of the Reported AVC data on each route except for US-101. The Reported AVC data may likely underestimate the frequency of these types of collisions.

Since the two sets of data overlap to a certain extent, attention must be paid to avoid duplicating the same accident records. One of the most effective ways to determine if a reported AVC datum has a match in the CR dataset is to compare its similarities in occurrence time and location.

Generally, the reported AVC data is recorded the same day when the AVC occurs. However, the carcasses are picked up by the WSDOT maintenance staff depending on when they find the carcass. Theoretically, the carcass pickup day should be the same as the day when the AVC is reported. In reality, a perfect match between two datasets rarely happens. The record of the same event typically looks different in time and/or location in each dataset. Such differences can be explained as follows:

- Animals that die off the roadway or far away from any residences might not be removed for several days or even longer. In essence, these are cases where the dead animal is not an immediate hazard to motorists and/or not an obvious and unpleasant sight. Therefore, reporting and/or response can be delayed or non-existent.

- The WSDOT maintenance staff generally does not remove carcasses over weekends, except during the winter. During the winter months, the WSDOT maintenance team patrols several times every day and night so the carcasses can be spotted sooner. However, heavy snowfalls may completely hide carcasses and delay the removal process for multiple months. During the summer months, the WSDOT maintenance staff does not patrol the highways every day because they have other priority duties. In this case, a carcass not affecting traffic movement significantly may not be reported or identified immediately and hence might not be picked up in a couple of days.

- In addition, human errors may be introduced to the two datasets when the records were input manually.

In summary, not all animal carcasses were removed and reported by transportation agencies. Meanwhile, not all AVCs were properly reported and recorded in the HSIS. Therefore, both datasets are very likely to underestimate the actual number of AVCs to some extent. Combining the two datasets will make the research data more complete and hence provide a better information base for AVC studies. Specifically, combining these two datasets will extend the data breadth (increase samples).

3.3 Fuzzy Logic Matching Method

3.3.1 Fuzzification

Three attributes are used in the data mapping process: animal type, date, location. The animal categories for Reported AVC data and CR data are a little different. The “non-domestic” animal type reported in AVC data is matched with the three deer types and elk in CR data. After the animal types had been matched, this algorithm will consider only “date difference” and “location difference” as the inputs.

Date difference refers to the difference between the date when the carcass was collected and the date when the collision was recorded in the Reported AVC dataset. Note that the date recorded in the CR dataset should have been the same date or later as that in the Reported AVC database because a carcass cannot be collected until after the collision has happened. Therefore, the date difference is mathematically defined as:

$$\text{Date difference} = \text{Date in the CR dataset} - \text{Date in the AVC dataset} \quad (3\text{-}1)$$

Location difference is the milepost difference between the Reported AVC location and the location where the carcass was collected. The State Route numbers in a data pair are required to be identical before mileposts could be compared. Therefore, the location difference is defined as the absolute value between the milepost in the AVC dataset and the milepost in the CR dataset:

$$\text{Location difference} = |\text{Milepost in the AVC dataset} - \text{Milepost in the CR dataset}| \quad (3\text{-}2)$$

These inputs are then translated into four fuzzy classes based on the level of difference: small, medium, big, and very big (S, M, B, and VB). VB presents the situation in which the input is larger than a critical range. For example, if the location difference is only considered within 3 miles, a 5 mile difference will be marked as VB. The determination of the critical range will be introduced in 3.3.4.

A membership function (Li and Yen, 1995) for each class needs to be determined during the fuzzification step. A membership function describes the membership degree, defined as the truth

extent to the respondent class and its value ranges from zero to one. Most research (Taylor et al., 1998; Nikunja, 2006; Naso et al., 2006) has assumed the membership function to be a triangle for simplification and has designed it based on subjective experiences. However, the triangular membership functions may be too simple to accurately reflect the reality. Therefore, this study adopted a survey based method (Li and Yen, 1995) to determine the membership functions for the fuzzy classes. Details about the membership function determination process will be described in the algorithm application section.

3.3.2 Rule Design

Fuzzy logic rules are needed for mapping inputs to outcomes. Eleven rules, shown in Table 3-2, are designed for this algorithm. The default rule weights reflect the relative importance of the rules. As mentioned earlier, the two inputs are milepost difference and date difference. The matching output between the AVC and the CR datasets is the outcome which is represented by six fuzzy classes: very very low (VVL), very low (VL), low (L), medium (M), high (H), and very high (VH). For example, VVL presents the situation in which the output class is very close to zero. In other words, the candidate data pair is too different to be a possible matching pair.

The output class decreases with the increase of milepost difference and/or date difference. Rules 1 through 9 cover normal matching conditions. For example, Rule 9 could be interpreted as

follows: if the milepost difference is big and the date difference is big, then their matching degree is very low. Rules 10 and 11 deal with situations that the output class will become VVL if either of the inputs is outside the limits.

3.3.3 Defuzzification

The defuzzification process converts the qualitative rule outcome into a numerical output. The centroid defuzzification method (a.k.a. Center-of-Area or gravity methods) (Runkler, 1996; Taylor and Meldrum, 1998) is used to determine the matching degree (*MD*) in this research:

$$MD = \frac{\sum_{i=1}^{n} w_i c_i I_i}{\sum_{i=1}^{n} w_i I_i} \tag{3-3}$$

where w_i is the rule weight representing the importance of the i^{th} rule; c_i is the centroid of the output class i, and I_i is the implicated area of the output class i. The centroid of each output class is defined in Table 3-3. Note that if the output classes include VVL, the output *MD* is set to zero. *MD* is calculated for all possible data pairs. In this study, a data pair is regarded as a match if *MD* $\geqslant$0.5. If multiple matches are found, then the matching with highest MD will be selected.

Table 3-2. Rule base for fuzzy mapping algorithm

Rule	Default Rule Weight	Input Classes		Output Classes
		Milepost difference	Date difference	
1	1	S	S	VH
2	1	S	M	H
3	1	S	B	M
4	1	M	S	H
5	1	M	M	M
6	1	M	B	L
7	1	B	S	M
8	1	B	M	L
9	1	B	B	VL
10	1	VB	– *	VVL
11	1	–	VB	VVL

* "–" means any input classes

Table 3-3. Centroid value for output classes

	VH	H	M	L	VL	VVL
c_i	1	0.8	0.6	0.4	0.2	0

3.3.4 Determination of Membership Function

Before applying the fuzzy logic based mapping algorithm, the membership functions need to be determined. In order to make the membership functions objective, an expert survey was conducted to collect necessary information to set them up properly.

The survey was conducted from Feb. 5th 2009 to Mar. 3rd, 2009. The CR and Reported AVC datasets differ significantly and have different sources, so it is difficult to find people familiar with both datasets. Because the Reported AVC data are more precise in location and date, as well as more physically and directly tied to incident location, the Reported AVC data were chosen as a baseline for comparison to the application of fuzzy logic to the CR data. Therefore, survey subjects are the WSDOT staff members who have been working on the CR data collection for more than three years. The survey questionnaire contains four questions

directly related to the determination of the fuzzy membership function. Questions included, "Based on your experience, how far away do you expect to find the carcass from the location where the actual collision took place?" and "What is the greatest discrepancy in distance you would expect to find between the actual and reported locations for a carcass removal report?" Similar questions about the date difference were also included.

Forty-eight out of the 54 received responses were considered valid. The six discarded surveys were incomplete in critical questions. From each expert's inputs, we were able to understand how these experts judge the date and location differences and the threshold values to be used. Figure 3-1 illustrates the fuzzification process of an expert. For example, if a location difference is smaller than the expert's expected location difference, then the current data pair's location difference is small, in this expert's opinion. The location difference of this same data pair may have been considered as big in another expert's view. These measured differences in experts' judgments offer a solid foundation to build up the membership functions.

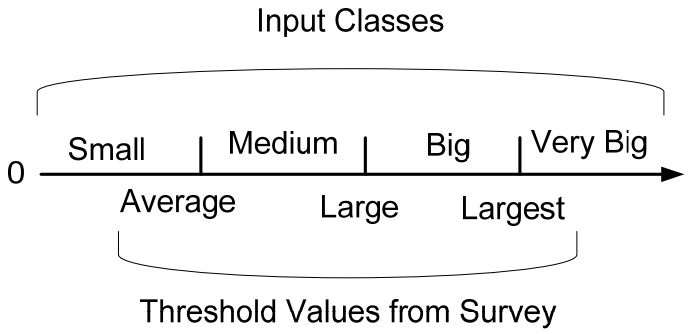


Figure 3-1 Determination of fuzzy classes.

The degree of membership of input value u (milepost difference or date difference) in fuzzy class A_i (i=1,2,3 representing the classes of S, M, B respectively) can be calculated by using the membership function for class A_i. The membership function is constructed as shown in Equation (3-4) by using the survey inputs from the WSDOT experts.

$$f_i(u) = n_{i,u} / K \tag{3-4}$$

where $n_{i,u}$ is the number of observations of $u \in A_i$ for class i and K is the total number of observations (valid responses received from the survey) for all classes (K =48 in this study).

The results for the constructed membership functions of the survey are shown in Figure 3-2 to Figure 3-4. Figure 3-2 shows the membership function for location difference between the AVC and CR datasets. For example, approximately 56% of the staff regarded one mile as a big difference while 38% of staff thought that it was a Medium difference and about 6% of staff regarded it as a Small difference.

Figure 3-3 and Figure 3-4 show the membership function for date difference on weekdays and weekends respectively. When an AVC happens during a weekend, the carcass is often collected on the following Monday or Tuesday, and therefore the date difference on weekends is

slightly larger than on weekdays. For example, approximately 60% of staff considered three days a big difference for weekdays but fewer staff (38%) considered the same period of time as a big difference for weekends.

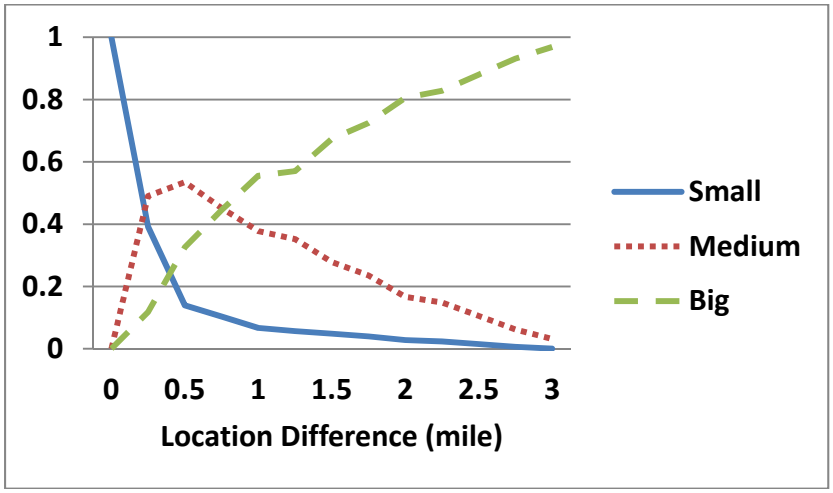


Figure 3-2 Membership function for location difference.

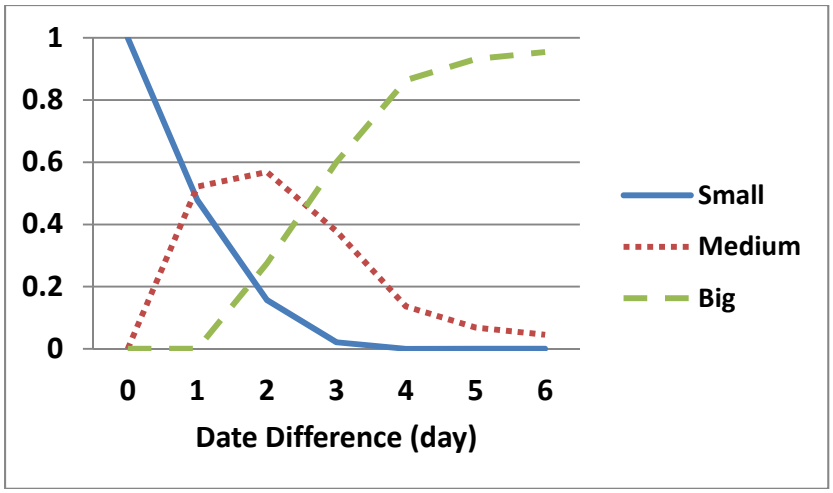


Figure 3-3 Membership function for time difference on weekdays.

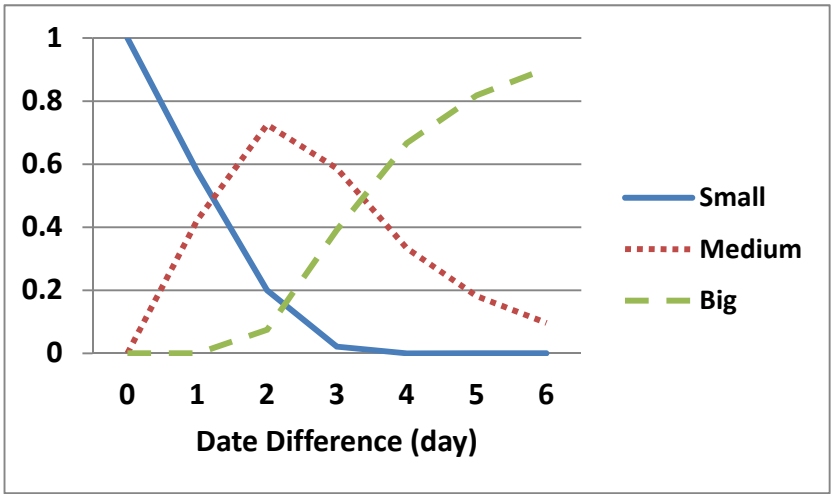


Figure 3-4 Membership function for time difference on weekends.

3.4 Mapping Application and Results

The fuzzy logic based mapping algorithm was used to combine the five-year (2002-2006) Reported AVC data and CR data for the ten study routes mentioned in the Research Data section.

Since the total records in the Reported AVC dataset are only about one third of those of the CR data, making use of the CR data can significantly increase the sample size for AVC studies. To merge the two datasets, their intersection needs to identify so that the same accidents will not be recorded twice in the combined dataset.

Additionally, the intersection of the two datasets can improve the richness of information because the combined dataset will have more variables on each matching AVC. As shown in Table 3-4, the fuzzy logic based mapping algorithm identified a matching percentage between 25%~35% for each year. This new dataset of matched records has variables combined from both datasets. The union of the two datasets can expand the data breadth. Compared to the original CR dataset, the new union dataset has about 15%~22% more records, as shown in the Improved Percentage column.

Table 3-4. Data mapping results for the study routes in five years (2002~2006)

Year	Total Number of Records	Matched	Matching	Union	Improved

	Reported AVC	Carcass Removal	Data Pairs	Percentage	Datasets	Percentage
2002	529	1876	152	28.7%	2253	20.0%
2003	508	1771	151	29.7%	2128	20.2%
2004	529	1702	139	26.3%	2092	22.9%
2005	544	2290	186	34.2%	2648	15.6%
2006	533	1944	144	27.0%	2333	20.0%

3.5 Algorithm Verification

After the proposed algorithm has been implemented, a major step is to verify whether the algorithm is able to reasonably imitate the experts' decision process and produce a combined quality dataset. However, because no ground-truth AVC data is available, it is nearly impossible to validate the performance of the algorithm by using the existing datasets. Therefore, another expert survey was also conducted from Mar. 5 to Mar 23, 2009 for verification purposes. Again, the survey participants are WSDOT employees who had collected CR data for more than three years. Each survey subject was asked to judge whether the data pairs listed on the questionnaire match. The disparity between the experts' results and the algorithm results can be a measure for the credibility of the proposed algorithm.

A total of 13 data pairs included in the survey questionnaire were extracted from the AVC dataset and the CR dataset. These data pairs are considered representative of both the day and location differences between the two datasets. As shown in Table 3-5, information about State Route, Milepost, Weekday, Month, and Day from the data pairs was also provided on the survey questionnaire. Many experienced WSDOT staffers were invited to fill out the questionnaire. They were asked to determine whether the data pairs match or not. Matching degree for each of the 13 listed data pairs was computed based on expert inputs. The computational results are then compared with the fuzzy logic based mapping algorithm outputs. The last three columns of Table 3-5 show the matching degrees from both the survey results and the fuzzy logic based mapping algorithm, as well as the percentage of the errors between survey and the results of the proposed algorithm. In the Matching Degree column, the gray cells indicate that the data pair should refer to the same collision; the clear cells indicate that the data pair does not match (In this study, the matching degree of a data pair should be 50% or higher to be marked as a match.).

The table 3-5 shows that the survey and algorithm results agree in all cases except data pair No. 11, which experts concluded was a match but the algorithm rejected. If the survey results are assumed accurate, then the accuracy rate (AR) for the proposed algorithm is:

$$AR = N_{accurate} / N_{total} = 12/13 = 92.3\% \quad (3\text{-}5)$$

where $N_{accurate}$ is the number of data pairs correctly matched by the algorithm; N_{total} is the total number of the data pairs evaluated. The matching rate of 92.3% is considered to be a very

encouraging result, given the complexity of this issue.

Mean Absolute Error (MAE), a quantity used to measure how close forecasts or predictions are to the eventual outcomes (Morris, 1986), was used as the error indicator. The MAE of the proposed algorithm can be calculated by using Equation (3-6):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|(f_i - y_i)\right| = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right| = 12\% \tag{3-6}$$

where f_i is the result estimated by the fuzzy logic-based data mapping algorithm; y_i is the ground truth matching degree values calculated from the survey result; and e_i is the MAE between the algorithm result and the survey result. The calculated error for each surveyed data pair is listed in the last column of Table 3-5.

Table 3-5. Survey and algorithm matching percentage for different data pairs

No	Route	Reported AVC Data				Carcass Removal Data				Matching Degree (%)		e_i*
		Milepost	Weekday	Month	Day	Milepost	Weekday	Month	Day	Survey	Algorithm	
1	2	302.1	Thu	Oct.	20	302	Thu	Oct.	20	100	96	0.04
2	2	327.2	Wed	May	25	325	Mon	Jun.	20	8	25	0.17
3	12	118.14	Mon	Feb.	14	118	Tue	Feb.	15	88	86	0.02
4	20	24.77	Wed	Oct.	26	24.1	Wed	Oct.	26	58	74	0.16
5	20	8.1	Thu	Nov.	10	5.5	Fri	Nov.	18	0	24	0.24

6	90	257.27	Sun	Sep.	25	257	Thu	Sep.	29	69	51	0.18
7	90	55.2	Sun	Jul.	31	56	Mon	Aug.	1	88	64	0.24
8	90	32.88	Thu	Mar.	31	34	Sat	Apr.	2	50	52	0.02
9	97	25.5	Wed	Jul.	20	24	Mon	Jul.	25	46	31	0.15
10	97	299.02	Sun	Sep.	10	299.7	Mon	Oct.	3	35	35	0
11	195	84.53	Mon	Nov.	14	83	Thu	Nov.	17	54	40	0.14
12	395	231.44	Fri	Apr.	29	233.8	Thu	May	12	12	24	0.12
13	970	2.21	Tue	Nov.	22	2	Wed	Nov.	23	96	82	0.14

* e_i is the absolute percentage error between the matching results

3.6 Summary

This chapter presented information regarding data collection and a fuzzy logic-based data mapping algorithm that aims to improve animal-vehicle collision (AVC) data by combining two types of data commonly used in AVC analysis: the Reported AVC data and carcass removal data. Two datasets collected from ten study routes in Washington State were used in this study.

The membership functions used in the fuzzy logic based mapping algorithm were formulated based on the survey responses from WSDOT experts who have been working in AVC-related work for years. Unlike predefined deterministic membership functions, the modified membership

functions can truly make the decision similar to the decision made by experts.

Using the proposed mapping algorithm, the carcass removal and the Reported AVC datasets can be combined to produce a more complete set of data. Through the use of this mapping algorithm, intersections of the two datasets can be identified as well. Records in the intersection of the two datasets contain more variables on the same accidents and can be used to support more detailed analysis of AVCs. About 25%~35% of the Reported AVC data can be matched to the CR data. The union of the two datasets can significantly increase the number of samples for AVC studies and hence expand breadth of data. Compared to the original CR dataset, the new union dataset increases the number of record by 15%~22%.

The proposed algorithm was verified by the expert judgment data on the surveyed AVC data pairs collected through another survey. The verification results showed that the accuracy of the proposed algorithm is approximately 90% for the limited pairs of data included in the survey. The fuzzy mapping algorithm was proved to be appropriate for increasing the quality and quantity of the AVC data. The improved dataset will benefit wildlife safety studies and countermeasure identifications. Since the design of the membership functions is adaptive in nature, the fuzzy logic based mapping algorithm introduced in this research can also be transferred for applications in other areas.

Chapter 4. Count Data Modeling Technologies

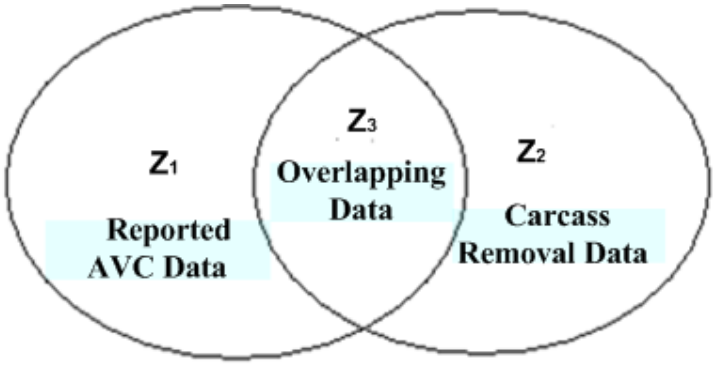
Three types of models are put forward in this chapter to address the last three key modeling issues identified in Section 1.3. The diagonal inflated bivariate Poisson (DIBP) models are proposed to dealing the issues of inconsistent data sources. The occurrence mechanism based probability models and GNM based models are developed to handle interaction behavior and non-linear relationships between dependent and independent variables.

4.1 Diagonal Inflated Bivariate Poisson Regression

The bivariate Poisson model and its diagonal inflated version, diagonal inflated bivariate Poisson regression (DIPB) model, have been used in the analysis of health care and sports data (Karlis and Ntzoufras, 2005). The DIBP model is chosen in this study for two reasons. First, both the bivariate Poisson and DIBP models are appropriate for modeling paired count data with correlation. The two data sets for AVC analysis, reported AVC data and carcass removal data, are different but should be correlated. Second, the DIPB model is capable of handling both over- and under- dispersed data (Karlis and Ntzoufras, 2005).

4.1.1 Bivariate Poisson Regression Model

Figure 4-1 shows the relationships between two types of data, the reported AVC data (left circle) and the carcass removal data (right circle) typically collected by transportation agencies. There are three regions of interest: Z_1, Z_2, and Z_3. Z_1 represents the AVC reports with no corresponding carcass removal data. Z_2 represents the carcass removal data with no counterparts in the reported AVC data. The records contained in both the reported AVC data and the carcass removal data are represented by Z_3. This area is the overlapping portion of the two data sets.



Z_1: AVC reports data without CR data; Z_2: CR data without reported AVC data; Z_3: overlapping portion

Figure 4-1 Relationship between the reported AVC and CR data sets

Let us assume that the count data sets Z_1, Z_2, and Z_3 follow independent Poisson distributions with parameters (means) λ_1, λ_2, and λ_3, respectively. Then the reported AVC data set $X = Z_1+Z_3$

and the carcass removal data set $Y = Z_2+Z_3$ follow a bivariate Poisson distribution (λ_1, λ_2, λ_3), with a joint probability mass function defined as (Karlis and Ntzoufras, 2005):

$$f_{BP}(x, y \mid \lambda_1, \lambda_2, \lambda_3) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{i=0}^{\min(x,y)} \binom{x}{i}\binom{y}{i} i! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right) \tag{4-1}$$

where x and y are the values that X and Y take on. The bivariate Poisson distribution is appropriate for modeling two random variables with positive dependence, which is the case for the reported AVC and carcass removal data sets. Its marginal distributions of X and Y follow Poisson distributions with $E(X) = \lambda_1+\lambda_3$ and $E(Y) = \lambda_2+\lambda_3$, respectively. Moreover, $COV(X, Y) = \lambda_3$, and hence λ_3 is a measure of dependence between the reported AVC data set and the carcass removal data set.

