The Pennsylvania State University

The Graduate School

Department of Civil and Environmental Engineering

METHODOLOGICAL APPROACHES TO INCORPORATE HETEROGENEITY IN TRAFFIC ACCIDENT FREQUENCY MODELS

A Thesis in

Civil Engineering

by

Sittipan Sittikariya

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2006

The thesis of Sittipan Sittikariya was reviewed and approved* by the following:

Venkataraman Shankar
Associate Professor of Civil Engineering
Thesis Advisor
Chair of Committee

Martin Pietrucha
Associate Professor of Civil Engineering

Thorsten Wagener
Assistant Professor of Civil Engineering

Evelyn Thomchick
Associate Professor of Supply Chain Management

Peggy Johnson
Professor of Civil Engineering
Head of the Department of Civil and Environmental Engineering

*Signatures are on file in the Graduate School

ABSTRACT

This dissertation addresses the impact of critical modeling issues on the statistical significance of specifications correlated with traffic accident frequencies. Also, this research intends to shed new light on the major factors contributing to the occurrence of median crossover accidents, which are particularly severe. Furthermore, the research aims to lay out methodologies that provide significant improvements in frequency predictions. The entire highway network of Washington State is used as the empirical setting. The database was developed over several years of observations and includes median crossover accident data, geometric and traffic volume information as well as weather data. In traffic accident databases that are comprised of cross-sectional panels, unobserved effects or heterogeneity, correlation due to shared unobserved effects through accident counts, and excess zero problems are common issues in the estimation of accident count models. I establish various model structures through classical frequentist and Bayesian approaches to explore these issues. Several hundred modeling specifications were assessed to determine the most meaningful and robust specifications. In the final analysis, the same set of specifications was used to compare the results across model structures. Preliminary results showed variabilities in statistical significance of key parameters; but the parameters themselves do not change significantly. Much of the variability in the standard errors of parameters results from error structure assumptions. For example, hierarchical structures, or multiplicative structures that are specified exogenously through Bayesian methods help improve predictions. The major contribution of this dissertation is the development of modeling taxonomies where assumptions under frequentist and Bayesian methods are incorporated through examination of model equivalencies. Such taxonomies are helpful in illustrating with greater clarity the relative usefulness of frequentist and Bayesian methods as they relate to all types of accident frequencies. In this sense, this dissertation offers a general framework. In addition, from a programming standpoint, the efficiency and consistency of highway safety prioritization can be broadly addressed through this framework.

Table of Contents

List of Tables

List of Figures

Glossary

β	Estimated Coefficient
σ	Standard Error
t	t-Statistic
AADT	Average Annual Daily Traffic
BHP	Bayesian Heterogeneity Poisson
HB	Hierarchical Bayes
HBHP	Hierarchical Bayesian Heterogeneity Poisson
NB	Negative Binomial Model
NM	Negative Multinomial Model
RENB	Random Effects Negative Binomial Model
REP	Random Effects Poisson Model
ZIP	Zero-Inflated Poisson Model
ZINB	Zero-Inflated Negative Binomial Model

Acknowledgements

I would like to express my heartfelt gratitude to my advisor, Professor Venky Shankar, for guidance, inspiration and lifetime fellowship. His guiding light always shines during the darkest times and never fails on the brightest days. I would like to thank all supervisory committee members, Professor Martin Pietrucha, Professor Thorsten Wagener and Professor Evelyn Thomchick, on their valuable contributions.

TIMG fellows have been providing tremendous contributions to this dissertation. I want to reiterate my thanks to Ming-Bang Shyu and Songrit Chayanan for providing a good company in the long days and nights spent in the laboratory, advice, support, knowledge and most of all lifetime friendship.

This dissertation would not have been possible without the encouragement from my family; my father Ananchai Sittikariya and my mother Wipawan Sittikariya. My parents are the ones who first taught me how to make my idea tangible. Now, this dissertation is finally tangible and I hope that it can represent some fraction of the immense gratitude and love I have for them.

Last but not least, I would like to express my highest appreciation to Napatskorn Poonlapyot and her mother, Thanutporn Poonlapyot, for unfailing support and endless optimism as well as unlimited love. Without both of them, I would not have been able to accomplish this dissertation.