In the bivariate Poisson model, λ_k with k = 1, 2, and 3 can be related to various explanatory variables by using the classical exponential link functions. Therefore, the bivariate Poisson regression model can take the following form:

$$\begin{aligned} &(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}), \\ &\ln(\lambda_{ki}) = \omega_{ki}^T \beta_k \end{aligned} \tag{4-2}$$

where i = 1, . . . , n, is the roadway segment number, ω_{ki} is the vector of explanatory variables for roadway segment i, β_k is the corresponding coefficient vector for Z_k. In this study, the roadway segments are separated by consistent geometric factors. It should be noted that the double

Poisson model is a special case of the bivariate Poisson model when λ_3=0.

4.1.2 Diagonal Inflated Bivariate Poisson Regression Model

A major disadvantage of the bivariate Poisson model is that its marginal distributions cannot handle over-dispersed or under-dispersed data since its marginal distributions are Poisson distributions that require the mean and the variance to be equal (Karlis and Ntzoufras, 2005). The diagonal inflated bivariate Poisson model proposed by Karlis and Ntzoufras (2005) can be used to fix this problem. This model uses a more general form developed on the basis of zero-inflated models and the probabilities of the diagonal elements are inflated in the probability table. The diagonal inflated bivariate Poisson model can be defined on the basis of the bivariate Poisson regression model as follows:

$$f_{IBP}(x,y)=\begin{cases}(1-p_m)f_{BP}(x,y\mid\lambda_1,\lambda_2,\lambda_3), & x\neq y\\ (1-p_m)f_{BP}(x,y\mid\lambda_1,\lambda_2,\lambda_3)+p_m f_D(x\mid\theta,J), & x=y\end{cases} \tag{4-3}$$

where p_m is the mixing Proportion. $f_D(x \mid \theta, J)$ is the probability mass function of a discrete distribution $D(x; \theta)$. $D(x; \theta)$ can be a Poisson, geometric, or a simple discrete distribution. That is, the data process has a probability of 1-p_m to follow a bivariate Poisson distribution and a probability of p_m to follow $D(x; \theta)$. Note that the bivariate Poisson (when p_m=0) and the zero-inflated double Poisson model (when λ_3=*0* and *J=0*) are special cases of diagonal inflated model

(Karlis and Ntzoufras, 2005). $f_D(x \mid \theta, J)$ can be defined as

$$f_D(x \mid \theta, J) = \begin{cases} \theta_x & for \quad x = 0,1,...,J \\ 0 & for \quad x \neq 0,1,...,J \end{cases} \tag{4-4}$$

where $\sum_{x=0}^{J} \theta_x = 1$. J is a parameter that controls the number in the diagonal cells in the cross-tabulation for the paired datasets considered by the model (cross-tabulation will be introduced in Section 3). If J = 0, only x=y=0 contributes to the inflated part ($f_D(x \mid \theta, J)$), then the model in Equation (4-3) become a zero-inflated model. If J = 1, both x=y=0 and x=y=1 contributes to the inflated part. In this case, cell (0, 0) and cell (1, 1) of the cross-tabulation are considered in the inflated part.

The marginal distributions of a DIPB model are mixtures of distributions with one Poisson component. For example, the marginal distribution of X is:

$$f_{IBP}(x) = (1-p) f_{P_0}(x \mid \lambda_1 + \lambda_3) + p f_D(x \mid \theta) \tag{4-5}$$

where $f_{Po}(x \mid \lambda)$ is the Poisson probability mass function with parameter $\lambda_1+\lambda_3$. The marginal distributions of the DIPB model can model either under-dispersed or over-dispersed count data, depending on the definition of $D(x; \theta)$. For example, if J=1, $\lambda_1+\lambda_3$=1 and p_m=0.5, the resulting distribution is under-dispersed. While J=0 (the simplest case of zero-inflated models), the resulting distribution is over-dispersed. This implies that the DIPB model is more flexible than the bivariate Poisson regression model and hence a clearly better choice for modeling the AVC

data in this study.

The parameters in most multivariate Poisson or related models are difficult to estimate because of the computational issues involved in their applications (Karlis and Ntzoufras, 2005; Ma and Kockelman, 2006). However, recent developments in statistical software models and computer hardware have provided several ways to estimate bivariate Poisson models. In this study, an open source statistical analysis package, R (http://www.r-project.org/, 2009), was used to estimate the models. The Expectation-Maximization (EM) approach (Dempster, Laird, and Rubin 1977; Borman, 2009) is used for estimating the parameters in the DIPB model. Details of the EM algorithm can be found in (Karlis, 2003; Karlis and Ntzoufras, 2005).

4.2 Possibility Modeling Considering Interaction Behavior

4.2.1 Extended Application of the Microscopic Probability (MP) Model

MP Model Structure

This study is based on the MP model proposed by Wang (1998). An overview of the MP model and its association with the AVC model are summarized in this section. The MP model describes the relationship between the presence of a leading vehicle and the ineffective response of a driver in the following vehicle. An important advantage of this approach is its capability of considering

the mechanism of accident occurrence in risk modeling. For example, this MP model separated the possibility of the obstacle been present on the road and the possibility of ineffective response from the driver. With the separation, this model can provide a more detail analysis on the response behavior from the driver. This approach has been successfully applied in many subsequent studies of accident risks (see for example Siddique 2000, Wang et al., 2003, and Kim et al., 2007) and achieved favorable results. Although animals' behavior exhibit different patterns than drivers', Wang' MP model will be applied to formulate the VAC before the new VAIP model is developed and investigated. Their performance will be examined and analyzed as follows.

In the MP model, the probability for a randomly selected vehicle to have an accident on a certain roadway section is the probability of the driver's ineffective response P_{vf} conditioned on the presence of an obstacle presenting P_o. In other words, the probability for a driver to have an AVC (P_{AVC}) can be expressed as the product of P_o and P_{vf} (Wang, 1998):

$$P_{AVC} = P_o \cdot P_{vf} \tag{4-6}$$

However, P_o, and P_{vf} are not directly observable, and require further estimation.

P_o Formulation

An animal becomes an obstacle for vehicles if its highway-crossing movement interrupts the smooth movement of vehicles. When an animal highway-crossing movement occurs within a

certain period, the animal may become an obstacle to the arriving vehicle. This period is called "effective time." As the arrival of an obstacle is discrete, nonnegative, and random, it is assumed to be a Poisson arrival process. In such a process, intervals between arrivals are independent and follow the same exponential distribution (Pitman, 1993). Assuming a disturbance *j* whose arrival rate is η_j and effective time is t_j, the density function is then:

$$f(t) = \eta_j e^{-\eta_j t_j} \qquad \text{for } t_j > 0 \tag{4-7}$$

According to the memoryless property of the exponential distribution (Pitman, 1993), the probability of having a disturbance *j* within t_j is independent of the time waited. Therefore, the probability for an arriving vehicle encountering disturbance *j* within t_j can be calculated by:

$$P_j = \int_0^{t_j} \eta_j e^{-\eta_j t} dt = 1 - e^{-\eta_j t_j} \tag{4-8}$$

Since any of the disturbances occurring in t_j may result in an AVC, the probability of encountering an obstacle animal, P_o, is equivalent to the probability that at least one disturbance occurs within the effective period. Therefore, P_o can be formulated as:

$$P_o = 1 - \prod_{j=1}^{J} (1 - P_j) \tag{4-9}$$

Replacing P_j with Equation (4-8), P_o becomes:

$$P_o = 1 - e^{-\sum_j \eta_j t_j} \tag{4-10}$$

In Equation (4-10), $\sum_j \eta_j t_j$ should always be positive and dependent on a set of variables. Thus, an exponential link function can be employed to reflect the effects of the explanatory factors as shown as:

$$\sum_j \eta_{dj} t_{dj} = e^{\beta_o x_o} \tag{4-11}$$

P_o then becomes:

$$P_o = 1 - e^{-e^{\beta_o x_o}} \tag{4-12}$$

where $\boldsymbol{\beta_o}$ and $\boldsymbol{x_o}$ are vectors of unknown parameters and explanatory variables of disturbance frequency, respectively. $\boldsymbol{\beta_o}$ does not change with location, while $\boldsymbol{x_o}$ does. Animal habitat integrity, habitat size, and animal population are very likely contribution variables to $\boldsymbol{x_o}$.

$\boldsymbol{P_{vf}}$ *Formulations*

It is assumed that a driver cannot avoid a collision if their Necessary Perception Reaction Time (NPRT) is longer than the Available Perception Reaction Time (APRT). The APRT refers to the

time a driver has for completing their perception and response under a given condition. The NPRT is the ability-oriented minimum required perception reaction time and typically varies from person to person. Both the APRT and the NPRT are random variables and are assumed to follow normal distributions. Since a normal distribution does not have a closed form for cumulative probability calculation, the Weibull distribution is used instead. The NPRT is assumed to follow the Weibull (α, λ) distribution, and the APRT is assumed to follow the Weibull (α, γ) distribution. Here, λ and γ are the scale parameters. The Weibull distribution shape parameter α is chosen to be 3.25 in this study because it has been empirically verified that when α=3.25, the Weibull distribution is a very good approximation to the normal distribution (Kao, 1960; Plait, 1962). Using the assumed distributions for the APRT and the NPRT, P_{vf} can be calculated as:

$$P_{vf} = \int_0^\infty \int_{t_{av}}^\infty f(\lambda,t) f(\gamma,t_{av}) dt dt_{av} = \int_0^\infty e^{-\lambda t_{av}^\alpha} \alpha\gamma t_{av}^{\alpha-1} e^{-\gamma t_{av}^\alpha} dt_{av} = \frac{1}{1+\lambda/\gamma} \tag{4-13}$$

where t_{av} is the variable used to represent the APRT. Equation (4-13) shows that P_{vf} is only dependent on λ/γ, and has no relationship to α. Since the parameters λ and γ are positive variables, λ/γ can be related to various factors by using an exponential link function as shown in Equation (4-13). Correspondingly, P_{vf} can be written as.

$$\lambda/\gamma = e^{-\boldsymbol{\beta}_{vh}\mathbf{x}_{vh}} \tag{4-14}$$

$$P_{vf} = \frac{1}{1+e^{-\boldsymbol{\beta}_{vh}\mathbf{x}_{vh}}} \tag{4-15}$$

where $\boldsymbol{\beta}_{vh}$ and $\boldsymbol{x}_{vh}$ are vectors of unknown parameters and explanatory variables, respectively, related to P_{vf}. Variables affecting drivers' task load and action complexity need to be included in $\boldsymbol{x}_{vh}$.

Integrated MP Model

The application of Wang's (1998) MP model in AVC only has the terms of the probability of an animal being present on the road (P_o) and the probability of an ineffective response by the driver (P_{vf}). Substituting Equations (4-12) and (4-15) into Equation (4-6), the probability of an individual vehicle being involved in an AVC is formulated as:

$$P_{AVC} = P_o P_{vf} = \frac{1-e^{-e^{\boldsymbol{\beta}_o \mathbf{x}_o}}}{1+e^{-\boldsymbol{\beta}_{vf}\mathbf{x}_{vf}}} \tag{4-17}$$

4.2.2 Vehicle-animal Interaction-based Probability (VAIP) Model

As discussed in the "Introduction" section, the AVC process is difficult to accurately model and interpret because many subjective and objective factors, such as human and animal factors,

cannot be properly reflected in the model. It is needed to have a modeling process that considers two significant AVC contributors: insufficient responses from drivers, such as a lack of deceleration, swerving and late responses from animals, such as freezing, running in the wrong direction. These two contributors interact with each other so that an AVC may be caused by either one or both. Since the MP model was originally developed for vehicle-to-vehicle collisions, the responses of animals were not considered in the modeling structure. An AVC could be avoided if drivers can react early and quickly to the obstacle or if the animals can notice to oncoming vehicles in a timely manner. Therefore, a third item addressing animal's response is desired in the MP model to enhance model rationality and applicability on AVCs. Thus, a vehicle-animal interaction-based probability (VAIP) model is proposed as an extension of the MP model.

4.2.2.1 VAIP Model Structure

This study considers that the occurrence of an AVC is conditioned on the presence of an animal in the roadway, ineffective response of the arriving vehicle driver, and the animal's failure to escape. Therefore, the vehicle-animal interaction probability can be formulated as

$$P_{AVC} = P_o \cdot P_{vf} \cdot P_{af} \tag{4-18}$$

where P_o is the probability of a hazardous crossing presence of an animal when vehicles travel

along roadways, P_{vf} is the probability of ineffective response of the driver, and P_{af} is the probability of the animal failing to escape being hit. Thus the probability for a randomly selected vehicle to have an AVC on a certain roadway section is the product of P_o, P_{vf}, and P_{af}. In this VAIP model, P_o and P_{vf} are defined according to the MP model and P_{af} describes animals' responses in a collision. In this study, the animals' responses are simplified by following a similar model structure of P_{vf} by comparing the animals' necessary perception reaction time with the available perception reaction time. Here, the available perception reaction time refers to the time an animal has for noticing and escaping from the approaching vehicle. The necessary perception reaction time is the minimum required perception reaction time depending on factors such as animal species and characteristics. Both are random variables and are assumed to follow normal distributions. By following the same modeling process with P_{vf} in Section 2.1.3 "P_{vf} Formulation", P_{af} can be written as:

$$P_{af} = \frac{1}{1+e^{-\boldsymbol{\beta}_{ah}\mathbf{x}_{ah}}} \tag{4-19}$$

where $\boldsymbol{\beta}_{ah}$ and $\boldsymbol{x}_{ah}$ are vectors of unknown parameters and explanatory variables, respectively, related to P_{af}. Variables affecting animal' action need to be included in $\boldsymbol{x}_{ah}$.

4.2.2.2 Integrated VAIP Model

By substituting Equations (4-12), (4-13), and (4-19) into Equation (4-18), the integrated VAIP risk model for each roadway section can be rewritten as:

$$P_{AVC} = P_o P_{af} P_{vf} = \frac{1 - e^{-e^{\boldsymbol{\beta}_o \mathbf{x}_o}}}{(1 + e^{-\boldsymbol{\beta}_{af} \mathbf{x}_{af}})(1 + e^{-\boldsymbol{\beta}_{vf} \mathbf{x}_{vf}})} \tag{4-20}$$

where, P_o is the probability of an animal being present on the road, P_{af} is the failure probability by the animal to escape from being hit, and P_{vf} is the probability of an ineffective response by the driver. One can see that the model contains not only road environment related factors, but also factors related to the behaviors of both humans and animals. The inclusion of human and animal factors is one of the major distinctions between the proposed model and most existing AVC models. Note that if the animals' reactions are dispensable as stationary objects, the probability, P_{af}=1, and the VAIP model reduces to the MP model.

4.2.3 P_{AVC} Formulation

It is assumed that vehicles within a traffic flow have a consistent AVC risk, P_{AVCi}. Thus, the number of AVCs occurring within this flow follows binomial distribution:

$$P(n_i) = \binom{f_i}{n_i} P_{AVCi}^{n_i} (1 - P_{AVCi})^{f_i - n_i} \tag{4-21}$$

where f_i is the annual traffic volume that can be calculated from the annual average daily traffic (AADT) for roadway section i, and n_i is the number of AVC occurred within f_i.

Since AVCs are very rare, P_{AVCi} should be very small while traffic volume f_i should be very large for the given span of time. Thus, the Poisson distribution is a good approximation to the binomial distribution (Pitman, 1993):

$$P(n_i) = \frac{m_i^{n_i} \cdot e^{-m_i}}{n_i!} \tag{4-22}$$

with Poisson distribution parameter:

$$m_i = E(n_i) = f_i \cdot P_{AVCi} \tag{4-23}$$

The mean and variance in a Poisson distribution need to be the same. However, in most cases, accident data are over-dispersed. An easy way to overcome this difficulty is to add an independently distributed error term, ε_i, to the log transformation of Equation (4-23). That is:

$$\ln m_i = \ln(f_i P_{AVCi}) + \varepsilon_i \tag{4-24}$$

We assume $\exp(\varepsilon_i)$ is a Gamma distributed variable with mean 1 and variance δ. Substituting Equation (4-23) into Equation (4-21) yields:

$$P(n_i \mid \varepsilon_i) = \frac{e^{(-f_i P_{AVCi} \exp(\varepsilon_i))} \cdot (f_i P_{AVCi} \exp(\varepsilon_i))^{n_i}}{n_i!} \tag{4-25}$$

Integrating ε_i out of Equation (4-24), a negative binomial distribution model can be directly derived as the following:

$$P(n_i) = \frac{\Gamma(n_i + \theta)}{\Gamma(n_i + 1)\Gamma(\theta)} \left(\frac{\theta}{f_i \cdot P_{AVCi} + \theta}\right)^{\theta} \left(\frac{f_i P_{AVCi}}{f_i \cdot P_{AVCi} + \theta}\right)^{n_i} \tag{4-26}$$

where $\theta = 1/\delta$. The expectation of this negative binomial distribution equals to the expectation of the Poisson distribution shown in Equation (4-22). The variance is now:

$$V(n_{ik}) = E(n_{ik})[1 + \delta E(n_{ik})] \tag{4-27}$$

Note that the Poisson regression model is regarded as a limiting NB regression model when δ approaches zero (Washington et al., 2003).

4.3 Generalized Nonlinear Models

In this section, the GLM-based modeling principles, formulated by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989), is introduced, and then the extended GNM-based approach is presented for crash data analysis. An elasticity estimation method is detailed for the significantly contributing factors based on the proposed GNM nonlinear predictor.

4.3.1 GLMs for Crash Data

In classical linear regression models, the expectation of crash rate (crash frequency per mile is used in this research) is formulated as an ordinary linear model. This model specification can be expressed as follows (McCullagh and Nelder, 1989)

$$E(y_i) = \mu_i = \sum_{j=1}^{n} x_{ij}\beta_j; \qquad \text{i=1,…,n} \qquad (4\text{-}28)$$

Where y_i is the crash rate along Roadway Segment i, E(y_i) or μ_i is the expected crash rate along Segment i during a certain time period; x_{ij} is the j^{th} explanatory variable for Segment i; β_j is the corresponding coefficient for the j^{th} explanatory variable; n is the total number of explanatory variables considered in the model. Compared to the simplest linear regression, more complicated models, such as Poisson, Gamma, Gaussian, Logit, Probit, Negative Binomial regressions, etc. have been used to enhance their capability of approximating and interpreting crash data. These models can be generalized by using a smooth and invertible linearizing link function to transform the expectation of the response variable, μ_i, to its linear predictor:

$$g(\mu_i) = \sum_{j=1}^{n} x_{ij}\beta_j \qquad (4\text{-}29)$$

Where, g(.) is the link function, which is monotonic, differentiable to connect the linear predictor of the explanatory variables with the expected crash rate in various formats, such as

identity, log, logit, etc. Its inverse function is expressed by $g^{-1}(.)$. In this research, the log function is used for rear-end crash analysis. The distribution of y_i is a member of a scaled exponential family, and its generalized density function can be expressed as

$$f(y_i;\theta;\psi) = exp\left\{\frac{y_i\theta - b(\theta)}{a(\psi)} + c(y_i;\psi)\right\} \quad (4\text{-}30)$$

Where, θ and ψ are parameters characterizing the density, and $a(\psi)$, $b(\theta)$, and $c(y_i;\psi)$ are real-valued functions. Ψ is the scale (or dispersion) factor. Depending on the specific formats of $a(\psi)$, $b(\theta)$, and $c(y_i;\psi)$, y_i could follow Gaussian, binomial, Poisson, gamma, inverse-Gaussian distributions, etc. The GLMs could be extended to multivariate exponential families (such as the multinomial distribution) and to certain non-exponential families (such as the two-parameter Negative-binomial distribution). One should note that the issues of data over-dispersion for crash data analysis can be addressed by the Quasi-Poisson GLMs. The Quasi-Poisson model utilizes the dispersion parameter, Ψ, in Equation (4-30) to model over-dispersion data compared to the Poisson model. If Ψ>1, the variance of y_i increases more rapidly than its mean.

4.3.2 GNMs for Crash Data

As discussed earlier, in many scenarios the relationship between the expected crash rate and its associated factors cannot be simply expressed by GLMs. GNMs are proposed as an extension of

a GLM to satisfy such specific requirement by changing the linear predictor, to be nonlinear in the parameters, β_j, in Equation (4-29). Without loss of any generality, assuming a Poisson model is formulated for analyzing rear-end crash data from 2002-2006 from ten highways (US2, SR8, US12, SR20, I90, US97, US101, US395, SR525 and SR970) in Washington State. Its non-linearizing link function is $g(\mu_i) = \log(\mu_i)$. Figure 4-2 illustrates the association between the logarithm of the expectation of crash rates (number of crashes per mile) and roadway grades. The solid marker curve shows crash rate change tendencies.

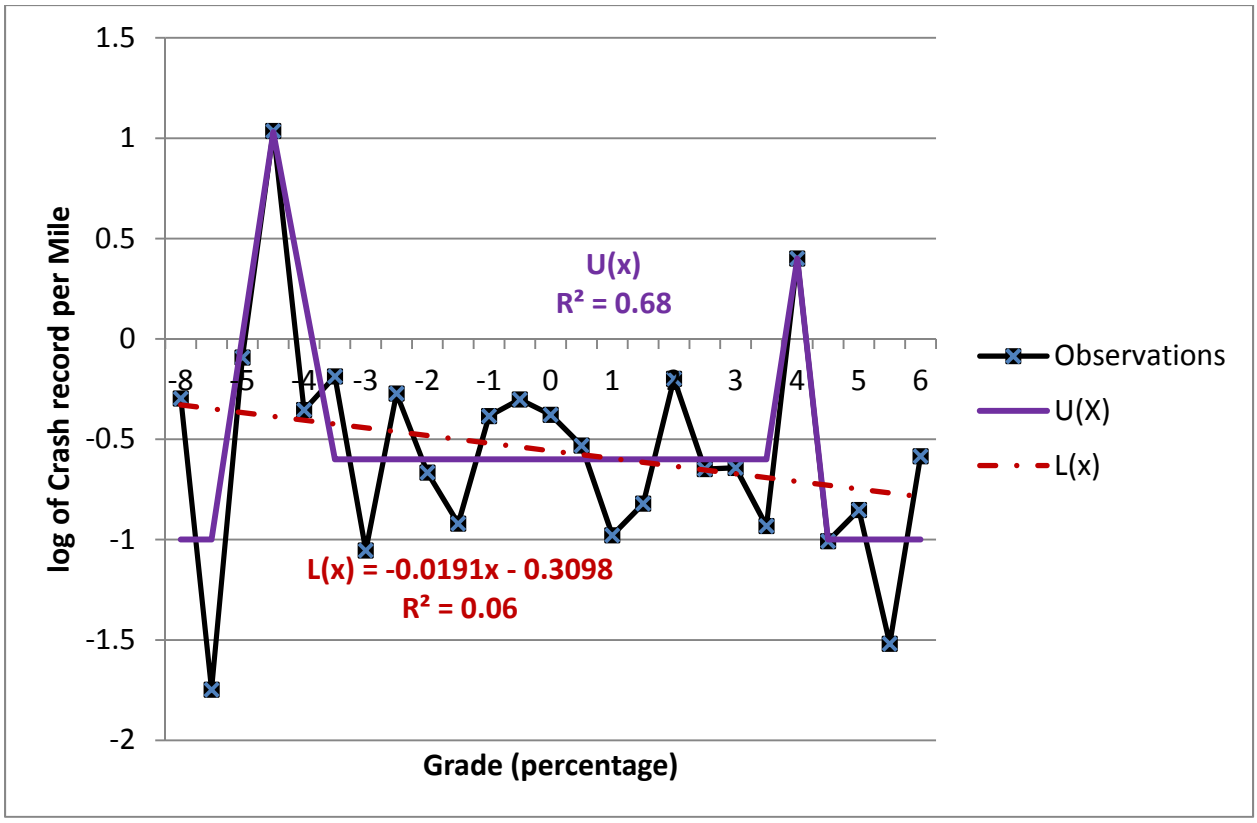


Figure 4-2. Rear-end crash rate (crash frequency per mile) in five years (2002-2006) from ten highways in Washington State, by grade

Previous GLM-based studies assume the monotonic relationship between the dependent variable, average Crash Rate and the independent variable, Grade through a linearizing link function, $\log(\mu_i)$. The logarithm of the expectation of crash rates is supposed to increase or decrease consistently when the variable, Grade, changes as a linear function. In Figure 4-2, such a linear relationship is modeled by the red dashed line $L(x)$. In order to quantify how well the different models fit the data and measure the model performance, the coefficient of determination, R^2, is used as the standardized model performance assessment criteria. R^2, for generalized non-linear models and piecewise linear functions, can be calculated by Equation (4-31)

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} \tag{4-31}$$

where the SS_{err} and the SS_{tot} are the sum of squares of residuals and the total sum of squares (proportional to the sample variance), respectively, which can be calculated as follows.

$$SS_{tot} = \sum_i (x_i - \bar{x})^2 \tag{4-32}$$

$$SS_{err} = \sum_i (x_i - \bar{\hat{x}_i})^2 \tag{4-33}$$

where, x_i is the i^{th} data value, $\bar{\hat{x}_i}$ is the associated modeled value based on the generalized non-linear models or piecewise linear functions, and $\bar{x}$ is the mean of the observed data.

The R^2 for the fitted line $L(x)$ is nearly equal to zero (R^2=0.06), showing a poor goodness-of-fit. Consequently, $L(x)$ could not model the fluctuation patterns associated with the variable, Grade, and misrepresent the true relationship between the expected crash rate and the geometric variable, Grade, assuming all other contributing factors are approximately equal across all grade levels.

Therefore, a GNM-based approach is needed to reasonably model the nonlinear relationship between the expected crash rate and its associated factors. Based on the visualized comparisons, a piecewise linear function $U(x)$ could be suitable to extract data change patterns. Based on the relationship between the logarithm of expected crash rate, and Grade, two peaks should be appropriately modeled when Grade=-4.4 and Grade=4. When the grade is small or moderate from -3.5 to 3.5, its impacts on crash occurrence are relatively stable. Its impacts become more severe when the grade increase beyond a certain threshold, for example Grade=-4.4 and

Grade=4. When the absolute value of the grade continuously increases (Grade>5 or Grade < -5.5) the crash rate tends to be lower. This may result from the facts that the drivers will drive much slower and pay more attention to handle considerable grade changes in these situations. Additionally, in a similar grade range, higher crash rate were observed on downhill highway sections than ones on uphill sections. This finding is consistent with the previous accident studies (Ahmed et al, 2011; Yu et al., 2013). In order to describe these unique impacts of Grade on crashes occurrence and a piecewise linear function is developed to fit the data as shown as the red dashed line in Figure 4-2:

$$U_g(x) = \begin{cases} -1 & -7 \le x < -5.5 \\ -1 + 2.5(x+5)/(5-4.4) & -5.5 \le x < -4.4 \\ -0.6 + 2.1(-x-3.5)/(4.4-3.5) & -4.4 \le x < -3.5 \\ 0.1 & -3.5 \le x < 3.5 \\ -0.6 + 1.2(x-3.5)/(4-3.5) & 3.5 \le x < 4 \\ -1 + 1.6(5-x)/(5-4) & 4 \le x < 4.5 \\ -1 & 4.5 \le x < 7 \end{cases} \quad (4\text{-}34)$$

In order to verify its goodness-of-fit of the piecewise linear function $U(x)$, its R^2 is calculated and aggregated for multiple linear pieces. The results of R^2= 0.68 indicate the piecewise linear function significantly outperforms the linear regression function and provides a better modeling performance for fitting the expected crash rate. For the other explanatory variables, the diverse nonlinear predictor, $U(x)$, such as polynomial function, exponential function, parabolic function, logarithmic function, etc. may be utilized to extract proper data features. In general, the model format of $U(x)$ can be determined based on statistical analysis of the crash rate and the specific explanatory variable.

Aggregating the nonlinear predictors for all the independent variables, and Equation (4-29) can be rearranged as

$$E(y_i) = \mu_i = \sum_{j=1}^{n} U_j(x_{ij}) w_j; \qquad \text{i=1,…,n,} \qquad (4\text{-}35)$$

where $U_j(x_{ij})$ is a nonlinear predictor for the j^{th} explanatory variable, and w_j is the corresponding weight for U_j. Consequently, the GNM link functions becomes

$$g(\mu_i) = \sum_{j=1}^{p} U_j(x_{ij}) w_j \qquad (4\text{-}36)$$

If all the $U_j(x_{ij})$ in the model are linear regressions of x_{ij}, a GNM will degrade to a GLM. Therefore, in this research, GLMs are special cases of GNMs.

4.4 Elasticity Calculation

In order to understand the influence and magnitude of contributing factors for potential safety improvements along a particular segment, the marginal effects of significant independent variables need to be calculated. Elasticity, a measure of a percentage change of the dependent variable resulting from a 1% change of an independent variable (Washington et al., 2003), is suitable for proportional marginal effect analysis here. The elasticity value can be calculated as

(Shankar et al., 1995; Abdel-Aty and Radwan, 2000; Washington et al., 2003):

$$E_{x_{ij}}^{\mu_i} = \frac{\partial \mu_i}{\partial x_{ij}} \frac{x_{ij}}{\mu_i} = \frac{\partial g^{-1}\left(\sum_{j=1}^{p} U_j(x_{ij}) w_j\right)}{\partial x_{ij}} \frac{x_{ij}}{g^{-1}\left(\sum_{j=1}^{p} U_j(x_{ij}) w_j\right)} \quad (4\text{-}37)$$

where μ_i is the expected crash rate for Roadway Segment i, and x_{ij} is the j-th explanatory variable for Segment i. One can find in Equation (4-37) that elasticity is determined by the value of the variable x_{ij}, the inverse link function, $g^{-1}()$, and the (non) linear predictor $U_j()$. Assume that crash rate follows the Poisson distribution, then the inverse link function $g^{-1}(.)=\exp(.)$, then Equation (4-37) can be exemplified as follows:

$$E_{x_{ij}}^{\mu_i} = \frac{\partial \exp\left(\sum_{j=1}^{p} U_j(x_{ij}) w_j\right)}{\partial x_{ij}} \frac{x_{ij}}{\exp\left(\sum_{j=1}^{p} U_j(x_{ij}) w_j\right)} = \frac{\partial (\sum_{j=1}^{p} U_j(x_{ij}) w_j)}{\partial x_{ij}} x_{ij} = w_j x_{ij} \frac{\partial U_j(x_{ij})}{\partial x_{ij}} \quad (4\text{-}38)$$

Equation (4-38) emphasizes the importance of the (non) linear predictor $U_j()$. It could introduce significant errors if inappropriate model specifications are used to estimate the values of elasticity without considering the nonlinear predictor. For GLM, $U_j(x_{ij}) = x_{ij}$. Then Equation (4-38) can be rewritten as:

$$E_{x_{ij}}^{\mu_i} = \frac{\partial \mu_i}{\partial x_{ij}} \frac{x_{ij}}{\mu_i} = w_j x_{ij} \quad (4\text{-}39)$$

The elasticity in Equations (4-38) and (4-39) apply when the explanatory variable $\boldsymbol{x}_{ij}$ is continuous. In case of a discrete variable, pseudo-elasticity is estimated as an approximate

elasticity of this variable (Washington et al., 2003):

$$E_{x_{ij}}^{\mu_i} = \frac{Exp(w_j)-1}{Exp(w_j)} \tag{4-39}$$

4.5 Measures of Goodness-of-fit

In order to evaluate the explanatory and predictive power of a model, several commonly used measures of goodness-of-fit (GOF) are adopted here for model comparisons: ρ^2 (rho-squared), adjusted ρ^2 (Ben-Akiva and Lerman, 1985), Akaike's Information Criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978； Liddle, 2007). These measures are summarized as follows:

ρ^2 (rho-squared) is the log-likelihood ratio index, and is used to evaluate a model's goodness-of-fit for random, discrete, and sporadic count data (Ben-Akiva and Lerman, 1985; Chin and Quddus, 2003). The index is formulated as

$$\rho^2 = 1 - \frac{\ln L(\hat{\boldsymbol{\beta}})}{\ln L(0)} \tag{4-40}$$

Adjusted ρ^2 (rho-squared) is the log-likelihood ratio index, and is used to evaluate model's GOF for random, discrete, and sporadic count data (Ben-Akiva and Lerman, 1985; Chin and Quddus,

2003; Washington et al., 2003). The index is formulated as:

$$\rho^2 = 1 - \frac{\ln L(\hat{\boldsymbol{\beta}}) - K}{\ln L(0)} \tag{4-41}$$

where $L(\hat{\boldsymbol{\beta}})$ is the maximum likelihood estimation of the compared model, $L(0)$ is the initial maximum likelihood estimation of the same model with only the constant term, and K is the number of parameters estimated in the model.

AIC is another measure of GOF for a statistical model (Akaike, 1974). AIC is often used for model selection. The model with the lowest AIC is considered the best model. In general, AIC is formulated as:

$$AIC = 2K - 2\ln(L) \tag{4-42}$$

where L is the maximum likelihood estimation of the model.

BIC (Schwarz, 1978) is also a criterion used for model selection among a group of models with different numbers of parameters. Compared to AIC, BIC has a stronger penalty for additional parameters. Similarly, the model with the lowest BIC is considered the best model. BIC is calculated as:

$$BIC = -2\ln(L) + k\ln N \tag{4-43}$$

where N is the number of observations in the data.

Chapter 5. Modeling Animal-Vehicle Collisions Using Diagonal Inflated Bivariate Poisson Regression

5.1 Data Description

Ten highways (US2, SR8, US12, SR20, I90, US97, US101, US395, SR525 and SR970) in Washington State were selected as study routes, as was recommended by WSDOT experts. Five years (2002-2006) of the reported AVC and the carcass removal datasets were analyzed for modeling crashes considering inconsistent data sources.

Table 5-1 shows all the explanatory variables used in the models. Annual average daily traffic (AADT) is converted into thousands of vehicles, and some variables, such as access control type, terrain type and rural or urban, are binary variables. Three types of animal habitats, while-tailed deer, mule deer and elk, are included in the variables because deer and elk are the most common animals involved in AVCs within Washington State.

The minimum, maximum, mean, and standard deviation (S.D.) are shown in the last four columns. One can find that both the reported AVC data and the carcass removal data are over-dispersed.

Table 5-1 Description of explanatory variables in the models

	Variable	Minimum	Maximum	Mean	S.D.
X[a]	Number of reported AVCs per segment [c]	0	22	0.24	0.81
Y[b]	Number of carcasses per segment [c]	0	95	0.94	3.88
z1	Annual average daily traffic (in thousands)	0.31	148.8	13.85	19.76
z2	Restrictive access control (Yes: 1; No: 0)			0.24	
z3	Posted speed limit (mph)	20	70	52.76	10.79
z4	Truck percentage (%)	0	52.28	14.05	8.29
z5	Median width (feet)	0	60	7.9	15.62
z6	Total number of lanes for both directions	1[d]	9	2.79	1.24
z7	Roadway length (feet)	0.01	6.99	0.22	0.4
z8	Terrain type (Rolling: 1; Otherwise:0)			0.720	
z9	Terrain type (Mountainous:1; Otherwise:0)			0.096	
z10	Lane width (feet)	10	20	12.5	1.88
z11	Left shoulder width (feet)	0	18	2.44	2.04
z12	Right shoulder width (feet)	0	20	4.03	3.52
z13	Rural or Urban (Urban:0; Rural:1)			0.758	
z14	White-tailed deer habitat (Yes: 1; No: 0)			0.31	
z15	Mule deer habitat (Yes: 1; No: 0)			0.51	
z16	Elk habitat (Yes: 1; No: 0)			0.31	

[a] Reported AVC data record; [b] Carcass removal data record; [c] Dependent variable; [d] Six out of 10475 segments have only one lane.

Table 5-2 is a cross-tabulation for the AVC data. Each cell represents the number of roadway segments that have corresponding numbers of AVC records in the reported AVC data and the carcass removal data in the five year study period. For example, 63 in the fifth row and the first column (the (0, 4) cell) indicates there are 63 roadway segments with four records in the carcass removal data and zero record in the reported AVC data. From this table, one can find that most roadway segments have zero records in both data sets. That is the (0, 0) cell has the largest number. It is reasonable in that most segments do not have AVCs observed during the study period. Among segments having at least 1 record in both the reported AVC and carcass removal data sets, the (1, 1) and (1, 2) cells contain the largest numbers of segments. Similarly, among the records with at least 2 records in each data set the (2, 2) cell contains the most records. Thus, the diagonal cells, cells (0, 0), (1, 1), and (2, 2) should be expected to play important roles in the data sets.

5.2 Model Estimation

To compare different models, the model details and evaluation criterion for double Poisson, bivariate Poisson, diagonal inflated bivariate Poisson (DIBP) and zero-inflated double Poisson models are listed in Table 5-3. In order to compare the effect of different J values on the diagonal

inflated bivariate Poisson model, three models— DIBP0, DIBP1 and DIBP2—with different J values are also estimated. Table 5-3 shows the details about the variables used for λ_k, the value of J as well as the number of parameters in each model. The insignificant variables were removed from the variables lists. For example, four variables ($z5$, $z10$, $z14$, $z15$), insignificant for λ_1, were removed from all the six fitted models (noted as "-$z5$-$z10$-$z14$-$z15$" in Table 5-3).

One can see that the bivariate Poisson model has a better fit than the traditional double Poisson model. The diagonal inflated bivariate Poisson and zero-inflated double Poisson models generally have better fits because they take the zero inflated portion into account. Overall, the DIBP1 model is considered the best-fitted model because it has the highest ρ^2, lowest AIC and lowest BIC. In comparison with DIBP1, DIBP2 does not show any improvement in its log-likelihood even when the J value becomes larger. That is because the DIBP2 model cannot benefit from the additional diagonal cell when the number of records in the (2, 2) cell of Table 5-2 is relatively low. Therefore, the selection of the J parameter should depend on the diagonal cell values in the AVC-carcass removal cross-tabulation as well as goodness-of-fit measures. The mixing proportions (p_m) in the last column indicate that the data in the diagonal of the AVC cross-tabulation should be over 66%. This result is also consistent with the statistical result in Table 5-2 where the sum of the diagonal value is about 79% of the total data.

Table 5-4 shows estimated values of θ and λ in the diagonal inflated bivariate Poisson models. θ values represent the proportion of the corresponding diagonal cells in the mixing proportion data; λ values denote the proportion of the three regions in Figure 4-1. All models have $\theta_0 > 0.99$,

indicating that more than 99% of mixing proportion data has zero AVC record and less than 1% of mixing proportion data has at least one AVC record for both data sets. This result is consistent with the statistics in Table 5-2 where both datasets have large numbers (more than 6698) of zero-accident roadway segments. Note that the value of λ_3 represents the average number of overlapped records per road segment. For the DIBP1 results, the overlapping percentage in the reported AVC data is about 13% (0.0664/(0.0664+0.4605)).

Table 5-5 shows the coefficient, standard deviation, *t*-value, and average elasticity values for each explanatory variable for λ_1, λ_2, and λ_3. All the listed variables are statistically significant at a 5% significance level.

Table 5-2 Cross-tabulation for AVC and CR data

		Number of Reported AVCs								Cumulated Record
		0	1	2	3	4	5	6	>6	
Number of Carcasses	0	**6698**	361	77	21	10	3	2	2	7174
	1	301	**67**	22	10	6	3	0	1	410
	2	228	69	28	5	6	2	2	0	340
	3	81	35	9	7	1	0	0	1	134
	4	63	26	10	5	1	2	1	0	108
	5	35	17	8	1	2	1	0	2	66
	6	26	17	7	7	0	0	2	0	59
	7	15	14	7	4	1	1	2	0	44
	8	17	8	7	4	2	0	0	0	38
	>8	81	64	43	31	22	13	10	16	280
Cumulated Record		7545	678	218	95	51	25	19	22	8653

Table 5-3 Details for the six fitted models

	Model details					Evaluation Criterion				
	λ_1	λ_2	λ_3	J	Par.	LL	ρ^2	AIC	BIC	p_m
DP	-z5-z10-z14-z15	-z1- z5-z10-z15-z16	—	—	25	-21802	0.313	43654	43852	—
BP	-z5-z10-z14-z15	-z1- z5-z10-z15-z16	-z5- z6-z10-z12-z15	—	37	-21173	0.333	42421	42715	—
DIBP0	-z5-z10-z14-z15	-z1- z5-z10-z15-z16	-z5- z6-z10-z12-z15	0	38	-17283	0.456	34642	34944	0.6612
DIBP1*	-z5-z10-z14-z15	-z1- z5-z10-z15-z16	-z5- z6-z10-z12-z15	1	39	-17275	0.456	34628	34938	0.6637
DIBP2	-z5-z10-z14-z15	-z1- z5-z10-z15-z16	-z5- z6-z10-z12-z15	2	40	-17275	0.456	34630	34948	0.6637
ZIDP	-z5-z10-z14-z15	-z1- z5-z10-z15-z16	—	—	27	-17415	0.451	34884	35099	0.6659

*Best-fitted model; -z5-z10-z14-z15 indicates variables z5, z10, z14, and z15 are removed from the model; (—): the parameter is set zero; *LL*: Log-likelihood; Par.: number of parameters; ρ^2: log-likelihood ratio index(Ben-Akiva and Lerman, 1985); AIC: Akaike's information criterion (Akaike, 1974); BIC: Bayesian information criterion (Schwarz, 1978； Liddle, 2007); p_m: mixing Proportion; DP: Double Poisson; BP: Bivariate Poisson; DIBP: Diagonal Inflated Bivariate Poisson; ZIDP: Zero-inflated Double Poisson.

Table 5-4 Estimated values of θ and λ in DIBP models

Models	θ estimation			Mean of parameter λ		
	θ_0	θ_1	θ_2	λ_1	λ_2	λ_3
DIBP0	1	—	—	0.4608	1.9205	0.0659
DIBP1*	0.9976	0.0023	—	0.4605	1.9359	0.0664
DIBP2	0.9976	0.0023	0.0000	0.4606	1.9359	0.0664

*Best-fitted model; DIBP: Diagonal Inflated Bivariate Poisson.

Table 5-5 The DIBP1 model for AVC

Explanatory variables	λ_1				λ_2				λ_3			
	Coeff	st.err	t-value	E_1	Coeff	st.err	t-value	E_2	Coeff	st.err	t-value	E_3
Constant	-2.904	0.093	-31.067	—	-3.164	0.101	-31.369	—	-26.763	2.769	-9.665	—
Annual average daily traffic (in thousands)	0.013	0.001	13.556	0.202	—	—	—	—	0.069	0.005	14.827	1.072
Restrictive access control (Yes: 1; No: 0)	-1.141	0.032	-35.753	-2.130	-0.986	0.062	-15.973	-1.680	-2.036	0.136	-14.988	-6.660
Posted speed limit (mph)	0.043	0.001	30.302	2.327	0.060	0.002	33.129	3.247	0.068	0.007	10.298	3.680
Truck percentage (%)	-0.049	0.001	-39.589	-0.634	-0.011	0.003	-4.055	-0.142	-0.069	0.004	-16.417	-0.892
Total number of lanes	-0.198	0.020	-9.761	-0.592	-0.395	0.017	-22.882	-1.180	—	—	—	—
Roadway segment length (feet)	0.499	0.009	58.069	0.105	0.471	0.028	17.042	0.099	0.912	0.030	30.785	0.192
Terrain type (Rolling: 1; Otherwise: 0)	-0.302	0.029	-10.543	-0.353	0.105	0.044	2.417	0.100	-1.925	0.096	-20.152	-5.855
Terrain type (Mountainous: 1; Otherwise: 0)	-0.958	0.037	-25.646	-1.606	-0.182	0.066	-2.755	-0.200	-2.027	0.182	-11.159	-6.591
Left shoulder width (feet)	0.036	0.004	9.718	0.189	0.038	0.004	8.836	0.199	0.092	0.012	7.416	0.482
Right shoulder width (feet)	0.034	0.003	12.466	0.310	0.032	0.003	11.340	0.291	—	—	—	—
Rural or Urban (Urban:0; Rural:1)	0.560	0.046	12.114	0.424	0.780	0.049	15.790	0.591	19.984	0.232	86.172	15.140
White-tailed deer habitat (Yes: 1; No: 0)	—	—	—	—	0.973	0.088	11.005	0.622	1.607	2.743	0.586	0.800
Elk habitat (Yes: 1; No: 0)	0.203	0.018	11.162	0.184	—	—	—	—	1.417	0.078	18.102	0.758

E_1 average elasticity value for λ_1; E_2 average elasticity value for λ_2; E_3 average elasticity value for λ_3.

5.3 Model Interpretation

Table 5-5 shows the DIBP1 model results, in which factors contributing to AVCs are identified. The positive values of the coefficients indicate that the increase of each of these explanatory variables increases the probability of AVC occurrences. Conversely, negative values of the corresponding coefficients show that the increases of these explanatory variables lower the probabilities of AVCs. In contrast to regular Poisson accident models, the diagonal inflated bivariate Poisson model contains three dependent variables: λ_1, λ_2 and λ_3. λ_1 and λ_2 quantify the effects on the reported AVC and the carcass removal portions, respectively, whereas λ_3 accounts for the combined effects on the overlapping carcass removal and the reported AVC data sets. The significance and interpretation of the explanatory variables for each dependent variable are discussed below.

Among the traffic elements, three variables are found to significantly contribute to the occurrence of AVCs. The estimated coefficients show that the variable of speed limit is the most significant variable affecting the occurrence of AVCs (λ_1: coef.= 0.043, t=30.302, E_1= 2.327; λ_2: coef.=0.06, t=33.129, E_2= 3.247; λ_3: coef.=0.068, t=10.298, E_3= 3.680). The elasticity values here show that a 1% increase in posted speed limit increases the λ_k by 2.327% for λ_1, 3.247% for λ_2, and 3.680% for λ_3. Higher speed limits tend to increase the likelihood of AVCs. This may be

because drivers travel at higher speeds under a higher speed limit, and high-speed vehicles require longer stopping distances. Therefore, drivers may not be able to stop quickly enough to avoid colliding with an animal on the road. This finding is consistent with most AVC related research that has concluded that speed limits have an increasing relationship with the AVC rates (Allen and McCullough, 1976; Rolley and Lehman, 1992).

AADT is found to have increasing effects on λ_1 (Coef.= 0.013, t=13.556, E_1= 0.202) and λ_3 (Coef.=0.069, t=14.827, E_3= 1.072), but have no significant effect on λ_2. This may be because that AVCs are more likely to be reported on a highway segment with heavier traffic since more travelers can observe and therefore call and report the AVCs. Meanwhile, once an AVC happens, the carcass could be removed by other agencies or persons other than WSDOT. This explains the reason why the AADT variable also contributes to AVC occurrences in the overlapping portion of the reported AVC and carcass removal data. Overall, a higher AADT increases the chance of AVCs because a higher volume elevates the level of accident exposure and shortens vehicle headways needed for animals to cross the road. This result is consistent with the accident research conducted by Chin and Quddus (2003).

A higher truck percentage is found to decrease the likelihood of reported AVCs and carcass removal for all λ's. One reason may be that drivers are more cautious when more trucks are on the road. Another reason may be that trucks are usually associated with louder noise which may scare animals away. Trucks also tend to have better driver visibility forward, which could provide more time for drivers to react. This result is similar to the motor vehicle accident

research by Milton and Mannering (1998), which identified a decreasing relationship between truck percentage and accident probability.

Among the geometric design elements, five variables are significantly associated with the occurrence of AVCs. Roadway segments with restrictive access control tend to have lower accident risk with fairly significant t-ratios for λ_1 (t=-35.753), λ_2 (t=-15.973), and λ_3 (t=-14.988). Usually, highways with higher restrictive access control (e.g. interstates) also have more physical obstructions along the highway that may limit the crossing of animals. In this case, animals find it more difficult to access highways protected by physical obstructions, and consequently, the number of AVCs is smaller for the highways with more restrictive access control.

The variable, total number of lanes, is found to be significant at a 5% significance level for λ_1 (t=-9.761) and λ_2 (t=-22.882) but not λ_3. With an increase in the total number of lanes, the roadway becomes wider, increasing the crossing difficulty for animals. Thus, wider road segments may be less attractive for animals to cross and hence reducing the likelihood of AVCs.

As expected, longer roadway segment length appears to increase the occurrence of AVCs (λ_1: coef.= 0.499, t=58.069, E_1= 0.105; λ_2: coef.=0.471, t=17.042, E_2= 0.099; λ_3: coef.=0.912, t=30.785, E_3= 0.192). This may be because the longer the roadway segment is, the more likely it is to segment animal habitats, between which animals will move. Similarly, more vehicle-miles are traveled on longer segments for the same traffic, number of lanes, etc. For the same per vehicle-mile traveled or per segment mile of length risk of collision a longer segment increases

the total AVCs.

Both left and right shoulders are found to have an increasing effect on λ_1 (t=9.718 for the left, t=12.466 for the right) and λ_2 (t=8.836 for the left, t=11.340 for the right). Generally, drivers may have a broader view on roadways with shoulders. However, the results indicate that shoulders do not give drivers enough time to react to the appearance of animals because drivers tend to drive faster on segments with shoulders.

In terms of area types, three variables have significant impacts on the occurrences of AVCs: rolling area, mountainous area, and rural areas. In comparison with level terrain, rolling areas are associated with low numbers of reported AVCs (λ_1: coef=-0.302, t=-10.543, E_1= -0.353) and AVCs in the overlapping portion of the reported AVC and carcass removal data sets (λ_3 : coef.=-1.925, t=-20.152, E_3= -5.855). However, rolling areas are found to be associated with a higher number of carcasses (λ_2: coeff=0.105, t=2.417, E_2= 0.100) than level terrain areas. The contradiction in the estimated coefficient values may imply that AVCs occurred in rolling areas are under reported compared with those occurred in level terrain areas.

In comparison with level terrain, mountainous areas tend to have a low likelihood of AVCs (λ_1: coef.= -0.958, t=-25.646, E_1=-1.606; λ_2: coef.=-0.182, t=-2.755, E_2=--0.200; λ_3: coef.=-2.027 t=-11.159). This may be because in mountainous areas, people drive more carefully, and vehicles are also slower. Another possible reason is that carcasses may not be easily found or require removal when they come to rest in areas off of roadways. Similarly, roadways in mountainous

terrain tend to be in valleys and tunnels which may limit the coverage of cell phone for reporting. Some might argue that animals should be also active in the mountainous areas, resulting in more AVCs. However, the valleys and tunnels may also impede animals' movements because these geometric characteristics may physically separate different habitats. Therefore, animal crossing activities could be reduced. Moreover, animal migration in mountains is also more predictable because of larger herds. WSDOT will often place warnings during the migration period to reduce crashes during peak animal travel.

Compared to highways in urban areas, those in rural areas are found to have more reported AVC and carcass removal records in both data sets (λ_1: coef=0.560, t=12.114, E_1= 0.424; λ_2: coef=0.780, t=15.790, E_2= 0.591; λ_3: coef=19.984, t=86.172, E_3= 15.140). This is under the expected values because animals are more active and populated in rural areas. However, looking at the overlapping portion of two data sets, this "rural effect" is more obvious (λ_3: coef=19.984, t=86.172, E_3= 15.140). This result highlights rural AVCs as a potential focus for future AVC research.

In terms of high density animal distribution areas, white-tailed deer habitat is associated with a higher λ_1 (t=11.005), as expected. Elk habitat is also found to have an increasing impact on λ_1 (t=11.162) and λ_3 (t=18.102). It makes sense that the areas with higher density animal distribution tend to have a higher AVC rate. However, mule deer habitat is not found to significantly affect the likelihood of AVCs. One main reason may be that the mule deer population distribution in Washington State is relatively uniformly and widely distributed and

covers a large portion of the study routes.

In summary, speed limit, restrictive access control, and roadway segment length are the most significant explanatory variables affecting all λ's (the absolute values of their t ratios are over 10). According to the average elasticity values, rural area, restrictive access control, and terrain type (rolling or mountainous) have the most significant marginal effects on λ_3 (the absolute values of their average elasticity values are all over 5%). It should be noted that the posted speed limit is the only variable with the absolute values of all the average elasticity values being over 2% for all λ's. Hence, reducing the posted speed limit could result in a reduction of AVCs effectively.

Based on the analysis above, it suggests that in areas where the highway segments the habitats of non-domestic animals, especially deer, transportation agencies should further examine speed limit and access control options to develop suitable countermeasures. Constructing fences and crossing infrastructure (e.g. tunnels and over bridges) along and within the hot spots could be helpful for connecting segmented animal habitats and preventing animals from interacting with vehicles in the areas with frequent AVCs (Donaldson, 2007).

5.4 Summary

Animal-Vehicle Collision (AVC) is an important roadway-safety concern in many areas around the world. In order to investigate the contributing factors of AVCs, reported AVC and carcass removal data sets are commonly used in previous studies. But these two significantly different data sets are usually analyzed separately. Although the two data sets complement each other, they have not been analyzed jointly. This research applies diagonal inflated bivariate Poisson (DIBP) regression models to fit these data sets concurrently. As an inflated version of the bivariate Poisson regression model, the DIBP models outperformed other models (double Poisson, bivariate Poisson, diagonal inflated bivariate Poisson, and zero-inflated double Poisson) studied in this research. The DIBP models are the best fitted models with the lowest AIC and BIC values. Functionally, the DIBP models not only can handle under- or over- dispersed count data but also can model paired data sets with correlation.

The contributing factors of AVCs are identified after the implementation of DIBP models. Three dependent variables (λ_1, λ_2 and λ_3) are each linked with a group of explanatory variables including traffic elements, geometric design factors, and geographic characteristics associated with AVCs. Two traffic elements, speed limit and AADT, and two geometric design factors, shoulder width and roadway segment length, are found to have an increasing effect on the likelihood of AVCs. In terms of the variables of geographic characteristics, rural area segments tend to have higher numbers of reported AVCs and carcass removals. The areas with dense

animal distributions, such as white-tailed deer and elk habitats, are also found to increase the occurrence probability of AVC.

In this study, the DIBP models have been found effective in modeling AVCs. The methodology developed in this study may be applied to model other types of accident with two datasets of similar characteristics. Since the datasets used in this study happen to be over-dispersed, DIBP's capability for handling under-dispersed data may be demonstrated in future studies. Moreover, comparisons between the DIBP models and other multi-variate models, such as bivariate negative binomial and bivariate Poisson-lognormal models, will be desired extensions of this study.

Although the DIBP models are effective in predicting and assessing contributing factors of AVCs using concurrently the reported AVC and carcass removal data sets collected from the ten study routes in Washington State, more data are needed to further investigate the approach and accident causation. Transferability testing is also needed when applying this model to different animal types or locations. Moreover, it will be more desirable to investigate the potential contribution of time factors (e.g. day vs. night) in the future research since AVCs associated with a specific animal type may be more frequent in certain periods of a day.

Chapter 6. Modeling Animal-Vehicle Collisions Considering Animal-Vehicle Interactions

6.1 Data Description

Three major data sources are used in this chapter for the modeling considering Animal-Vehicle interactions:

- Carcass removal data by Washington State Department of Transportation (WSDOT) stores the information of animal carcass being collected. The information includes location (by milepost), date, weather, animal type, sex, age, etc. Carcass removal data have been commonly used in AVC research (Reilley and Green, 1974; Allen and McCullough, 1976; Knapp and Yi, 2004, Lao et al., 2010). This study used five years (2002-2006) of carcass removal data from ten highway routes (US 2, SR 8, US 12, SR 20, I-90, US 97, US 101, US 395, SR 525 and SR 970) as the study routes following the recommendation from WSDOT experts.
- Deer distribution data by Washington Department of Fish Wildlife (WDFW) is in the form of GIS-based maps for mule deer, white-tailed deer, and elks.
- Roadlog data by Highway Safety Information System (HSIS) provides geometric information for the roadway, such as median width, number of lanes and shoulder width.

Table 6-1 lists all explanatory variables used in the modeling process. Most of the quantitative and dummy variables were directly selected from the combined dataset. Several variables were created based on the observed data. For example, the variable "Speed Level" was created based on posted speed limits. This variable is a dummy variable. The variable is set to 1 when the posted speed limit is greater than 50 mph and 0 otherwise. This is because a dramatic increase in AVCs was found when the speed limit > 50mph. Other examples, such as variables $z14$, $z15$, and $z16$, were created for representing habitats of different types of animal.

The minimum, maximum, mean, and standard deviation (S.D.) of each variable are shown in Table 6-1. One can find that the reported collision data is over-dispersed as indicated by the variance being higher than the mean.

Table 6-1: Description of explanatory variables in the models

	Variable	Min	Max	Mean	S.D.
Y[a]	Number of carcasses per segment [b]	0	16	0.095	0.564
z1	Annual average daily traffic (in thousands)	0.31	148.8	15.11	21.07
z2	Restrictive access control (Yes: 1; No: 0)			0.24	
z3	Speed level (>50mph: 1; otherwise: 0)			0.68	
z4	Truck percentage level (>5%: 1; otherwise: 0)			0.78	
z5	Median width (> 6 feet: 1; others: 0)			0.33	
z6	Total number of lanes (in both directions)	2	9	2.96	1.32
z7	Roadway length (mile)	0.01	6.99	0.22	0.4
z8	Terrain type (Rolling: 1; Otherwise: 0)			0.72	
z9	Terrain type (Mountainous: 1; Otherwise: 0)			0.095	
z10	Lane width (feet)	10	20	12.5	1.88
z11	Left shoulder width (feet)	0	18	2.44	2.04
z12	Right shoulder width (feet)	0	20	4.03	3.52
z13	Rural area (Rural: 1; Urban: 0)			0.76	
z14	White-tailed deer habitat (Yes: 1; No: 0)			0.31	
z15	Mule deer habitat (Yes: 1; No: 0)			0.51	
z16	Elk habitat (Yes: 1; No: 0)			0.31	
z17	Sex of animal (Male: 1; Female: 0)			0.328	
z18	Horizontal curve (Curve degree>3: 1; otherwise: 0)			0.16	
z19	Vertical curve (Grade percentage>3%: 1; otherwise: 0)			0.22	

[a] Specific to carcass removal data only; [b] Dependent variable, number of carcasses within two years (2005-2006); Min: Minimum; Max: Maximum; S.D.: standard deviation

6.2 Model Estimation

For the purpose of comparison, both a Poisson regression model (Equation (4-22), when δ approaches zero in Equation (4-26)) and a negative binomial regression model (Equation (4-26)) were produced for the MP and VAIP model estimation using the carcass removal data. An open source statistical analysis package, R (Institute for Statistics and Mathematics, 2011), was used for model estimation in this research.

Table 6-2 shows the coefficients of explanatory variables and statistical test results of the convergence MP model, estimated by negative binomial regression. Variables significantly associated with the probability of a hazardous crossing of an animal, P_o and the probability of the driver's ineffective response, P_{vf}, are shown as the explanatory variables in the models.

Table 6-2: Description of explanatory variables in the MP model

Explanatory variables	Coeff[a]	st.err[b]	t-value
Variables affecting the probability of a hazardous crossing of an animal (P_o)			
Constant	-16.359	0.268	-60.945
Median width (> 6 feet: 1; others: 0)	-1.016	0.137	-7.444
Total number of lanes	-0.290	0.057	-5.119
Terrain type (Rolling: 1; Otherwise: 0)	0.248	0.070	3.525
Rural area (Rural: 1; Urban: 0)	1.890	0.133	14.197
White-tailed deer habitat (Yes: 1; No: 0)	1.516	0.056	26.963
Animal sex (Male: 1; Female: 0)	-0.720	0.056	-12.876
Variables affecting the probability of ineffective response of the driver (P_{vf})			
Speed level (>50mph: 1; otherwise: 0)	1.954	0.277	7.042
Truck percentage Level (>5%: 1; otherwise: 0)	-1.219	0.183	-6.646
Model Evaluation			
AIC at base model[#]			26,861
AIC at convergence with Poisson regression			19,653
AIC at convergence with NB regression (δ =1.66)			17,177
ρ^2			0.36

[a]coefficients in the model; [b]standard error; ρ^2 was calculated by comparing the log-likelihood with the base model; base model [#]: δ approaches zero and β=0.

Similarly, the coefficients of the explanatory variables and their significance are shown for the VAIPM model in Table 6-3. In addition to the probabilities, P_o and P_{vf}, the probability of the animal's failure to escape from being hit, P_{af}, is explicitly formulated. One variable, the sex of animal, is identified significant by P_{af}. Additionally, to fully understand the marginal effects of each independent variable, their elasticity values are calculated as (Shankar et al., 1995;Abdel-Aty and Radwan, 2000; Washington et al., 2003):

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\partial x_{ik}} \frac{x_{ik}}{\lambda_i} = \beta_k x_{ik} \tag{6-3}$$

where λ_i is the expected number of accidents for roadway segment i, x_{ik} is the k-th variable in the vector of explanatory variables for roadway segment i, and β_k is the corresponding coefficient of the k-th variable. The elasticity in Equation (25) applies when the explanatory variable $\boldsymbol{x}_{ik}$ is continuous. In case of an indicator variable, pseudo-elasticity is estimated as an approximate elasticity of this variable (Washington et al., 2003):

$$E_{x_{ik}}^{\lambda_i} = \frac{Exp(\beta_k) - 1}{Exp(\beta_k)} \tag{6-4}$$

Table 6-3: Description of explanatory variables in the VAIP model

Explanatory variables	Coeff[a]	st.err[b]	t-value	E[c]
Variables affecting the probability of a hazardous crossing of an animal (P_o)				
Constant	-15.666	0.268	-58.363	
Median width (> 6 feet: 1; others: 0)	-1.016	0.137	-7.444	-1.762
Number of lanes	-0.290	0.057	-5.119	-0.336
Terrain type (Rolling: 1; otherwise: 0)	0.248	0.070	3.525	0.220
Rural area (Rural: 1; Urban: 0)	1.890	0.133	14.197	0.849
White-tailed deer habitat (Yes: 1; No: 0)	1.516	0.056	26.963	0.780
Variables affecting the probability of the animal failure to escape from being hit (P_{af})				
Animal sex (Male: 1; Female: 0)	-1.134	0.074	-15.347	-2.108
Variables affecting the probability of ineffective response of the driver (P_{vf})				
Speed level (>50 mph: 1; otherwise: 0)	1.954	0.277	7.042	0.858
Truck percentage level (>5%: 1; otherwise: 0)	-1.219	0.183	-6.646	-2.384
Model Evaluation				
AIC at base model[e]			26,861	
AIC at convergence with Poisson regression			19,653	
AIC at convergence with standard NB regression[d]			19,484	
AIC at convergence with NB regression (δ =1.66)			17,177	
ρ^2			0.36	

[a]Coefficients in the model; [b]standard error; [c]average elasticity value; δ is referred to as the overdispersion parameter; ρ^2 was calculated by comparing the log-likelihood with the base model; base model [e]: δ approaches zero and β=0; standard NB regression; [d]the traditional NB regression model with the standard structure.

6.3 Model Interpretation

The estimated coefficients, their t–values, and GOF for the MP model and the VAIP model are shown in Tables 6-2 and 6-3 respectively. Comparing the estimation results from Tables 6-2 and 6-3, one can find that the GOF of these two models are almost the same: both the adjusted ρ^2 values are 0.36, and the AIC values are undistinguished. Based on the AIC values within table 6-2 or 6-3, the negative binomial regression outperformed the corresponding Poisson regression. The estimate results show that the δ value is 1.66 in both MP and VAIP models and their p value is 0.00, which verifies that δ is significantly greater than 0, and the carcass removal data are over-dispersed. In this case, the model estimated with Poisson regression should not be used because it requires the mean and variance of the carcass removal data to be the same. Model estimated with the NB regression is a better choice for this study.

For both the MP and VAIP models estimated by the NB regression, a total of eight variables are identified as significant, including the number of lanes, terrain type, rural area, white-tailed deer habitat, median width, sex of animals, "truck percentage level", and "speed level". Among them, two variables, "truck percentage level" and "speed level", have significant impacts on P_{vf}, the probability of drivers' ineffective response, and the other six variables play significant roles in determining P_o, the probability of encountering a disturbance animal in the MP model. However, in the VAIP model, one variable, sex of animal, is explicitly identified as significant by P_{af}, the

probability of the animal's failure to escape from being hit, instead of P_o in the MP model. Although both models show the similar GOF, further analyses show that the VAIP model demonstrates more capability of interpreting the AVC process and the impacts of explanatory variables. Therefore, the detailed explanations and discussions regarding the VAIP model follows.

6.3.1 Interpretation of Estimation Results for P_o

The five significant variables affecting the probability of an animal's presence reflects both roadway geometric characteristics and animal distribution features as shown in Table 6-3. Compared with the level terrain type, rolling terrain tends to have an increasing effect on the possibility of the presence of an animal on the road P_o (Coef. = 0.248, t = 3.525, E = 0.220). This may be because rolling terrain has a higher animal population than that of level terrain. The elasticity value here shows that an incremental change of 0.22% to the AVC accident risk is caused by the changes from level terrain to rolling terrain. Similarly, compared to the highways in urban areas, those in rural areas also tend to have a higher P_o (Coef. = 1.890, t = 14.197, E = 0.849). This may also be due to the higher animal population and activity levels in rural areas. The elasticity value here shows that an incremental change of 0.849% to the AVC accident risk is caused by the changes from urban area to rural area.

Among all the variables, white-tailed deer habitat was found to be the most significant

explanatory variable affecting AVCs (Coef. = 1.516, t = 26.963, E =0.780). This may be due to the higher animal population in the white-tailed deer habitat, contributing to the increased probability of animal crossing P_o. If a highway section segments a white-tailed deer habitat area, a driver using this section will have a higher probability of encountering an animal. Compared with white-tailed deer habitats, the variable of elk habitats is insignificant at 95% significance level. This can be explained by the fact that the total number of collisions with elk only contributes a small part of the whole AVC records for the study period. Mule deer habitat also was insignificant in the model. The reason for this may be because the mule deer habitat distribution is relatively uniformly in Washington State and covers a large portion of the study routes. The elasticity value for the white-tailed deer habitat indicates an incremental change of 0.780% on the AVC accident risk caused by the changes from other areas to white-tailed deer habitat areas. The finding is consistent with another AVC study (Lao et al., 2011a).

The number of lanes is the significant factor having a negative effect on the presence of animals, P_o (Coef. = -0.290, t = -5.119, E = -0.336). With an increase in the total number of lanes, the probability of animals present on the road tends to be lower. This is understandable because roadway sections with more travel lanes are typically wider, which might increase the crossing difficulty for animals. Therefore, animals would be reluctant to cross a wider segment and thus the P_o is lower. The elasticity value here shows that a 1% increase in the number of lane decreases the AVC accident risk by 0.336%. In addition, wider lanes also provide additional area for evasive maneuvers, which will reduce the AVC risk.

The variable of median width is related to roadway geometric design elements. A median width of greater than 6 feet was found to have a significant decreasing effect on P_o (Coef. = -1.016, t = -7.444, E = -1.762). This variable is similar to the number of lanes in that a wider median will increase the crossing hesitation for animals, and hence reduce the likelihood of AVCs. The elasticity value here shows a decrement change of 1.762% on the AVC accident risk caused by the changes from median width less than 6 feet to median width more than 6 feet.

6.3.2 Interpretation of P_{vf}

Among the factors affecting the probability of the driver's ineffective response, P_{vf}, two explanatory variables, "Speed level" and "Truck percentage level", were found to be significant. The speed limit level has a positive estimated coefficient (Coef. = 1.954, t = 7.042, E = 0.858). This implies that when a highway segment had a speed limit greater than 50 mph, the probability of a driver's ineffective response would increase. A vehicle running at a higher speed requires a longer stopping distance. Hence, when an animal is perceived, the distance traveled under normal reaction time for a faster vehicle is increased and therefore the time before impact is decreased. This explains why speed limit has an increasing effect on P_{vf}. This finding is consistent with many previous AVC related studies, e.g. Rolley and Lehman (1992) and Allen and McCullough (1976). The elasticity value here indicates an incremental change of 0.858% to the AVC accident risk is caused by the changes from the highways of speed limit lower than 50 mph to the highways of speed limit higher than 50 mph.

The truck percentage level was found to have an increasing impact on the probability of driver's effective response (decreased failure to avoid collision, Coef, = -1.219, t = -6.646, E = -2. 384). This is presumably because truck drivers drive at relatively lower speeds. Trucks have taller profiles, which allow their drivers to have longer sight lines and better visual abilities. Another possible reason may be that trucks usually make more noise, which can drive animals away from the roadway, though more research will be necessary to confirm this finding and the mechanic by which it functions. This result is supported by the motor vehicle accident research (Milton and Mannering, 1998) in which the increase in the percentage of trucks may decrease the accident probabilities. The elasticity value here indicates a decrement change of 2.384% to the AVC accident risk is caused by the changes from the areas with lower truck percentage to the areas with higher truck percentage.

6.3.3 Interpretation of P_{af}

Turning to the factors affecting the probability of animal's response, P_{af}, one variable, sex of animal, were found to affect P_{af} significantly. Compared with female animals, male animals tend to have lower collision risk (Coef. = -1.025, t = -12.877, E = -1.787). This may be because male animals require less response time than female animals. The elasticity value here indicates a decrement change of 1.787% on the AVC accident risk caused by the change from female animals to male animals. The modeling capability of the MP model is extended by the item, P_{af} to explicitly explain unique animal response behavior with different attributes. For instance,

animal species and gender may play significant roles in determining their reactions when a vehicle is approaching. Some animal species may detect the approaching vehicles much earlier than the others. Male animals may respond and run faster than females. The proposed VAIPM model is capable of capturing specific animal responses in an AVC and enhances the MP model's ability in data interpretation. Due to data constraints, animal species data are not available for model calibration and estimation. Table 6-3 shows that sex of animal is considered significant in describing animal response behavior.

6.4 Spatial and Temporal Transferability Test

The relationship between AVCs and their associated factors may change temporally and spatially. Thus, a concern with the model is whether its estimated coefficients are transferable spatially or temporally. When testing spatial and temporal transferability, the following likelihood ratio test can be conducted (Washington et al., 2003):

$$X^2 = -2[LL(\beta_T) - LL(\beta_a) - LL(\beta_b)] \tag{6-1}$$

where $LL(\beta_T)$ is the log likelihood at convergence of the model using the data from both regions (or time periods), $LL(\beta_a)$ is the log likelihood at convergence of the model using the data from a region (or time period a), $LL(\beta_b)$ is the log likelihood at convergence of the model using the data

from b region (or time period b). This X^2 statistic is a χ^2 distribution with degrees of freedom equal to the summation of the number of coefficients in region a and region b minus the number of coefficients in the overall model.

We statistically tested spatial and temporal transferability for the model. Table 6-4 shows the transferability test results. The number of segments, number of accidents, and the log-likelihoods for different data sets are also show in Table 6-4.

For both transferability tests, the null hypothesis was that the coefficients are transferable. For the spatial test, the first data set was routes SR 8, US 12, I-90, US 101, and SR 970, with the second data set having the remaining five routes. For the temporal test, the first data set was the year 2005, and the second the year 2006. Following Equation (26), the data sets were estimated separately and then together. For the spatial test, X^2 was 168 with 9 degrees of freedom, which is greater than 16.92 at a 95% confidence level. For the temporal test, X^2 was 42 with 9 degrees of freedom, which is greater than 16.92 at a 95% confidence level. Thus, the coefficients were found to not be transferable between routes or years.

Table 6-4: Spatial and temporal transferability test results for AI model 3

	# of segments	# of accidents	log-likelihood
	Spatial transferability test		
*First five routes	10,415	1,290	-3,369
#Second five routes	9,993	2,607	-5,132
Overall data	20,408	3,897	-8,585
X^2	$= -2[LL(\beta_T) - LL(\beta_a) - LL(\beta_b)]$=-2[-8,585+5,132+3,369]=168		
	Temporal transferability test		
2005	9,942	2,110	-4,572
2006	10,466	1,787	-3,992
Overall data	20,408	3,897	-8,585
X^2	$= -2[LL(\beta_T) - LL(\beta_a) - LL(\beta_b)]$=-2[-8,585+4,572+3,992]=42		

*First five routes: SR 8, US 12, I-90, US 101, and SR970; #Second five routes: US 2, SR 20, US 97, US 395, and SR 525.

Although the estimated coefficients could not transfer from year to year or from location or location, the significant explanatory variables and their sign (positive or negative) converged from these different data sets are basically exact the same. The poor transferability may be a reflection of the performance and characteristic differences among animals in different time

period or location. Thus, if we want to estimate a more accurate elasticity for different explanatory variables, we need to recalibrate the model using the data set in a particular time and location. However, the impacts from those variables, either being with a decreasing or an increasing factor on the AVCs, remain the same in different time periods and different locations. This implies that the model can still be applied to develop AVC countermeasures in practice.

6.5 Summary

A series of count data models have been used in AVC analysis throughout past studies. However, most of these models used in vehicle collisions seldom include human factors or animal characteristics in their analysis process, although these attributes are critical to the occurrence of AVCs. Thus, further investigation on modeling crash data considering interaction behavior is still desired.

This research presents the microscopic probability (MP) and Vehicle-animal Interaction-based Probability (VAIP) models and their estimation results. Both models consider the probability of drivers' ineffective response and animals' presence. As an improvement, the VAIPM models include a third term, the probability of an animals' response failure to escape, to capture animals' reaction characteristics in AVCs. The test results show that VAIPM models can provide a better explanation of the relationship among human factors, animal distributions, roadway design

factors, and AVCs. Key research findings are summarized as follow:

- Compared with urban areas, the probability for a vehicle to encounter an animal is high in rural areas. This is likely due to the animal population difference between the two areas.
- The probability for a vehicle to hit a deer is much higher when driving on a highway through a white-tailed deer habitat.
- The probability of a driver's ineffective response will increase with speed limit. It goes up significantly when speed limit is greater than 50 mph.
- Compared with female animals, male animals are more alert and have a better chance to escape from potential AVCs.

Results from the model are useful to transportation agencies for determining countermeasures against AVCs. This research recommends that transportation agencies should further examine some key associated variables, including speed limit, and develop suitable countermeasures in the areas where the highway crosses the habitat of non-domestic animals, such as white-tailed deer.

Chapter 7. Application of Generalized (non) Linear Models for Rear End Accident

7.1 Generalized (non) Linear Models for Rear End Accident

7.1.1 Data Description

Rear-end crashes are among the more common accident types, accounting for almost one-third of all reported accidents in the U.S. (National Transportation Safety Board, 2001). In addition to traffic safety issues, when crashes occur, they temporarily reduce roadway capacity and cause congestion. The 2010 Urban Mobility Report indicates the annual delay per person was 37 hours and an average of $808 per traveler resulted from congestion in the 439 surveyed-urban areas in 2009 (Schrank and Lomax 2010). These data clearly illustrate the urgent needs of rear-end crash modeling and analysis for traffic safety improvements. Based on appropriate model specifications, we can identify significant factors to improve highway design and traffic operation regulation development, leading to a decrease in the frequency and severity of rear-end crashes.

To verify the effectiveness of the GNM-based approach, the rear-end crash data from ten highway routes (US 2, SR 8, US 12, SR 20, I-90, US 97, US 101, US 395, SR 525 and SR 970) in Washington State from 2002 to 2006 are used in this study. The rear-end crash data have been

extracted from the Washington State accident file provided by Highway Safety Information System (HSIS), which is operated by the University of North Carolina Highway Safety Research Center and the LENDIS corporation under a contract with Federal Highway Administration (FHWA) (HSIS, 2012). The HSIS collision data of Washington were compiled from the State Trooper filed field reports and citizen reports. Roadlog data provided by HSIS provides geometric information for the roadway, such as terrain type, lane width and shoulder width. Additionally, traffic flow data, and accident-involved vehicle data are also collected from the Washington State Department of Transportation (WSDOT). Table 7-1 lists all explanatory variables used in the modeling process. A total of 2950 crashes are recorded on 10466 highway segments during the five years. The minimum, maximum, mean, and standard deviation (S.D.) of each variable are calculated as shown in Table 7-1. One can find that the number of rear-end accident per mile is over-dispersed as indicated by the variance (25) being higher than the mean (4.7).

Table 7-1: Description of explanatory variables in the models

	Variable	Min	Max	Mean	S.D.
Y	Number of rear end accident per mile [a]	0	900	4.7	25
z1	Annual average daily traffic (in thousands)	0.31	148.8	15.11	21.07
z2	Restrictive access control (Yes: 1; No: 0)	0	1	0.24	—
z3	Speed limit (mph)	20	70	54.1	12.04
z4	Truck percentage	0	60.37	12.9	9.06
z5	Median width (> 6 feet: 1; others: 0)	0	1	0.33	—
z6	Total number of lanes (in both directions)	2	9	2.96	1.32
z7	Roadway length (mile)	0.01	6.99	0.22	0.4
z8	Terrain type (Rolling: 1; Otherwise: 0)	0	1	0.72	—
z9	Terrain type (Mountainous: 1; Otherwise: 0)	0	1	0.095	—
z10	Lane width (feet)	10	20	12.5	1.88
z11	Left shoulder width (feet)	0	18	2.44	2.04
z12	Right shoulder width (feet)	0	20	4.03	3.52
z13	Rural area (Rural: 1; Urban: 0)	0	1	0.76	—
z14	Vertical curve (Grade percentage)	0	9.87	1.039	1.167

[a] Dependent variable; Min: Minimum; Max: Maximum; S.D.: standard deviation

7.1.2 Model Estimation

One of the benefits of a GNM-based approach is to provide the sufficient modeling flexibility to extract the complex relationship between dependent and independent variables. The model specifications and function structures can be defined and developed as needed. For example, when using the variable, AADT, the logarithmic function should be based on the statistical analysis, the observation relationship between AADT and the log (crash rate), and the visual illustration showed in Figure 7-1. Then the parameters for the logarithmic function can be estimated during the model calibration and validation processes. For the other specific explanatory variables, the diverse functions, such as polynomial function, exponential function, parabolic function, logarithmic function, etc. may be utilized to extract proper data features. Thus, the model estimation consists of two interactive steps: 1) Model specification estimation of nonlinear predictors, and 2) Parameter calibration and estimation, which are detailed as follows.

7.1.2.1 Estimation of Nonlinear Predictors

It is necessary to determine the appropriate predictors U_j in Equations (4-35) and (4-36) before the corresponding coefficients w_j are calibrated. To better illustrate the nonlinear predictor estimation process, an example is detailed to formulate the predictor for the variable, AADT, as follows. Assume the number of rear-end crashes follows the Poisson model, and all other

dependent variables are approximately consistent across different AADT levels when the sample data is large enough. To develop an appropriate format of the predictor U_j, the visualized comparisons between the logarithm of the expectation of crash rate (number of crashes per mile) and AADT are illustrated in Figure 7-1. The solid marker line shows the crash rate from ten highways in Washington State in five years (2002-2006), classified by AADT. As we can see, the logarithm of the average crash rate tends to increase when the AADT increases at a variable rate. The increase rate becomes smaller with the higher AADT, which indicates the inappropriateness of using a linear predictor. To address this issue, a logarithmic calculation as the nonlinear predictor is adopted to approximate the impacts of AADT on crash rate:

$$U_A(x)=2.346\ln(x)-3.4057 \tag{7-1}$$

Compared to the linear predictor $L(x)$=0.3008x-1.6054, the value of R^2 increases from 0.8452 to 0.9673 when the nonlinear predictor, $U(x)$, is utilized as shown in Figure 7-1. Therefore, the nonlinear predictor is more suitable to describe the relationship between crash rate and AADT, and is employed in this study. The logarithm of AADT and its impacts on rear-end crash frequencies have been found significant and this results are consistent with many previous studies (e.g. Wong et al., 2007; Abdel-Aty and Haleem, 2011).

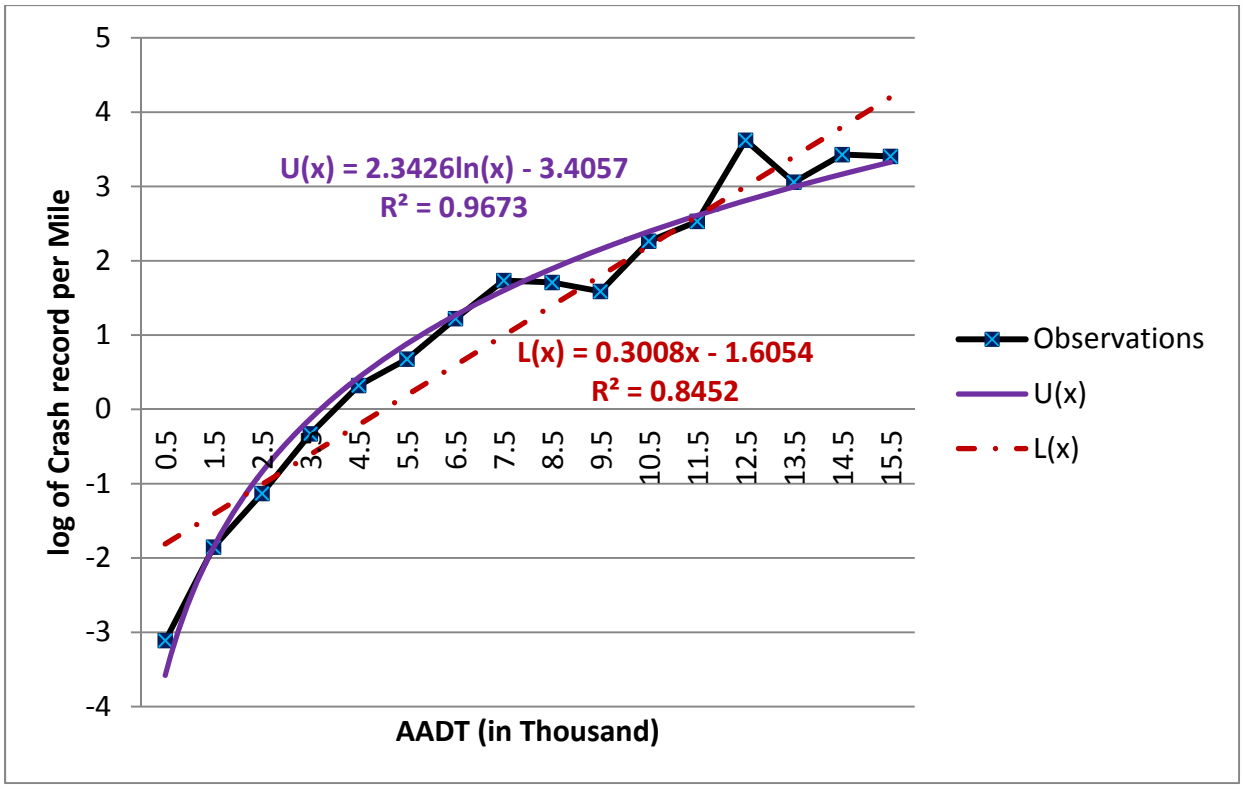


Figure 7-1. Rear end crash rate (crash frequency per mile) in five years (2002-2006) from ten highways in Washington State, by AADT

Similar studies can be conducted to other contributing factors. In addition to the nonlinear predictor for the variable, Grade, is established in Equation (4-34), the other predictor for the variables, truck percentage, and right shoulder width are developed in Equations (7-2), and (7-3) respectively:

$$U_{tr}(x) = \begin{cases} 1.2 + 1.8(x-1)/2 & x \le 3 \\ -1.786\mathrm{Ln}(x/2 - 0.5) + 2.8438 & x > 3 \end{cases} \quad (7\text{-}2)$$

$$U_{rw}(x) = \begin{cases} 0 & x \le 1 \\ 1 & x > 1 \end{cases} \quad (7\text{-}3)$$

For the variables, right shoulder width, the step functions are verified appropriate by using different thresholds. For the other variables, no strong nonlinear associations are observed and linear predictors are sufficient to characterize their impacts on crash rate.

7.1.2.2 Parameter Estimation

For comparison purposes, various GLMs (Equations (4-28) and (4-29)) and GNMs (Equations (4-35) and (4-36)) are established to model and analyze rear-end crash data. The parameters for both GLMs and GNMs are estimated based on Maximum-Likelihood Estimation (MLE) methods. The dispersion parameter, Ψ, in Equation (4-30) is obtained based on the Method of Moments Estimator (MME). To illustrate their significantly different performance between GLM and GNM, a specific Quasi-Poisson regression model with a logarithm link function is estimated for both GLM and GNM. An open source statistical analysis package, R (Institute for Statistics and Mathematics, 2011), was used for model estimation and calibration in this study.

The explanatory variables were analyzed based their significance in the models. Residual deviance is utilized to examine and measure the model's goodness-of-fit. Table 7-2 shows the coefficients of significant variables and statistical test results of the Quasi-Poisson GLM. The calculated elasticity for each variable listed in the last column. Similarly, the coefficients of the significant variables and their significance are shown for the Quasi-Poisson GNM in Table 7-3.

The variable, Grade is identified significant in the Quasi-Poisson GNM addition to the other variables recognized by the GLM. Rather than the deterministic elasticity values calculated for the GLM significant variables, the sign of the elasticity values for the nonlinear predictors of the GNM variables are changeable depending on their corresponding values and function specification. Equations (7-4) and (7-5) detail their values which are further interpreted in the following section. Table 7-4 shows the F-test for both Quasi-Poisson GLM and GNM. The F-value and its significant level of p=0.000 indicate that in terms of the goodness-of-fit measure of residual deviance, the difference between the Quasi-Poisson GLM and GNM is statistically significant, and the Quasi-Poisson GNM outperforms its GLM counterpart.

$$E_G = \begin{cases} 0 & -7 \le \boldsymbol{x}_{ij} \le -5.5, \boldsymbol{x}_{ij} = -4.4, -3.5 \le \boldsymbol{x}_{ij} \le 3.5, \boldsymbol{x}_{ij} = 4, 4.5 \le \boldsymbol{x}_{ij} < 7 \\ 2.17\boldsymbol{x}_{ij} & -5.5 < \boldsymbol{x}_{ij} < -4.4 \\ -1.14\boldsymbol{x}_{ij} & -4.4 < \boldsymbol{x}_{ij} < -3.5 \\ 1.17\boldsymbol{x}_{ij} & 3.5 < \boldsymbol{x}_{ij} < 4 \\ -0.78\boldsymbol{x}_{ij} & 4 < \boldsymbol{x}_{ij} < 4.5 \end{cases} \tag{7-4}$$

$$E_T = \begin{cases} 0.23\boldsymbol{x}_{ij} & \boldsymbol{x}_{ij} \le 3 \\ -0.47\boldsymbol{x}_{ij}/(\boldsymbol{x}_{ij} - 1) & \boldsymbol{x}_{ij} > 3 \end{cases} \tag{7-5}$$

Table 7-2: Description of explanatory variables in the GLMs

Explanatory variables	Coeff[a]	st.err[b]	t-value	E[c]
Constant	2.589	0.174	14.899	
Right shoulder width (ft)	-0.014	0.005	-2.936	-0.128
AADT (in thousands)	0.016	0.001	15.171	0.257
Rural area (Rural: 1; Urban: 0)	-1.727	0.109	-15.810	-4.626
Speed limit (mph)	-0.019	0.004	-4.840	-1.025
Truck percentage	-0.031	0.006	-5.563	-0.399
Model Evaluation				
Dispersion parameter for Quasi-Poisson GLM		61.586		
Null deviance on 20,931 degrees of freedom		317,792		
Residual deviance on 20,926 degrees of freedom		227,811		

[a]coefficients in the model; [b]standard error; [c]average elasticity value.

Table 7-3: Description of explanatory variables in the GNMs

Explanatory variables	Coeff[a]	st.err[b]	t-value	E[c]
Constant	1.932	0.167	11.602	
U_j for Right shoulder width	-0.413	0.082	-5.009	-0.511
U_j for AADT	0.592	0.033	17.984	1.390
U_j for Grade percentage	0.488	0.103	4.740	E_G
Rural area (Rural: 1; Urban: 0)	-1.025	0.114	-9.013	-1.786
Speed limit (mph)	-0.010	0.003	-2.953	-0.010
U_j for Truck percentage	0.261	0.037	7.098	E_T
Model Evaluation				
Dispersion parameter for Quasi-Poisson GNM	57.405			
Null deviance on 20,931 degrees of freedom	317,792			
Residual deviance on 20,925 degrees of freedom	218,113			

[a]coefficients in the model; [b]standard error; [c]average elasticity value.

Table 7-4: F-Test between the GLM and GNM

	Resid.Df	Resid.Dev	Df	Deviance	F-Value	Pr(>F)
GLM	20,926	227,811				
GNM	20,925	218,113	1	9,698	168.9	<2.20E-16

7.1.3 Model Interpretations and Discussions

The statistical analysis results are shown in Tables 7-2, 7-3, and 7-4. Classical statistical theory indicates that residual deviance has an asymptotic chi-squared distribution and significantly larger values of residual deviance indicate a model deficiency. The residual deviances are 218,113 (20,925 degrees of freedom) and 227,811 (20,926 degrees of freedom) indicating a good fit of Quasi-Poisson GNM outperforms its GLM counterpart. The mean-variance relationship for the Poisson GLM and GNM (mean/variance = 1) is not valid for rear-end crash data. The dispersion parameters for the GLM and GNM are 61.586 and 57.405 respectively, which are significantly different from one and verify the appropriateness of the Quasi-Poisson regression for both GLM and GNM. Apparently, the sample variance is larger than the sample mean for the rear-end crash data used in this study.

A total of five explanatory variables are identified significant, including Right Shoulder Width, AADT, Rural Area, Speed Limit and Truck Percentage, for both Quasi-Poisson GLM and GNM. One more variables, Grade is also identified significant in the Quasi-Poisson GNM. Due to more reasonable model specifications, the GNM includes more explicitly significant variables and performs better in extracting data patterns. The detailed model interpretations and discussions will be mostly based on the Quasi-Poisson GNM.

Three traffic flow-related variables are found significant for rear-end crash occurrence formulated in the GLM and GNM. The variable, Speed Limit, has a negative coefficient. Its estimation results indicate that the highway sections with higher speed limits are associated with lower rear-end crash rate, which may result from the facts that high speed limits are normally set up for relatively desirable roadway conditions and geometric characteristics and the less rear-end crashes are observed along such roadways. The coefficients of the variable, AADT, are 0.016 and 0.592 for the GLM and GNM, respectively, which illustrate increased traffic volumes may lead to more rear-end crashes. Furthermore, the AADT nonlinear predictor represented by Equation (7-1) in the GNM method further indicates that a higher value of AADT is associated with a lower increasing rate of crash occurrence. The Truck Percentage is found significant to have a negative effect on rear-end crash occurrence in the GLM method. Its nonlinear function (Equation (7-2)) and its elasticity (Equation (7-5) in the GNM further indicates that the highest rear-end crash rate is associated with the traffic composition with trucks: 3% and passenger cars: 97%. The probability of rear-end crash occurrence increases when truck percentage increased before it reaches the turning point of 3%. Then increased truck percentage will have a negative impact on rear-end crash occurrence. The GNM results could provide better explanations that the number of rear-end crashes tend to increase when more trucks are observed due to its inferior operation performance compared to passenger cars. However, when truck percentage increase to a certain level, driver tend to pay more attention and drive cautiously to lead to a reduction in crash rate.

Among the geometric characteristics and design elements, two variables, including Right

Shoulder Width, and Grade, are found significant for rear-end crash occurrence in the GNM. On the other hand, only one variable, Right Shoulder Width, is significant in the GLM. The GLM results show that the crash rates tend to decrease along the roadway sections with wider right shoulders. The GNM results indicate that rear-end crash rate significantly decrease when roadway sections are characterized with right shoulders compared to ones without right shoulders. However, no significant difference is observed after the right shoulders are bigger than certain value (in this research the critical right shoulder width is equal to 1 foot). The variable, Grade, is not found significant in the GLM but its significant attributes are identified by the GNM. Its nonlinear predictor (Equation (4-34)) and its elasticity (Equation (7-4)) in the GNM method indicates that rear-end crash rate reaches its two peak values when Grade increases to the certain thresholds (4.4% for the downhill and 4% for the uphill in this study). Its impacts on crash occurrence decrease after it has exceeded these critical values because drivers may pay more attention on abnormal geometric characteristics and roadway conditions. The nonlinear predictor also indicates its minor impacts on crash occurrence when the grade is small (from -3.5% to 3.5%) or extremely large (from -7% to -5.5% for downhill sections and from 4.5% to 7% for uphill sections).

In terms of area types, only one variable is found significant: Rural Areas in both the GLM and the GNM. The estimation results indicate that compared to urban highways, rural highways are associated with lower rear-end crash rate.

In summary, the GNM outperforms the GLM by illustrating more reasonable model

specifications and flexible result interpretations. More sensible and applicable explanations for the relationship between crash rate and its contribution factors are supported and justified by the GNM. The GNM-based approaches overcome the constraints associated with the previous GLMs in describing the monotonous relationship between significant contributing factors and crash occurrence. The findings are useful to identify significant contributing factors and develop more robust and applicable countermeasures against rear-end crash occurrence.

7.2 Generalized (non) Linear Models for Animal-Vehicle Collisions

7.2.1 Data Description

The method of GNMs can be implemented in modeling a wide variety of count data. To further elaborate the advantages of GNMs, this study also uses five years (2002-2006) of AVC data from ten highway routes (US 2, SR 8, US 12, SR 20, I-90, US 97, US 101, US 395, SR 525 and SR 970) following the recommendation from WSDOT experts. Three major data sources are included in this study:

- Combined AVC data from reported AVC data and carcass removal (CR) data (Lao et al., submitted for publication). Reported AVC data can be extracted from the Washington

State accident file provided by Highway Safety Information System (HSIS), which is operated by the University of North Carolina Highway Safety Research Center and the LENDIS corporation under a contract with Federal Highway Administration (FHWA) (HSIS, 2011). The HSIS collision data of Washington were compiled from the State Trooper filed field reports and citizen reports. Carcass removal data was provided by the maintenance team of Washington State Department of Transportation (WSDOT) stores. Both reported AVC data (Hubbard et al., 2000; Malo et al., 2004; Seiler, 2005) and CR data (Reilley and Green, 1974; Allen and McCullough, 1976; Knapp and Yi, 2004; Lao et al., 2011) have been commonly used in AVC research.

- Roadlog data received from the Highway Safety Information System (HSIS) provides geometric information for the roadway, such as terrain type, lane width and shoulder width.
- Deer distribution data provided by Washington Department of Fish Wildlife (WDFW) is in the form of GIS-based maps for mule deer, white-tailed deer, and elks.

Table 7-5 lists all explanatory variables used in the modeling process. Most of the quantitative and dummy variables were directly selected from the combined dataset. Several variables were created based on the observed data. For example, the variable "Truck Percentage Level" is a dummy variable and was created based on the percentage of trucks on the traffic flow. The variable is set to 1 when the percentage of trucks is greater than 5% and is set to 0 when equal or greater. The variable determination is set in this manner because a dramatic increase in AVCs was found when the percentage of trucks in traffic is less than 5%. Other examples, such as

variables $z14$, $z15$, and $z16$, were created for representing habitats of white-tailed deer, mule deer, and elk.

The minimum, maximum, mean, and standard deviation (S.D.) of each variable are shown in Table 7-5. The number of AVCs per segment is over-dispersed as indicated by the variance (4.23) being higher than the mean (1.14).

Table 7-5: Description of explanatory variables in the models

	Variable	Min	Max	Mean	S.D.
Y	Number of AVCs per segment [a]	0	104	1.14	4.23
z1	Annual average daily traffic (in thousands)	0.31	148.8	15.11	21.07
z2	Restrictive access control (Yes: 1; No: 0)			0.24	
z3	Speed level (>50mph: 1; otherwise: 0)			0.68	
z4	Truck percentage level (>5%: 1; otherwise: 0)			0.78	
z5	Median width (> 6 feet: 1; others: 0)			0.33	
z6	Total number of lanes (in both directions)	2	9	2.96	1.32
z7	Roadway length (mile)	0.01	6.99	0.22	0.4
z8	Terrain type (Rolling: 1; Otherwise: 0)			0.72	
z9	Terrain type (Mountainous: 1; Otherwise: 0)			0.095	
z10	Lane width (feet)	10	20	12.5	1.88
z11	Left shoulder width (feet)	0	18	2.44	2.04
z12	Right shoulder width (feet)	0	20	4.03	3.52
z13	Rural area (Rural: 1; Urban: 0)			0.76	
z14	White-tailed deer habitat (Yes: 1; No: 0)			0.31	
z15	Mule deer habitat (Yes: 1; No: 0)			0.51	
z16	Elk habitat (Yes: 1; No: 0)			0.31	
z17	Vertical curve (Grade percentage)	0	9.87	1.039	1.167

[a] Dependent variable, combination of both carcass data and reported AVC data; Min: Minimum; Max: Maximum; S.D.: standard deviation

7.2.2 Model Estimation

7.2.2.1 Estimation of Nonlinear Predictors

If the nonlinear predictor U_j are unknown in Equations (4-35) and (4-36), the correspondent weight parameter w_j for U_j will be unable to estimate. Thus, similar to the application on rear-end collision data, it is necessary to determine U_j based on some preliminary statistical analysis on the crash rate and its associated factors. That is the model estimation should follow two steps mentioned in 7.1.2.: 1) Model specification estimation of nonlinear predictors, and 2) Parameter calibration and estimation, which are detailed as follows.

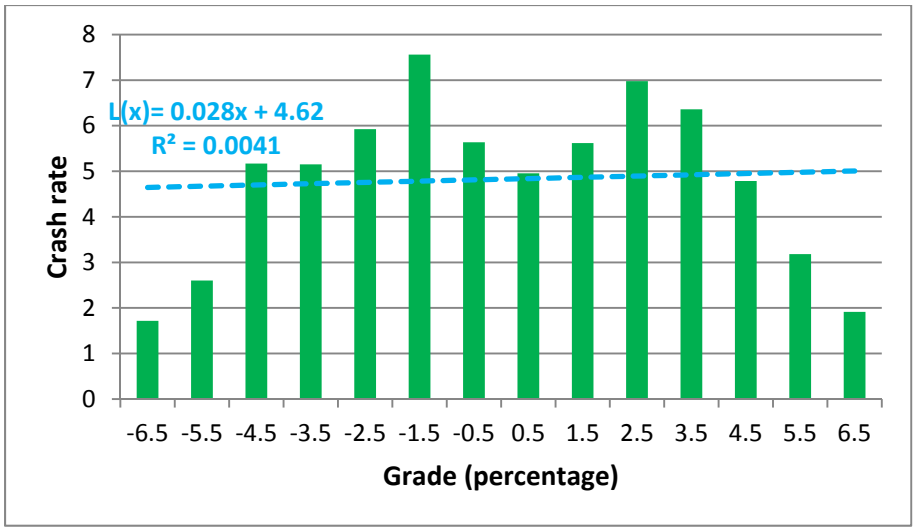
During the preliminary analysis process, if the relationship between the crash rate and the associated factor is found to be significantly different from a linear relationship, a new nonlinear contribution function U_j is needed to be determined; otherwise, $U_j(x_{ij})$ can keep the same value as x_{ij}.

Take the associated factor of grade as an example. Assume all other contributing factors are approximately equal across different grade levels. This assumption is reasonable when the sample data in different grade levels is large enough. The five year crash data from the ten routes has 11981 crash records and a total of 10466 segments. Thus, the assumption is acceptable. Figure 7-2 shows the bar chart of crash rate in five years (2002-2006) from ten highways in Washington State, by grade. Figure 7-2 (a) treats the downhill and uphill separately and Figure 7-

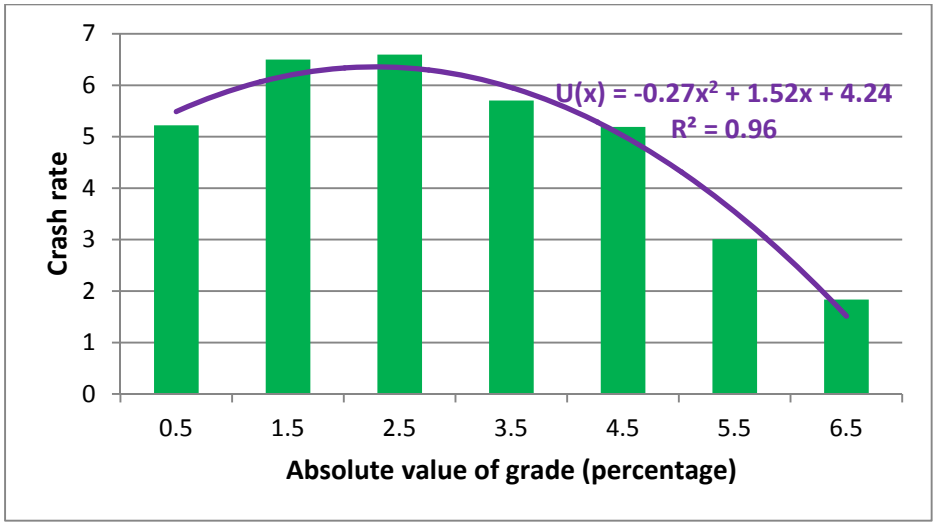
2 (b) combines the downhill and uphill together. From Figure 7-2 (a), for both downhill and uphill, one can find that the crash rate tends to increase with the increasing of the absolute value of grade at the beginning, after a certain point, i.e. around a grade of 2.5%, the crash rate tend to decrease with the increasing of the absolute value of grade. A linear line, like the dashed line $L(x)$ in Figure 7-2 (a), is not appropriate to be used for describing the relationship between the crash rate and grade because it merely describes the monotonic relationship.

The relationship between crash rate and AADT is another example. The bar chart in Figure 7-3 shows the Animal-Vehicle Crash (AVC) rate for five years (2002-2006) from ten highways (US2, SR8, US12, SR20, I90, US97, US101, US395, SR525 and SR970) in Washington State. The x-axis is the AADT (in thousand) and the y-axis is the average crash rate (number of crashes per mile).

Previous studies using GLM assume the relationship to be in a generalized linear form and show that the expected crash rate should increase with an increasing AADT (Chin and Quddus, 2003; Lao et al., 2011). This increasing linear relationship is shown as the dashed line $L(x)$ in Figure 7-3. The R^2 for the fitted line $L(x)$ is nearly equal to zero (R^2=4E-5). Accordingly, $L(x)$ could not reflect the true relationship between the expected crash rate and AADT, assuming all other contributing factors are approximately equal across AADT levels.



(a)Downhill is treated as minus grade



(b)Absolute value of grade

Figure 7-2. Animal-Vehicle crash rate (crash frequency per mile) in five years (2002-2006) from ten highways in Washington State, by grade

It is necessary to identify the nonlinear relationship between the expected crash rate and some of its associated factors. Assume that each associated factor x corresponds with a contribution value $U(x)$ related with crash data. Here, the correspondent value $U(x)$ is defined as a nonlinear predictor. For the same associated factor AADT, as shown in Figure 7-3, the solid curve $U(x)$ is a second order polynomial curve with an R^2=0.63. This curve $U(x)$ does a much better job than $L(x)$ for fitting the expected crash rate. Thus, the nonlinear predictor $U(x)$ can be used for modeling more flexible relationships between the expected crash rate and the contributing factors. For a specific explanatory variable, the $U(x)$ can be straight lines, any curves, or combination of both. In general, the determination of $U(x)$ is based on statistical analysis of the crash rate and the explanatory variable.

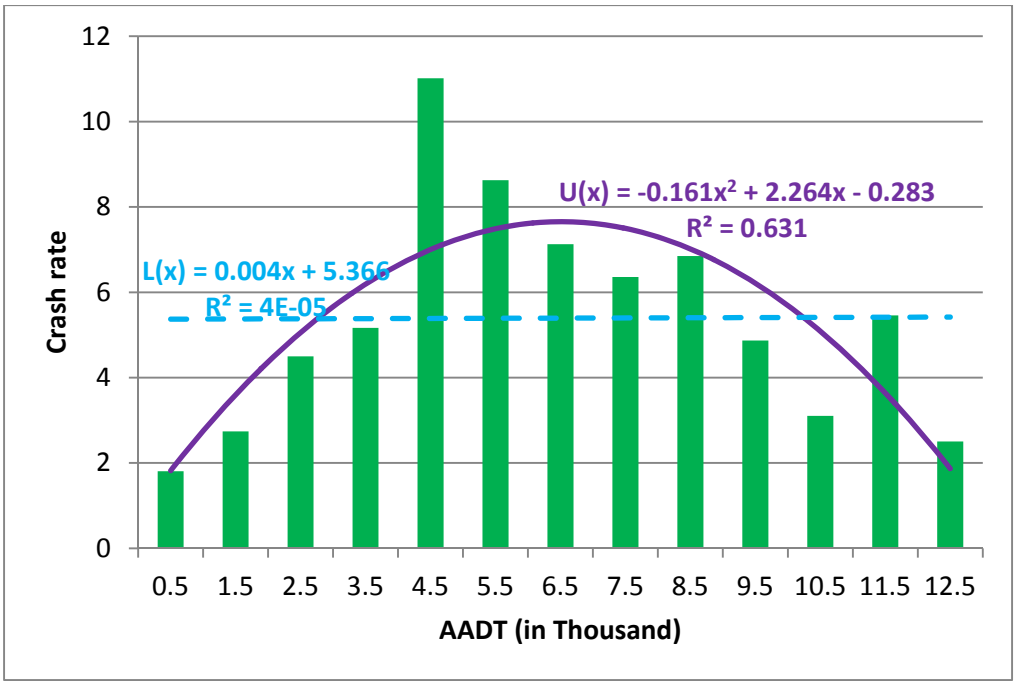


Figure 7-3. Animal-Vehicle crash rate (crash frequency per mile) over five years (2002-2006) from ten highways in Washington State, by AADT

Therefore, nonlinear predictors are necessary in order to better describe the relationship between crash rate and grade percentage or the AADT. To simplify the question for the grade, this research combines the downhill and uphill by only considering their absolute value. The reasons are: (1) as shown in Figure 7-2 (a), the trends of grade effect on crash rate for downhill and uphill are similar and (2) when applied to the case study, it doesn't make any improvement in the models after separating the downhill and uphill into different nonlinear predictors.

As shown in Figure 7-2 (b), the second order polynomial curve, $U(x) = -0.27x^2 + 1.52x + 4.24$, is identified as the nonlinear predictor for grade associated factor for the absolute value of grade percentage. With an R^2=0.96, this second order polynomial curve does a much better job than $L(x)$ for fitting the expected crash rate.

Similar approach is applied to other associated factors. Table 7-6 shows a summary of the nonlinear predictor estimation for grade and AADT. For other associated factors, the relationship between crash rate and these factors is not found to have any special pattern and be significantly different from a linear relationship. Therefore, their $U_j(x_{ij})$ are kept to be identical with x_{ij}.

Table 7-6: Nonlinear predictor estimation for grade and AADT

Variable (x)	Nonlinear predictor U(x)	R^2
Vertical curve (Grade percentage)	$-0.27x^2 + 1.52x + 4.24$	0.96
Annual average daily traffic (in thousands)	$-0.16x^2 + 2.26x - 0.28$	0.63

7.2.2.2 Parameter estimation

For the purpose of comparison, both Poisson regression model and a negative binomial regression model were used for estimating parameters in GLM (Equations (4-28) and (4-29)) and

GNM (Equations (4-35) and (4-36)).

The variables were assigned based on preliminary analysis and their significance in the models. The final model was selected based on the AIC value. Table 7-7 shows the coefficients of explanatory variables and statistical test results of the convergence GLM from Equation (4-29), estimated by negative binomial regression. Variables significantly associated with the crash rate are shown as the explanatory variables in the model. The calculated elasticity for each associated factor is listed in the last column.

Similarly, the coefficients of the explanatory variables and their significance are shown for the GNMs in Table 7-8. Nonlinear predictors for grade percentage and AADT are also significant in GNMs, whereas these two variables are insignificant in GLMs. Whether the elasticity values of the nonlinear predictors are positive or negative correlates with the value the correspondent variable, which will be discussed more in the following section.

Table 7-7: Description of explanatory variables in the GLMs

Explanatory variables	Coeff[a]	st.err[b]	t-value	E[c]
Variables affecting the probability of a hazardous crossing of an animal (P_o)				
Constant	-1.733	0.194	-8.943	
Speed level (>50mph: 1; otherwise: 0)	1.618	0.072	22.544	0.802
Truck percentage Level (>5%: 1; otherwise: 0)	-0.453	0.086	-5.234	-0.573
Total number of lanes	-0.253	0.044	-5.791	-0.756
Median width (> 6 feet: 1; others: 0)	-0.374	0.112	-3.337	-0.454
Terrain type (Rolling: 1; Otherwise: 0)	0.260	0.065	3.987	0.229
Rural area (Rural: 1; Urban: 0)	1.208	0.087	13.930	0.701
White-tailed deer habitat (Yes: 1; No: 0)	0.892	0.060	14.926	0.590
Model Evaluation				
AIC at based model[#]			58,168	
AIC at convergence with Poisson regression			45,989	
AIC at convergence with NB regression (δ = 6.105)			21,353	
ρ^2 for NB regression model			0.6329	

[a]coefficients in the model; [b]standard error; [c]average elasticity value; ρ^2 was calculated by comparing the log-likelihood with the base model; base model[#]: δ approaches zero and β=0.

Table 7-8: Description of explanatory variables in the GNMs

Explanatory variables	Coeff[a]	st.err[b]	t-value	E[c]
Variables affecting the probability of a hazardous crossing of an animal (P_o)				
Constant	-2.368	0.250	-9.465	
Speed level (>50mph: 1; otherwise: 0)	1.576	0.072	21.925	0.793
Truck percentage Level (>5%: 1; otherwise: 0)	-0.499	0.087	-5.725	-0.647
U_j for Annual average daily traffic	0.038	0.012	3.213	$0.037(2.26-0.32\boldsymbol{x}_{ij})$
Total number of lanes	-0.230	0.044	-5.211	-0.687
Median width (> 6 feet: 1; others: 0)	-0.296	0.112	-2.629	-0.344
U_j for Grade percentage	0.105	0.033	3.172	$0.100(1.52-0.55\boldsymbol{x}_{ij})$
Terrain type (Rolling: 1; Otherwise: 0)	0.252	0.066	3.837	0.223
Rural area (Rural: 1; Urban: 0)	1.160	0.089	13.016	0.687
White-tailed deer habitat (Yes: 1; No: 0)	0.895	0.060	14.976	0.591
Model Evaluation				
AIC at based model[#]			58,168	
AIC at convergence with Poisson regression			45,852	
AIC at convergence with NB regression (δ = 6.053)			21,338	
ρ^2 for NB regression model			0.6331	

[a]coefficients in the model; [b]standard error; [c]average elasticity value; ρ^2 was calculated by comparing the log-likelihood with the base model; base model[#]: δ approaches zero and β=0.

7.2.3 Model interpretations and discussions

Based on the AIC values in Table 7-7 or 7-8, the negative binomial regression outperformed the corresponding Poisson regression. Meanwhile, the estimate results show that the δ value is about 6 for NB regression in both GLM and GNM models, and their p value is 0.00. This verifies that δ is significantly greater than 0, and that the AVC data are over-dispersed. In this case, the model estimated with Poisson regression should not be used because it requires the mean and variance of the number of AVCs per segment data to be the same. Hence, a model estimated with the NB regression is a better choice.

Several variables are identified as significant, including median width, number of lanes, terrain type, rural area, white-tailed deer habitat, speed level and truck percentage level, in both GLM and GNM models with NB regression. Whereas two variables, grade and AADT, are insignificant in a GLM model with NB regression, the nonlinear predictors of grade and AADT, are found to be significant in the GNM model. Moreover, both the AIC value and adjusted ρ^2 for the GNMs outperformed those of GLMs. Therefore, GNMs include more associated factors and perform better in fitting the data than that of GLMs. The detailed explanations and discussions will be mostly based on GNM with NB regression.

Among the traffic elements, three variables are significantly associated with the occurrence of

AVCs in the GNM and only two variables are significant in the GLM. The estimation results of two significant variables, speed level and truck percentage level, from the GLM and the GNM are similar in this study. Another variable, AADT is not found to be significant in the GLM. However, the nonlinear predictor of AADT was found to have significant effect on the crash data in the GNM.

The estimated coefficients show that the variability of speed level is the most significant in affecting the occurrence of AVCs (Coef. = 1.576, t = 21.925, E = 0.793). This implies that when a highway segment had a speed limit greater than 50 mph, the probability of an AVC occurrence increased. A reason for this is that a vehicle running at a higher speed requires a longer stopping distance; hence the reaction time is shorter. This finding is consistent with many previous AVC related studies (e.g. Rolley and Lehman, 1992; Allen and McCullough, 1976; and Lao et al., 2011b). The elasticity value here indicates the changes from the highways of speed limit lower than 50 mph to the highways of speed limit higher than 50 mph will cause an incremental change of 0.793% to the AVC risk.

The truck percentage level was found to have a negative relationship with the occurrence of AVCs (Coef. = -0.499, t = -5.725, E = -0.647). This is presumably so because most truck drivers drive at relatively lower speeds, are professionally trained, and are well experienced. This result is consistent with the motor vehicle crash research (Milton and Mannering, 1998) in which the increase in the percentage of trucks may decrease the crash probabilities. The elasticity value here indicates a decrement change of 0.647% to the AVC risk is caused by the changes from the

areas with a lower truck percentage to the areas with a higher truck percentage.

The nonlinear predictor of AADT was also found to have significant effect on the crash data (Coef. = 0.038, t = 3.213, E = 0.037(2.26-0.32$\boldsymbol{x}_{ij}$)). Based on the formula of elasticity, the elasticity is positive when the AADT is smaller than 7000, whereas the elasticity is negative when the AADT is bigger than 7000. This indicates that the probability of a crash is higher when AADT is around 7000. A low crash rate with a really low AADT may because the chance of an animal conflicting with a vehicle is diminished in a low volume condition. The low crash rate with a really high AADT may be caused by low travel speed with a really high AADT; the animals may be driven away when the traffic is busy.

Among the geometric design elements, three variables are found to significantly contribute to the AVC frequency in the GNM and only two variables are significant in the GLM. The estimation results of two significant variables, number of lanes and median width, from the GLM and the GNM are similar in this study. Another variable, grade, is not found to be significant in the GLM. However, the nonlinear predictor of grade was found to have a significant effect on the crash data in the GNM.

The number of lanes a significant factor in having a decreasing effect on crash data (Coef. = -0.230, t = -5.211, E = -0.687). This is understandable because roadway sections are typically wider with more travel lanes, which might increase the crossing difficulty for animals and thus the crash rate is lower. The elasticity value here shows that a 1% increase in the number of lane

decreases the AVC risk by 0.687%.

A median width of greater than 6 feet was found to have a significant negative effect on the AVC frequency (Coef. = -0.296, t = -2.629, E = -0.344). Similar to the number of lanes, a wider median will increase the crossing hesitation for animals, and hence reduce the probability of AVCs. The elasticity value here shows a decrement change of 0.344% on the AVC accident risk caused by the changes from median width less than 6 feet to median width more than 6 feet.

The nonlinear predictor of grade was also found to have a significant effect on the crash data (Coef. = 0.105, t = 3.172, E =0.100(1.52-0.55$\boldsymbol{x}_{ij}$)). Based on the formula of elasticity, the elasticity is positive when the grade percentage is smaller than 2.76%, whereas the elasticity is negative after the grade percentage is bigger than 2.76%. This indicates that the probability of AVC frequency is highest when grade percentage is 2.76%. Crash rates tend to increase initially when the grade percentage increases because the vehicle is harder to control in a steeper roadway. However, if the grade becomes steeper, the crash rates tend to decrease due to travel speeds lowering as well as increased driver awareness.

In terms of area types, two variables have significant impacts on the occurrences of AVCs: rolling areas, and rural areas. The estimation results of these two variables are similar in both the GLM and the GNM. In comparison with level terrain, rolling areas are associated with higher numbers of AVCs (Coef. = 0.252, t = 3.837, E = 0.223). This may be due to the higher animal population and activity levels in rolling terrain areas. The elasticity value here shows that an

incremental change of 0.223% to the AVC risk is caused by the changes from level terrain to rolling terrain.

Similarly, compared to highways in urban areas, those in rural areas are found to have a higher number of AVCs (Coef. = 1.160, t = 13.016, E = 0.687). This may also be because rural areas tend to have a higher animal population than urban areas. The elasticity value here shows that an incremental change of 0.687% to the AVC risk is caused by the changes inflicted by going from an urban area to a rural area.

In terms of high density animal distribution areas, white-tailed deer habitat is associated with a higher number of AVCs (Coef. = 0.895, t = 14.976, E =0.591). This may because a higher animal population in the white-tailed deer habitat will increase the chance of animal crossing the highway and thus increase the probability of AVC. Compared with white-tailed deer habitats, elk habitats are insignificant at a 95% significance level. This can be explained by the fact that the total number of collisions with elk only contributes a small portion of the whole AVC records in the study period. Mule deer habitats also were insignificant in the model. The reason for this may be the mule deer habitat distribution is relatively uniform in Washington State and covers a large portion in the study route area. The elasticity value for the white-tailed deer habitat indicates an incremental change of 0.780% on the AVC risk caused by the changes from other areas to white-tailed deer habitat areas. This finding is consistent with another AVC study (Lao et al., 2011a). The estimation results of white-tailed deer habitat from the GLM and the GNM are similar in this study.

In summary, among all the significant variables, speed level, rolling terrain type, rural area type, and white-tailed deer habitat have positive effects on AVC risk. Three variables, truck percentage level, total number of lanes, and median width, tend to reduce the probability of AVC risk when the values of these variables increase. Two other variables, nonlinear predictors for AADT and grade, are also found to significantly contribute to the occurrence of AVCs in the GNMs. With the increase of AADT, the crash probability tends to initially increase and then decrease after the AADT surpasses a certain value. Similarly, with the increase of grade, the probability of crash tends to increase initially, yet decrease once the grade reaches a certain value. The findings of AADT and grade from GNMs are useful because the original models, like GLMs, can only describe the monotonous relationship and thus the effect from AADT and grade on the crash rate could not be accurate described in this case. After getting a better relationship between the crash rate and associated factors, results from this model can be used to compile countermeasures against AVCs.

7.3 Model Validation and Transferability Test

In order to further compare the performance between the GLMs and GNMs. Model validation and transferability tests are conducted by using rear-end collision data. The GNM and GLM performance is examined and compared based on the validation dataset as well. As shown in Table 7-9, two test scenarios are designed: Scenario 1: The entire data sets are separated into to

sub-datasets: 70% of data are used to estimate and calibrate the models (the training dataset), and 30% are used to validate the models (the test dataset); Scenario 2: 80% of data are used as the training dataset, and 30% are used as the test dataset. Safety data from each route are separated proportionally into the training and test datasets in order to minimize the data heterogeneity between two datasets.

Table 7-9 Comparisons of the GNM and GLM performance

	Scenario 1 (70%/30%)		Scenario 2 (80%/20%)	
Total Validation Data[1]	6279		4186	
Models	GLM	GNM	GLM	GNM
MSPE[2]	0.564	0.336	0.736	0.419
Accurately Estimated Segments[3]	5647	5798	3630	3773
AES Percent	89.9%	92.3%	86.7%	90.1%

[1]Total Validation Data are the total number of the roadway segments used for the validation; [2]Mean Squared Prediction Error (Residual); [3]Accurately Estimated Segments are counted when the predicted and observed accidents are within 10% or less than 1.

As we can see from Table 7-9, the Mean Squared Prediction Error (MSPE) of GNMs is much smaller than that of GLMs for Scenario 1 (0.336 vs. 0.564) and Scenario 2 (0.419 vs. 0.736). The GNMs produce more accurate prediction results illustrated by the number of accurately estimated segments and their corresponding percentages. These results show that GNMs outperform GLMs based on the validation results under Scenarios 1 and 2.

Additionally, in order to test the model generality and transferability, the data is separated into two sets: Dataset A with five highway routes (SR 8, US 12, SR 20, US 97, and US 101) and Dataset B with the other five highway routes (US 2, I-90, US 395, SR 525 and SR 970). The model transferability issue is examined based on the model residual deviance ratio calculated in Equation 7-6.

$$R = (RD_A + RD_B)/ RD_{All} \tag{7-6}$$

where RD_A, RD_B, RD_{All} are residual deviance for Dataset A, Dataset B and all dataset respectively. If the model obtained from one dataset is transferable to apply for the other dataset, the sum of the residual deviance from Dataset A and Dataset B should be equal to the residual deviance from the total datasets. That is R should ideally be equal to 1. Table 7-10 shows the residual deviance results for both the GLM and GNM. Based on the results, we can see that the residual deviance ratios of the GNM and GLM are 99.03% and 96.66%, respectively, which indicate that the GNM performances a little better than that of the GLM in term of the location transferability.

Table 7-10: Residual deviance results between GLM and GNM

	GNM	GLM
All data	54,593	59,499
Dataset A	18,280	19,421
Dataset B	35,786	38,090
Residual deviance ratio	99.03%	96.66%

7.4 Summary

To improve highway safety, many studies have been conducted to identify the relationship between contributing factors and crash occurrences by using various modeling techniques. Most of these studies are GLM-based approaches, which assume that the crash rate and the linear combination of its contributing factors can be formulated through a link function. However, this assumption is violated in many situations due to the limitations of the linear combination. This research focused on GNM-based approach development enabling a nonlinear, including non-monotonic, relationship linking the crash rate and its associated factors. In the GNM-based approach, nonlinear predictors are used to replace the GLM linear predictors. If the nonlinear predictors are identical with the corresponding variables, the GNMs will degrade to GLMs.

The GNMs can be used to model more complicated relationships between any count data and its associated factors. In this research, the GNMs were applied to two types of datasets: rear-end crash data and AVC data. For both datasets, 5-year (2002-2006) crash data in Washington State was used. In the application of rear-end crash data, one variable, Grade, was found to be insignificant in GLMs, but significant in GNMs. Results from the GNMs indicates that crash risks tend to increase when the Grade value increases initially, then after a certain point, the crash risks tend to decrease while Grade continues increasing. This non-monotonic relationship is also found between crash rate and truck percentage by GNMs. This non-monotonic relationship between the crash rate and its associated factors could not be described in the original GLMs. Similarly, in the application of AVC, two variables, Grade and AADT, were found to be insignificant in GLMs, but significant in GNMs. These two variables all have a parabolic impact on the crashes, which could not be captured by the GLMs. This indicates that GNMs are more powerful in modeling the relationship between crash rates and their associated factors. Moreover, the calculation of the elasticity in the GNMs also can reveal that the non-monotonic relationship depends on the predictor format and the values of the variables. The elasticity from GNMs can provide more accurate information for transportation agencies to improve traffic safety.

Chapter 8. Conclusions

To improve highway safety, many studies have been conducted to identify the relationship between collisions and their contributing factors by using various modeling techniques. The most commonly used crash data models include Poisson regression, Gamma regression, and Negative Binomial (NB) regression. Although these count data models have been developed for decades, some major issues are still needed to be further investigated. Five key modeling issues are identified in this research: data dispersion, excess zeros, inconsistent observations, interaction behavior, and nonlinear relationships. The first two issues, data with dispersion and excess zeros, have been well studies in the previous research. However, the latter three still need expansion on the current research.

The issues of inconsistent observations brings to light that the crash data collected from different data sources do not match with each other. The interaction issue is that the characteristics of the collisions with interaction behavior, such as Animal-Vehicle Collision (AVC), are different from these of single vehicle crashes. The nonlinear relationship issue is that the relationships between the crash rate and its contributing factors will not be linear or monotonic.

To address the issues of inconsistent observations, this research puts forward two methods. In the first method, a fuzzy logic-based data mapping algorithm is proposed to match the data from two datasets. The membership functions of the fuzzy logic algorithm are estimated based on a survey from the Washington State Department of Transportation carcass removal staff, who have been

gaining experience in the field for many years. By applying this algorithm to 5-year (2002-2006) AVC datasets in Washington State, the combined dataset increased the number of records 15%-22% when compared to the original CR dataset. The proposed algorithm was verified by the expert judgment data on the surveyed AVC data pairs collected through another survey. The verification results showed that the accuracy of the proposed algorithm is approximately 90% for the limited pairs of data included in the survey. The fuzzy mapping algorithm proved appropriate to increase the quality and quantity of the AVC data and benefit wildlife safety studies and countermeasure identifications. Due to the fact that the design of the membership functions is adaptive in nature, the fuzzy logic based mapping algorithm introduced in this research can also be easily transferred for applications in other areas.

In the second method, a diagonal inflated bivariate Poisson regression (DIBP) method is adopted to fit two datasets together. The proposed model technique was applied to the reported AVC and carcass removal data sets collected in Washington State during 2002-2006. As an inflated version of the bivariate Poisson regression model, the diagonal inflated bivariate Poisson model outperformed other models (double Poisson, bivariate Poisson, diagonal inflated bivariate Poisson, and zero-inflated double Poisson) studied in this research. It was determined the DIBP models are the best fitted models with the lowest AIC and BIC values. The DIBP method demonstrates its capability of fitting two data sets with remarkable overlapping portions resulting from the same stochastic process. Therefore, the DIBP model provides researchers a new approach to investigating AVCs from a different perspective involving the three distribution parameters (λ_1, λ_2 and λ_3). Functionally, the DIBP model not only can handle under- or over-

dispersed count data but also can model paired data sets with correlation.

To address the interaction issue, a new probability model which explicitly formulates the interactions between the objects is presented. The proposed method was applied to the AVC data and this method can explicitly formulate the interactions between animals and drivers. This method can better capture the relationship among drivers' and animals' attributes, roadway and environmental factors, and AVCs. The proposed method includes a term (the probability of an animals' response failure to escape) to capture animals' reaction characteristics in AVCs. This method can be further developed to model other types of collisions with interaction behavior, such as Pedestrian-Vehicle Collisions, Vehicle-Vehicle Collisions, and Bicycle-Vehicle Collisions.

To address the nonlinear relationship issue, a Generalized Nonlinear Model (GNM)-based approach for modeling crash data is formulated using Washington State traffic safety data. Most of previous models focused on contributing factor identification and crash risk modeling are GLM-based approaches, which assume that the crash rate and the linear combination of contributing factors can be formulated through a link function. However, this assumption is violated in many situations due to the requirement of linear combination. This research focused on GNM-based approach development enabling a nonlinear, including non-monotonic, relationship linking crash rates and their associated factors. In the GNM-based approach, nonlinear predictors are used to replace the GLM linear predictors. If the nonlinear predictors and the corresponding variable are identical, the GNMs will degrade to GLMs. This indicates

that GNMs are more powerful in modeling the relationship between crash rates and their associated factors. Moreover, the calculation of the elasticity in the GNMs also can reveal that the non-monotonic relationship depends on the predictor format and the values of the variables. The elasticity from GNMs can provide more accurate information for transportation agencies to improve traffic safety.

The GNMs can be used to model more complicated relationships between any count data and its associated factors. In this research, the GNMs were applied to two types of datasets: rear-end crash data and AVC data. For both datasets, 5-year (2002-2006) crash data in Washington State was used. In the application of rear-end crash data, one variable, Grade, was found insignificant in GLMs, but significant in GNMs. Results from the GNMs indicates that the crash risks and grade have the non-monotonic relationship. The non-monotonic relationship is also found between the crash rate and truck percentage by GNMs. This non-monotonic relationship between the crash rate and its contributing factors could not be described in the original GLMs. When applied to the AVC data, two variables, Grade and AADT, were found insignificant in GLMs, but significant in GNMs. These two variables all have parabolic impact on the crashes, which could not be captured by the GLMs. The application of these two types of datasets indicates that GNMs are more powerful in modeling the relationship between crash rates and their associated factors. Moreover, the calculation of the elasticity in the GNMs also can reveal that the non-monotonic relationship depends on the predictor format and the values of the variables. The elasticity calculations from GNMs can provide more accurate information for transportation agencies to improve traffic safety. As a new modeling approach, the GNM may better extract the non-linear

relationships among variables and provide new perspectives of interpreting crash causal factors in a more complete and comprehensive manner.

These proposed solutions of the three major issues in crash modeling are critical for collision studies. The fuzzy-logic based data mapping algorithm can combine partial observations from different processes and achieve a more complete dataset for a thorough analysis. The DIBP models can directly take two data observation processes into account and provide an explanation of different contributing factors within both datasets. The occurrence mechanism based probability models and GNM based models are effective methods for handling interaction behavior and non-linear relationships between dependent and independent variables. Although the search results indicate that the proposed methods have their advantages of modeling crash data with rear-end data and AVC data when considering the inconsistent observations, interaction behavior and nonlinear relationships, further tests with other types of crashes, such as Pedestrian-Vehicle Collisions and Bicycle-Vehicle Collisions, are still desired.

References

1. Abdel-Aty M. Haleem K., 2011 Analyzing Angle Crashes at Unsignalized Intersections Using Machine Learning Techniques, Accid. Anal. Prev., 43(1), pp. 461-470.
2. Abdel-Aty, M. A., Radwan, A. E., 2000. "Modeling traffic accident occurrence and involvement." Accid. Anal. Prev. 32(5): 633-642.
3. Abdel-Aty, M., Abdelwahad, H., 2004. Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure. Accid. Anal. Prev., 36 (3), pp. 447–456.
4. Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models, Transp Res Rec 2061, 55–63.
5. Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. Accid. Anal. Prev. 43(1), 429-438.
6. Akaike, H., 1974. "A new look at the statistical model identification". IEEE Trans Automat Contr 19 (6): 716–723.
7. Allen, R.E., McCullough. D.R., 1976. Deer-Car Accidents in Southern Michigan. Journal of Wildlife Management, Volume 40, pp. 317–325.
8. Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159.
9. Anderson, R.W.G. A.J. McLean, M.J.B. Farmer, B.H. Lee, C.G. Brooks. 1997. Vehicle travel

speeds and the incidence of fatal pedestrian crashes. Accident Analysis and Prevention, 29 (5), pp. 667–674.

10. Ben-Akiva, M. E., Lerman, S. R. 1985. Discrete choice analysis: Theory and application to travel demand. MIT Press series in transportation studies, 9. Cambridge, Mass: MIT Press, 167-168.
11. Borman, S. 2009. The Expectation Maximization Algorithm - A short tutorial, http://www.seanborman.com/publications/EM_algorithm.pdf.
12. Cameron, A., Trivedi, P., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK.
13. Chin, H.C., Quddus, M.A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Anal. Prev. 35, 253–259.
14. Cohen, A. C., 1963. Estimation in Mixtures of Discrete Distributions. in Proceedings of the International Symposium on Discrete Distributions, Montreal, New York: Pergamon Press, 373-378.
15. Curtis, P.D., Hedlund J.H., 2005. Reducing Deer-Vehicle Crashes. Wildlife Damage Management Fact Sheet Series. Cornell Cooperative Extension, Ithaca, N.Y.
16. Daniels, S., Brijs, T., Nuyts, E., Wets, G., 2010. Explaining variation in safety performance of roundabouts. Accid. Anal. Prev. 42(2), 292-402.
17. Danielson, B.J., Hubbard. M.W., 1998. A Literature Review for Assessing the Status of Current Methods of Reducing Deer-Vehicle Collisions. A report prepared for The Task Force on Animal Collisions, Iowa Department of Transportation and Iowa Department of Natural

Resources.

18. Davis G.A., 2001. Relating severity of pedestrian injury to impact speed in vehicle-pedestrian crashes: simple threshold model. Transportation Research Record, 1773, pp. 108–113.
19. Dempster, A, Laird, N, Rubin, D., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. R. Stat. Soc., Ser. B, Stat. Methodol. 39, 1–38.
20. Donaldson, B.M., 2007. Use of Highway Underpasses by Large Mammals and Other Wildlife in Virginia: Factors Influencing Their Effectiveness. Transp Res Rec 2011, 157–164.
21. Donnell, E.T., Mason, J.M., 2006. Predicting the frequency of median barrier crashes on Pennsylvania interstate highways. Accid. Anal. Prev. 38 (3), 590–599.
22. El-Basyouny,, K., Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models. Transp Res Rec 1950, 9-16.
23. El-Basyouny,, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accid. Anal. Prev. 41 (5), 1118-1123.
24. El-Basyouny, K. and Sayed, T., 2010. Application of generalized link functions in developing accident prediction models. Safety Science. 48, 410–416.
25. El-Basyouny,, K., Sayed, T., 2011. A full Bayes multivariate intervention model with random parameters among matched pairs for before–after safety evaluation. Accid. Anal. Prev. 43 (1), 87-94.
26. Elvik, R., 2011, Assessing Causality in Multivariate Accident Models, Accident Analysis and Prevention, Vol. 43, pp. 253-264.
27. Erke, A., 2009. Red light for red-light cameras?. A meta-analysis of the effects of red-light

cameras on crashes. Accid. Anal. Prev. 41 (5), pp. 897-905.

28. Fox H., 1996. Crashes at Signal Controlled Junctions and Pelican Crossings in Glasgow Scottish Office Central Research Unit, Glasgow.
29. Garber, N.J., Wu, L., 2001. Stochastic Models Relating Crash Probabilities with Geometric and Corresponding Traffic Characteristics Data (UVACTS-5-15-74). Charlottesville, VA: Center for Transportation Studies, University of Virginia. HSIS., 2009. http://www.hsisinfo.org/. Accessed June 15, 2009
30. Garber, N.J., Miller, J.S., Abel, R.E., Eslambolchi, S., Korukonda, S.K., 2007. The impact of red light cameras (photo-red enforcement) in Virginia. Final report VTRC 07-R2. Virginia Transportation Research Council, Charlottesville, VA.
31. Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The Negative-Binomial-Lindley Generalized Linear Model: Characteristics and Application using Crash Data. Accident Analysis and Prevention, 45(2), pp. 258-265.
32. Gkritza, K., Baird, M., Hans. Z. N., 2010. Deer-vehicle collisions, deer density, and land use in Iowa's urban deer herd management zones. Accid. Anal. Prev. 42(6), pp. 1916-1925.
33. Gonzales-Barron, U., and Butler, F., 2011. A comparison between the discrete Poisson-gamma and Poisson-lognormal distributions to characterise microbial counts in foods. Food Control, 22 (8), pp. 1279–1286.
34. Harb, R., Radwan, E., Yan, X., Pande, A. Abdel-Aty M., 2008. Freeway work-zone crash analysis and risk identification using multiple and conditional logistic regression. ASCE J. Transp. Eng., 34 (5), pp. 203–214.
35. Hauer, E., 2004. Statistical Road Safety Modeling, Transportation Research Record: Journal

of the Transportation Research Board, No. 1897, Transportation Research Board of the National Academies, Washington, DC.

36. HSIS. 2012. http://www.hsisinfo.org/. Accessed April 10, 2012
37. Huijser, M. P., P. McGowen, J. Fuller, A. Hardy, A. Kociolek, A. P. Clevenger, D. Smith, and R. Ament, 2007a. Wildlife-Vehicle Collision Reduction Study. Report to Congress. Federal Highway Administration, U.S. Department of Transportation, Washington D.C.
38. Huijser, M.P., Fuller J., Wagner M.E., A. Hardy, and A.P. Clevenger., 2007b. Animal-vehicle Collision Data Collection: A Synthesis of Highway Practice. National Cooperative Highway Research Board Program: Synthesis 370. Transportation Research Board, Washington, DC.
39. Huijser, M.P., and P.J.M. Bergers, 2000. The Effect of Roads and Traffic on Hedgehog (Erinaceus europaeus) Populations. Biological Conservation, Vol. 95, pp. 111–116.
40. Hubbard, M.W. B.J., Danielson, Schmitz, R.A., 2000. Factors influencing the location of deer-vehicle accidents in Iowa. J Wildl Manage 64, 707–712
41. Johnson, N. L., Kotz, S., 1969. Distributions in Statistics: Discrete Distributions, Boston: Houghton Muffin.
42. Institute for Statistics and Mathematics. 2011. http://www.r-project.org/. Accessed June 28, 2011.
43. Jovanis, P. P., Chang, H., 1986. Modeling the Relationship of Accidents to Miles Traveled, Transp Res Rec 1068, 42-51
44. Kao, J. H. K., 1960. A summary of some new techniques in failure analysis, Proceedings of Sixth National Symposium on Reliability and Quality Control, Washington DC, 190-201.
45. Kadane, J.B., Shmueli, G., Minka, T.P., Borle, S., Boatwright, P., 2006. Conjugate analysis of

the Conway–Maxwell–Poisson distribution. Bayesian Anal 1, 363–374.

46. Karim, E., Tarek, S., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accid. Anal. Prev. 41 (4), 820-828.
47. Karlis, D., 2003. An EM algorithm for multivariate Poisson distribution and related models. J Appl Stat 30(1), 63-77.
48. Karlis D, Ntzoufras I (2003). "Analysis of Sports Data by Using Bivariate Poisson Models." Journal of the Royal Statistical Society D (The Statistician), 52, 381 – 393.
49. Karlis, D., Ntzoufras, I., 2005. Bivariate Poisson and diagonal inflated Poisson regression models in R. J Stat Softw 14(10).
50. Kim, J., Wang, Y., Ulfarsson, G. 2007. Modeling the Probability of Freeway Rear-End Crash Occurrence. ASCE J. Transp. Eng., 133 (1), 11-19.
51. Kim, J.-K. G.F. Ulfarsson, V.N. Shankar, F. L., 2010. Mannering. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accident Analysis & Prevention. 42(6), Pages 1751–1758.
52. Kim, J.-K. G.F. Ulfarsson, V.N. Shankar, S. Kim. 2008. Age and pedestrian injury severity in motor-vehicle crashes: a heteroskedastic logit analysis. Accident Analysis and Prevention, 40 (5), pp. 1695–1702.
53. Knapp, K.K., Lyon, C., Witte, A., Kienert, C., 2007. Crash or Carcass Data: A Critical Definition and Evaluation Choice. Transp Res Rec 2019, 189-196.
54. Knapp, K.K., Yi, X., 2004. Deer-Vehicle Crash Patterns and Proposed Warning Sign Installation Guidelines. In: Proceedings of the TRB 2004 Annual Meeting, TRB 04-000571.
55. Kocherlakota, S., Kocherlakota, K., 1992. Bivariate Discrete Distributions. New York:

Marcel Dekker.

56. Kumara, S.S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Inj. Prev. 3 (4), 53–57.
57. Lao,Y., Wu, Y., Corey, J., Wang, Y.,. 2011a. Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression. Accid. Anal. Prev., 43(1), 220-227.
58. Lao Y., Zhang G., Wu Y., Wang Y., 2011b. Modeling animal–vehicle collisions considering animal–vehicle interactions. Accid. Anal. Prev., 43(6): 1991-1998.
59. Lao, Y., Wu, Y., Wang, Y., McAllister, K. 2012a. Fuzzy Logic-based Mapping Algorithm for Improving Animal-Vehicle Collision Data. Journal of Transportation Engineering. 138(5). Pp. 520-526.
60. Lao Y, Zhang G, Wang Y., 2012b. Modeling Vehicle Crashes Using Generalized Nonlinear Model, Submitted to Accident Analysis and Prevention, April. 2012.
61. Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accid. Anal. Prev. 34 (2), 149–161.
62. Li, C.C., Lu, J.C., Park, J., Kim, K., Brinkley, P.A., Peterson, J.P., 1999. Multivariate zeroinflated Poisson models and their applications. Technometrics, 41(1), 29-38.
63. Liddle, A. R., 2007. Information criteria for astrophysical model selection. Monthly Notices of the Royal Astronomical Society: Letters 377:1, L74-L78.
64. Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37(1), 35-46.
65. 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. Accid. Anal. Prev. 38 (4), 751–766.

66. Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. Accid. Anal. Prev. 40 (3), 1123–1134.
67. Lord, D., Manar, A., Vizioli, A., 2005. Modeling crash-flow-density and crashflow- V/C ratio for rural and urban freeway segments. Accid. Anal. Prev. 37 (1), 185–199.
68. Lord, D., Geedipally, S.R., 2011. The negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. Accident Analysis and Prevention 43 (5), 1738–1742.
69. Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Accid. Anal. Prev. 44 (5), 291-305.
70. Lord, D., Miranda-Moreno, L.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. Saf Sci 46 (5), 751–770.
71. Ma, J., Kockelman, K., 2006. Bayesian multivariate Poisson regression for models of injury count, by severity. Transp Res Rec 1950, 24–34.
72. Maher, M.J., Summersgill I.A., 1996. Comprehensive methodology for the fitting of predictive accident models. Accident Analysis & Prevention, 28(3) 281-296.
73. Malo, J.E., Suarez, F., Diez, A., 2004. Can we mitigate animal–vehicle accidents using predictive models?. J. Appl Ecol 41. 701–710.
74. Malyshkina, N., Mannering, F., 2010. Empirical assessment of the impact of highway design

exceptions on the frequency and severity of vehicle accidents, Accid. Anal. Prev. 42 (1), 131-139.

75. Matsui Y., 2005. Effects of vehicle bumper height and impact velocity on type of lower extremity injury in vehicle-pedestrian accidents. Accid. Anal. Prev., 37 (4), pp. 633–640.
76. Maycock, G., Hall, R. D., 1984. Accidents at 4-arm roundabouts. Laboratory Report LR1120. Crowthorne, Berks, U.K.: Transport Research Laboratory.
77. McCullagh, P. and J.A. Nelder. 1989. Generalized Linear Models. 2nd ed. Chapman & Hall/CRC, Boca Raton, Florida.
78. Meng, Q., Weng, J., 2011. Evaluation of rear-end crash risk at work zone using work zone traffic data. Accid. Anal. Prev., 43 (4), pp. 1291-1300.
79. Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accid. Anal. Prev. 26 (4), 471–482.(1), 131–139.
80. Miaou, S.-P., Bligh, R.P., Lord, D., 2005 Developing median barrier installation guidelines: a benefit/cost analysis using Texas data. Transp Res Rec 1904, 3–19.
81. Miaou, S., Hu, P.S., Wright, T., Rathi, A.K., and Davis, S.C. 1992. Relationship between truck accidents and highway geometric design: a Poisson regression approach. Transp Res Rec 1376, 10-18.
82. Miaou, S.P., Lord, D. 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and bayes versus empirical bayes. Transp Res Rec 1840, 31-40
83. Miaou, S.P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design

relationships. Accid. Anal. Prev. 25 (6), 689–709.

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

85. Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accid. Anal. Prev. 37 (4), 699–720.

86. National Transportation Safety Board, 2001. Special Investigation Report—Vehicle- and Infrastructure-Based Technology for the Prevention of Rear-End Collisions, NTSB/SIR-01/01, PB2001-917003.

87. Nelder, J., Wedderburn, R. 1972. "Generalized Linear Models". Journal of the Royal Statistical Society. Series A (General) 135 (3), 370–384

88. NHTSA, 2011. Traffic Safety Facts 2009. U.S. Department of Transportation, DOT HS 811 392.

89. NHTSA, 2012. Traffic Safety Facts 2010. Data. DOT HS 811 630.

90. Noland, R.B., Quddus, M.A., 2004. A spatially disaggregate analysis of road casualties in England. Accid. Anal. Prev. 36(6), 973-984.

91. Oh, C., Kim, T., 2010. Estimation of rear-end crash potential using vehicle trajectory data. Accident Analysis and Prevention, 42 (6), pp. 1888-1893.

92. Oh, J.,Washington, S.P., Nam, D., 2006. Accident prediction model for railway highway interfaces. Accid. Anal. Prev. 38 (2), 346–356.

93. Öström, M. A. Eriksson. Pedestrian fatalities and alcohol. Accident Analysis and Prevention, 33 (2) (2001), pp. 173–180.

94. Park, E.-S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transp Res Rec, 2019, 1–6.

95. Pitman J. 1993. Probability, Springer-Verlag New York, Inc.

96. Plait, A. 1962. The Weibull Distribution – with the tables, Industrial Quality Control, No. 19, 19-26.

97. Poch, M., Mannering F. 1996. Negative binomial analysis of intersection accident frequencies, J.Transp. Eng., 122(2), 105-113.

98. Proctor, M.F., 2003. Genetic Analysis of Movement, Dispersal, and Population Fragmentation of Grizzly Bears in Southwestern Canada, Ph.D. dissertation, The University of Calgary, Calgary, AB, Canada.

99. Qin, X., Ivan, J.N., Ravishankar, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accid. Anal. Prev. 36 (2), 183–191.

100. Reilley, R.E. and H.E. Green. 1974. Deer Mortality on a Michigan Interstate Highway. Journal of Wildlife Management, Volume 38, pp. 16–19.

101. Retting, R.A. Ferguson, S.A. Hakkert, A.S. 2003. Effects of red light cameras on violations and crashes: a review of the international literature. Traffic Injury Prevention, 4 (2003), pp. 17–23.

102. Rodriguez, D.A., Rocha, M., Khattak, A.J., and Belzer, M.H. 2003. Effects of truck driver wages and working conditions on highway safety: Case study. Transp Res Rec 1833, 95-102.

103. Rolley, R. E., Lehman, L. E., 1992. Relationships Among Raccoon Road- Kill Surveys, Harvests, and Traffic. Wildl Soc Bull 20, 313–318.

104. Romin, L. A., and Bissonette J. A., 1996. Deer–Vehicle Collisions: Status of State

Monitoring Activities and Mitigation Efforts. Wildlife Society Bulletin, Volume. 24, No. 2, pp. 276–283.

105. Schrank, David, and Timothy Lomax. 2009. The 2009 Urban Mobility Report. Texas A&M University, Texas Transportation Institute.
106. Schwarz, G., 1978. "Estimating the dimension of a model". Ann Stat 6 (2): 461–464.
107. Seiler A., 2005. Predicting locations of moose–vehicle collisions in Sweden. J Appl Ecol 42, 371–382.
108. Shankar, V., Mannering, F., Barfield, W. 1995. Effect of roadway geometric and environment factors on rural freeway accident frequencies. Accid. Anal. Prev. 27(3), 371–389.
109. Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accid. Anal. Prev. 29 (6), 829–837.
110. Shankar V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Saf Sci 41(7), 557-640.
111. Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P., 2005. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. J R Stat Soc Ser C Appl Stat 54, 127–142.
112. Shively, T., Kockelman, K., Damien, P., 2010. A bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics. Transportation Research Part B 44, 699–715.
113. Siddique, Z. Q. 2000. Accident Risk Modeling of Vehicle-to-Bicycle and Vehicle-to-Pedestrian Accidents at Four-Legged Signalized Intersections, Master Thesis, Asian Institute of Technology.

114. Trépanier, M., Leroux, M.H., Marcellis-Warin, N. 2009. Cross-analysis of hazmat road accidents using multiple databases. Accident Analysis & Prevention, 41(6), pp. 1192-1198.

115. Ulfarsson, G.F. Kim, S., Booth. K.M., 2010. Analyzing fault in pedestrian–motor vehicle crashes in North Carolina. Accident Analysis and Prevention, 42 (6), pp. 1805–1813.

116. Van der Zee, F.F., J. Wiertz, C.J.F. ter Braak, R.C. van Apeldoorn, J. Vink, 1992. Landscape Change as a Possible Cause of the Badger Meles meles L. Decline in The Netherlands. Biological Conservation, Volume 61, pp. 17–22.

117. Ver Hoef, J.M., Boveng, P.L., 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology, 88 (2007), pp. 2766–2772.

118. Wang, X. Abdel-Aty, M. 2006. Temporal and spatial analysis of rear-end crashes at signalized intersections. Accid. Anal. Prev., 38 (6), pp. 1137–1150

119. Wang, Y, Ieda, H., Mannering, F.L., 2003. Estimating Rear-End Accident Probabilities at Signalized Intersections: An Occurrence-Mechanism Approach. J. Transp. Eng. 129(4), 1-8.

120. Wang, Y., 1998. Modeling vehicle-to-vehicle accident risks considering the occurrence mechanism at four-legged signalized intersections. Ph.D. dissertation, The Univ. of Tokyo, Tokyo.

121. Wang, Y., Nihan, N.L. 2004. Estimating the Risk of Collisions between Bicycles and Automobiles at Signalized Intersections. Accident. Anal. Prev. 36(3), 313-321.

122. Wang, Y., Lao,Y., Wu, Y., Corey, J., 2010. Identifying High Risk Locations of Animal-Vehicle Collision for Washington State Highways. Transportation Northwest (TransNow) and Washington State Department of Transportation (WSDOT) Research Report WA-RD 752.1/TNW 2010-04.

123. Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16 (2), 275–289.

124. Washington, S.P., Karlaftis, M. G., Mannering, F. L., 2003. Statistical and econometric methods for transportation data analysis. Boca Raton: Chapman & Hall/carcass removal C, Boca Raton, Florida, 242-243.

125. Winkelmann, R., Zimmermann, K., 1995. Recent developments in count data modeling: theory and applications, J. Econ. Surveys 9(1), 1–24.

126. Wong, S.C., Sze, N.N., Li, Y.C. 2007. Contributory factors to traffic crashes at signalized intersections in Hong Kong. Accid. Anal. Prev., 39 (6), pp. 1107-1113.

127. Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accid. Anal. Prev. 50, 371-376.

128. Zamani, H., Ismail, N., 2010. Negative binomial-Lindley distribution and its application. Journal of Mathematics and Statistics 6 (1), pp. 4–9.

Curriculum Vitae

1. PERSONAL INFORMATION

1.1 EDUCATION

Doctoral:
Jun. 2008~Present

Major: Transportation Engineering
Dept. of Civil and Environmental Engineering, University of Washington, Seattle
Dissertation*: Nonlinear modeling for highway safety issues*

Master of Science: *Sep. 2005~Mar.2008*

Major: Transportation Engineering
School of Transportation Engineering, Tongji University, Shanghai, China
Thesis: *Dynamic forecasting method for traffic congestion at signalized intersection*

Bachelor of Science: *Sep. 2001~Jun.2005*

Major: Transportation Engineering
School of Transportation and Traffic Engineering, Jilin University, Jilin, China
Thesis: *Background updating algorithm for video-based vehicle detection under complicated environment*

1.2 RESEARCH INTERESTS

- Traffic Safety Analysis and Accident Modeling
- Sustainable Transportation System
- Transit Operation
- Intelligent Transportation Systems
- Transportation Planning
- Traffic Operation and Management
- Traffic Simulation
- Transportation Data Management and Analysis

1.3 PROFESSIONAL HISTORY

- **Intern Transportation Engineer**, Urban Planning Office, Washington State DOT, Seattle WA. (Aug. 2012~ Now)

- In charge of traffic dynamic assignment simulation and model calibration
- **Research/Teaching Assistant**, Smart Transportation and Application Research Laboratory (STAR Lab), Department of Civil and Environmental Engineering, University of Washington, Seattle. (Jun. 2008~ Now)
  - Safety analysis and modeling; traffic data management and quality control; traffic detection analysis and improvement; travel time reliability analysis and modeling; managed lane simulation
- **Research Assistant**, Transportation System Engineering & Intelligent Transportation System (ITS) Lab, School of Transportation Engineering, Tongji University, Shanghai, China. (Sep. 2005~Mar.2008)
  - Transportation infrastructure design for urban arterial; transportation facility planning, such as transit network planning and parking planning; traffic control and management; traffic simulation and modeling
- **Undergraduate Research Assistant**, School of Transportation and Traffic Engineering, Jilin University, Jilin, China. (Oct. 2004~Jun.2005)
  - Traffic video data processing and analysis
- **Undergraduate Research Assistant**, School of Transportation and Traffic Engineering, Jilin University,, Jilin, China. (Jun. 2004~Aug.2004)
  - Data analysis for highway snow disaster prevention techniques

2. PUBLICATIONS AND PRESENTATIONS

2.1 PEER-REVIEWED JOURNALS

1. Lin Hou, **Yunteng Lao**, Yinhai Wang, Zuo Zhang, Yi Zhang, and Zhiheng Li. Modeling Freeway Incident Response Time: A Mechanism-based Approach. *Transportation Research Part C*. Volume 28, pp. 87-100. Mar. 2013.

2. Yong Wang, Xiaolei Ma, **Yunteng Lao**, Yinhai Wang, and Haijun Mao. Vehicle Routing Problem: Simultaneous Deliveries and Pickups with Split Loads and Time Windows. Accepted for publication in *Transportation Research Record: Journal of the Transportation Research Board*, Aug. 2013.

3. **Yunteng Lao**, Guohui Zhang, Jonathan Corey and Yinhai Wang. Gaussian Mixture Model-Based Speed Estimation and Vehicle Classification Using Single Loop Measurements. *Journal of Intelligent Transportation Systems*. 16(4), pp. 184-196. 2012.

4. **Yunteng Lao**, Yao-Jan Wu, Yinhai Wang, and Kelly McAllister. Fuzzy Logic-based Mapping Algorithm for Improving Animal-Vehicle Collision Data. *Journal of Transportation Engineering*. 138(5), pp.520-526. 2012.

5. Runze Yu, **Yunteng Lao**, Xiaolei Ma, and Yinhai Wang. Short-Term Traffic Flow Forecasting for Freeway Incident Induced Delay. Accepted for publication in *Journal of Intelligent Transportation Systems*. Sep. 2012.

6. Xiaoyue Liu, Guohui Zhang, **Yunteng Lao**, Yinhai Wang. Modeling Traffic Flow Dynamics on Managed Lane Facility: A Cell Transmission Model Based Approach. *Transportation Research Record: Journal of the Transportation Research Board.* No. 2289, pp. 163-170. 2012.

7. **Yunteng Lao**, Yao-Jan Wu, Yinhai Wang, and Xiaoguang Yang. Applicability of Signalized Single and Double Phase Midblock Crossings in Highly Populated Areas. *Proceedings of the ICE – Transport*. 10.1680/tran.10.00027. 2012.

8. **Yunteng Lao**, Guohui Zhang Yao-Jan Wu, and Yinhai Wang. Modeling for Animal-Vehicle Collisions Considering Animal-Vehicle Interactions. *Accident Analysis and Prevention Journal*. 43(6), pp.1991-1998, Nov. 2011.

9. Jonathan Corey, **Yunteng Lao**, Yao-Jan Wu, and Yinhai Wang. Detection and Correction of Inductive Loop Detector Sensitivity Errors Using Gaussian Mixture Models. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2256, pp. 120-129. 2011.

10. **Yunteng Lao**, Yao-Jan Wu, Jonathan Corey, and Yinhai Wang. Modeling Animal-Vehicle Collisions Using Diagonal Inflated Bivariate Poisson Regression, *Accident Analysis and Prevention Journal*. 43(1), pp. 220-227. 2011.

11. Xiaoyue Liu, Guohui Zhang, **Yunteng Lao**, Yinhai Wang. Quantifying the Attractiveness of High-Occupancy-Toll Lane Under Various Traffic Conditions Using Traffic Sensor Data. *Transportation Research Record: Journal of the Transportation Research Board*. No. 2229, pp. 102-109. 2011.

12. Meiping Yun, **Yunteng Lao**, Xiao-Guang Yang. Game Theory-based Analysis on Utility of Transportation Management Policy. *Journal of Tongji University* (NATURAL SCIENCE). 38(4), pp. 527 -532. 2010. (Chinese Version)

13. Meiping Yun, Long Kejun, **Yunteng Lao**, Xiaoguang Yang. Modeling on Scale of Public Parking Lots Considering Parking Choice Probability. *Systems Engineering*. 26(2), pp. 84-88. 2008. (Chinese Version)

14. Jing Zhao, Xiaoguang Yang, Yu Bai and **Yunteng Lao**. Investigation on Impacts of Bus Lane on Capacities of Signalized Intersections. *Urban Transport of China*. 6(5), pp. 74-79, Sep. 2008. (Chinese Version)

15. Xiaoguang Yang, **Yunteng Lao** and Meiping Yun. Application of Different Pedestrian Cross Pattern to No-signal Controlled Segment. *Journal of Tongji University* (NATURAL SCIENCE). 35(11), pp. 1466-1469. 2007. (Chinese Version)

16. Xuemei Zhou, Xianzun Zhang, Xiaoguang Yang, **Yunteng Lao**. Travel Mode Choice-Based Prediction of Public Transit Demand. *Journal of Tongji University* (NATURAL SCIENCE). 35(12), pp. 1627-1631. 2007. (Chinese Version)

17. **Yunteng Lao**, Xiaoguang Yang et al. Evaluation of Traffic State Detection Method. *Computer and Communications*, pp. 74-77, Dec 2006. (Chinese Version)

18. Zhihui Li, **Yunteng Lao** and Dianhai Wang et al. Research Region Selective Update Mixture Gaussian Background Model In Hybrid Traffic Flow Video Detection Process. *Journal of ITS communication*. Vol. 7, pp.25-27. 2005. (Chinese Version)

2.2 JOURNAL ARTICLES UNDER REVIEW

19. Yong Wang, Xiaolei Ma, **Yunteng Lao**, Yinhai Wang, and Haijun Mao. A Hybrid Algorithm for Two-echelon Logistics Distribution Region Partitioning Problem. Submitted to *Journal of Computers & Operations Research*, Nov. 2012.

20. **Yunteng Lao**, Jonathan Corey, and Yinhai Wang. Applicability of Conditional Left Turn Phase Reservice Strategies. Submitted to *IEEE Transactions on Intelligent Transportation Systems*, Nov. 2012.

21. Yong Wang, Xiaolei Ma, **Yunteng Lao**, Yinhai Wang, and Haijun Mao. A Fuzzy-based Customer Clustering Approach with Hierarchical Structure for Logistics Network Optimization. Submitted to *Transportation Research Part E*, Nov. 2012.

22. Yong Wang, Xiaolei Ma, **Yunteng Lao**, Yinhai Wang, and Haijun Mao. A Two-stage Heuristic Method for Vehicle Routing Problem with Split Deliveries and Pickups. Submitted to *Journal of Annals of Operations Research*, Nov. 2012.

23. **Yunteng Lao**, Aivis Grislis, Yao-Jan Wu, and Yinhai Wang. Parameters Influencing Single-vehicle Large Truck Accidents on Rural Two Lane Roads in Washington State. Submitted to *Safety Science*. Jul. 2012.

24. Perrine Kenneth, **Yunteng Lao**, Jun Wang, and Yinhai Wang. Area-wide System for Coordinated Ramp Meter Control. Submitted to *ASCE Journal of Computing in Civil Engineering. Jul*. 2012.

25. **Yunteng Lao**, Guohui Zhang and Yinhai Wang. Modeling Vehicle Crashes Using

Generalized Nonlinear Model, Submitted to *Accident Analysis and Prevention*, April. 2012.

26. Lin Hou, **Yunteng Lao**, Yinhai Wang, Zuo Zhang, Yi Zhang, and Zhiheng Li. Exploring Time-varying Effects Of Influential Factors On Incident Clearance Time Using A Non-proportional Hazard-based Model. Submitted to *Transportation Research Part A*. Sep. 2011.

27. Jonathan Corey, **Yunteng Lao**, and Yinhai Wang. Quantifying and Comparing Left Turn Strategy Performance. Submitted to *Transportation Research Record: Journal of the Transportation Research Board*, Aug. 2012.

2.3 PEER-REVIEWED CONFERNCE PAPERS

28. Yong Wang, Xiaolei Ma, **Yunteng Lao**, Yinhai Wang, and Haijun Mao. Location Optimization of Multiple Distribution Centers Under Fuzzy Environment. Accepted by the 92th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2013.

29. Jonathan Corey, **Yunteng Lao**, Xin Xin and Yinhai Wang. Improving Intersection Performance with Left Turn Phase Reservice Strategies. 15th International IEEE Conference on Intelligent Transportation Systems (ITSC 2012) 2012. pp. 403-408.

30. Jian Xu, Xiaoguang Yang and **Yunteng Lao**. Study on the Capacity of Left-Through Shared Lane with Permitted Left-Turn Phasing. 2008 International Conference on Intelligent Computation Technology and Automation (ICICTA). 2008. pp. 498 – 503.

31. Meiping Yun, **Yunteng Lao**, Yingying Ma, and Xiaoguang Yang. Optimization Model on Scale of Public Parking Lot Considering Parking Behavior. Proceedings of the Eighth International Conference of Chinese Logistics and Transportation Professionals. Chengdu, China. 2008.

32. **Yunteng Lao** and Xiaoguang Yang. Quantification of congestion on signalized intersection based on detector data. The 10th International IEEE Conference on Intelligent Transportation Systems. Seattle, USA. 2007.

33. Wu, Su Feng, Meiping Yun, Xiaoguang Yang, Yang Yang and **Yunteng Lao**. The travel choice behavior with traffic state information. The 10th International IEEE Conference on Intelligent Transportation Systems. Seattle, USA. 2007.

34. **Yunteng Lao** and Xiaoguang Yang. Identification for road traffic state. The 11th World Conference on Transport Research, Berkeley, USA. 2007.

35. **Yunteng Lao**, Tang Shoupeng, Xiaoguang Yang et al. Evaluation method for accuracy of

road traffic state information. The First International Conference of Transportation Engineering. Chengdu, China. 2007.

36. Xuemei Zhou, Xiaoguang Yang, and **Yunteng Lao**. Public Transport Service Level Influence on Travel Mode Choice. The First International Conference of Transportation Engineering. Chengdu, China. 2007.

37. **Yunteng Lao**, Xiaoguang Yang and Chu Hao. Optimization Method for Traffic State Detection Algorithm Based on Driver Evaluation. The 14th World Congress on ITS. Beijing, China. 2007.

38. **Yunteng Lao**, Tang shoupeng, Xiaoguang Yang et al. Research on Determination for Identification Indexes of Traffic State. 7th national youth conference on transport. Tianjing, China. 2007. pp.396-399. (Chinese Version)

39. Chen Rongkang, **Yunteng Lao** and Li Zhihui. The Study of Video Vehicle Detection for Background Update Algorithm under Complicated Environment. Research and Practice on Intelligent transportation System. Shanghai, China. 2005. pp.80-86, August 2005. (Chinese Version)

2.4 RESEARCH REPORTS

40. Yinhai Wang, Jonathan Corey, **Yunteng Lao**, and Xin Xin. Criteria for the Selection and Application of Advanced Traffic Signal Control Systems. Transportation Northwest (TransNow) Research Report TNW2012-. Jun. 2012.

41. Yinhai Wang, **Yunteng Lao**, Cathy Liu and Guangning Xu. Simulation-Based Testbed Development for Analyzing Toll Impacts on Freeway Travel. Transportation Northwest (TransNow) Research Report TNW 2011-11. Nov. 2011.

42. Yinhai Wang, Runze Yu, **Yunteng Lao**, and Timothy Thomson. Quantifying Incident-Induced Travel Delays on Freeways Using Traffic Sensor Data: Phase II. Transportation Northwest (TransNow) and Washington State Department of Transportation (WSDOT) Research Report WA-RD 752.3/TNW 2010-07. June 2010.

43. Yinhai Wang, **Yunteng Lao**, Yao-Jan Wu, and Jonathan Corey. Identifying High Risk Locations of Animal-Vehicle Collision for Washington State Highways. Transportation Northwest (TransNow) and Washington State Department of Transportation (WSDOT) Research Report WA-RD 752.1/TNW 2010-04. June 2010.

44. Yinhai Wang, Jonathan Corey, **Yunteng Lao**, and Yaojan Wu. Development of a Statewide Online System for Traffic Data Quality Control and Sharing. Transportation Northwest (TransNow) Research Report TNW2009-12. Oct. 2009.

45. Yinhai Wang, Kenneth Perrine, and **Yunteng Lao**. Developing an Area-Wide System for Coordinated Ramp Meter Control. Transportation Northwest (TransNow) Research Report TNW2008-11. Oct. 2008.

2.5 PROFESSIONAL PRESENTATIONS

1. **Yunteng Lao**, Guohui Zhang and Yinhai Wang. Gaussian Mixture Model-Based Speed Estimation and Vehicle Classification Using Single Loop Measurements. Presented at 91st Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2012.

2. Lin Hou, **Yunteng Lao**, Yinhai Wang, et al. Exploring Time-varying Effects Of Influential Factors On Incident Clearance Time Using A Non-proportional Hazard-based Model. Presented at 91st Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2012.

3. Lin Hou, **Yunteng Lao**, Yinhai Wang, et al. Modeling Freeway Incident Response Time: A Mechanism-based Approach. Presented at 91st Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2012.

4. Xiaoyue Liu, Guohui Zhang, **Yunteng Lao**, Yinhai Wang. Modeling Traffic Flow Dynamics on Managed Lane Facility: A Cell Transmission Model Based Approach. Presented at 91st Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2012.

5. **Yunteng Lao**, Guohui Zhang Yao-Jan Wu, and Yinhai Wang. Modeling for Animal-Vehicle Collisions Considering Animal-Vehicle Interactions. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2011.

6. **Yunteng Lao**, Yao-Jan Wu, Yinhai Wang, and Kelly McAllister. Fuzzy Logic-based Mapping Algorithm for Improving Animal-Vehicle Collision Data. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2011.

7. **Yunteng Lao**, Yao-Jan Wu, Jonathan Corey, and Yinhai Wang. Modeling Animal-Vehicle Collisions Using Diagonal Inflated Bivariate Poisson Regression. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2011.

8. Jonathan Corey, **Yunteng Lao**, Yao-Jan Wu, and Yinhai Wang. Detection and Correction of Inductive Loop Detector Sensitivity Errors Using Gaussian Mixture Models. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan.

2011.

9. Xiaoyue Liu, Guohui Zhang, **Yunteng Lao**, Yinhai Wang. Quantifying the Attractiveness of High-Occupancy-Toll Lane Under Various Traffic Conditions Using Traffic Sensor Data. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2011.

10. Runze Yu, **Yunteng Lao**, Xiaolei Ma, Yinhai Wang. Short-Term Traffic Flow Forecasting for Improved Estimates of Freeway Incident Induced Delays. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2011.

11. **Yunteng Lao**, Yao-Jan Wu, Yinhai Wang, and Xiaoguang Yang. Applicability of Signalized Single and Double Phase Midblock Crossings in Highly Populated Areas, Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2009.

12. Perrine Kenneth, **Yunteng Lao**, Wang Yinhai. Areawide System for Coordinated Ramp Meter Control. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2009.

13. **Yunteng Lao.** Congestion Quantification of Two Kinds of Traffic Bottlenecks for Urban Road. The 4th Joint Seminar between Tongji and Kyoto University in Transportation & Urban Management. Shanghai, China. Mar. 2008.

14. **Yunteng Lao** and Yang Xiaoguang. Quantification of Congestion on Signalized Intersection Based on Detector Data. The 10th International IEEE Conference on Intelligent Transportation Systems. Seattle, USA. Oct. 2007.

15. **Yunteng Lao** and Yang Xiaoguang. Identification for Road Traffic State. The 11th World Conference on Transport Research, Berkeley, USA. Jun. 2007

16. **Yunteng Lao**, Tang Shoupeng, Yang Xiaoguang et al. Evaluation Method for Accuracy of Road Traffic State Information. The First International Conference of Transportation Engineering. Chengdu, China. 2007.

17. **Yunteng Lao**, Yang Xiaoguang and Chu Hao. Optimization Method for Traffic State Detection Algorithm Based on Driver Evaluation. The 14th World Congress on ITS. Beijing, China. 2007.

18. **Yunteng Lao**, Tang shoupeng, Yang Xiaoguang et al. Research on Determination for Identification Indexes of Traffic State. 7th national youth conference on transport. Tianjin, China. 2007.

Reprint Permissions

Permissions have been granted by the publishers to reuse the contents in the papers listed below in this dissertation.

1. **Yunteng Lao**, Guohui Zhang Yao-Jan Wu, and Yinhai Wang. Modeling for Animal-Vehicle Collisions Considering Animal-Vehicle Interactions. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2011.

2. **Yunteng Lao**, Yao-Jan Wu, Yinhai Wang, and Kelly McAllister. Fuzzy Logic-based Mapping Algorithm for Improving Animal-Vehicle Collision Data. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2011.

3. **Yunteng Lao**, Yao-Jan Wu, Jonathan Corey, and Yinhai Wang. Modeling Animal-Vehicle Collisions Using Diagonal Inflated Bivariate Poisson Regression. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., USA. Jan. 2011.