Chapter 1
INTRODUCTION

1.1 Motivation and Research Objectives

There are some common modeling problems in the estimation of accident count models. Unobserved heterogeneity is a common theme in accident occurrence, leading to the well-known overdispersion problem (Shankar et al., 1995). Two major sources causing overdispersion are high-valued crash counts or excess zeros due to under-reporting. The negative binomial (NB) model is suitable for overdispersed accident frequencies, as shown in several prior works (see for example Poch and Mannering, 1996; Milton and Mannering, 1996).

Another common issue that arises in accident data contexts is the issue of correlation among accident counts. Multiple years of cross-sectional data on highway accident occurrences are often available from public domains, including time series information on traffic volumes, accident counts and roadway geometrics as well as roadside characteristics. In addition, weather information is also available from national databases (see for example Shankar et al., 2004). It is noted here that correlation among accident counts and unobserved heterogeneity can be viewed as a simultaneity problem. That is, due to explicit correlation among dependent variables, or implicit correlation among the error structures, a simultaneity issue that argues for joint density functions arises. Complicating the decomposition of this simultaneity issue is the "mathematical overlap" between correlation and heterogeneity. From an empirical standpoint, it is highly possible that serial correlation may be spuriously captured as heterogeneity; the converse is also true. Fundamentally however, heterogeneity and serial correlation can be classified broadly as violations of the "independently and identically distributed" (IID) assumption. In the context of count models, as is the case in this dissertation, the empirical question relates to the proportion of the heterogeneity component of the IID violation problem. Some empirical examples follow to illustrate why treating the IID problem in count contexts as primarily a heterogeneity issue is useful.

In the presence of accident count correlation in multiple years of cross-sectional accident count data, the efficiency of parameter estimates comes into question. Similar to the classical linear regression model, one can expect parameter estimates to be inefficient in the presence of correlation in count models (Guo, 1996). A method to adjust for repeated observation effects on parameter estimates is necessary to adjust for heterogeneity. One such method relates to the use of the cluster heterogeneity model or negative multinomial (NM) model (Ulfarsson and Shankar, 2003). In that model, a joint likelihood based on repeated observations is constructed to modify the traditional negative binomial likelihood. As a result, parameter estimates reflect an adjustment in their standard errors, with much of the adjustment resulting from proximate years. The second type of adjustment for correlation involves the development of random effects, where segment-specific heterogeneity is accounted for. In the traffic accident context, the random effects approach has been shown to be reasonable (See for example, Shankar et al., 1998). The empirical evidence in the above-mentioned literature points heavily to the usefulness of characterizing parameter estimation in longitudinal accident datasets as fundamentally heterogeneity problems. This is not to say serial correlation is irrelevant. Rather, one has to argue the case for serial correlation in count data contexts related to traffic accidents as primarily an error lag problem. The reason for this argument is that the exogenous regressors in longitudinal accident datasets seldom vary. With the exception of traffic volumes and weather effects, roadway geometrics and other design related variables are practically constant. Much of the dynamics in longitudinal accident datasets arises from overlapping heterogeneity effects captured by the error term as a "group of omitted variables" effect. Akin to the linear context, if the omitted variables are correlated with the exogenous regressors, parameter bias is highly likely. Hence, at the very least, short of rigorous mathematical treatment through alternative estimators, treating the omitted variables problem as a heterogeneity effect as group specific effects over time or segment mitigates the potential for parameter bias.

The last common issue that arises in traffic accident contexts is one that pertains to partial observability. Partial observability relates to under-reporting of less severe accidents. As a result, some roadway segments appear to have perfect safety histories in the form of

zero counts. However, this problem of zero counts may be a deceptive one. For example, especially in events of potentially high severity such median crossover accidents, the problem of excess zeros is a significant one. A highway segment with historically excess zeros may appear to be a safe location, but there is no guarantee that no median crossover accidents will occur in the future. As interactions between traffic volumes and geometrics increase, and geometrics and weather conditions exacerbate accident "proneness," one would expect that over time, with increase in traffic volumes and adverse weather conditions, non-zero count probabilities would be expected to rise. Limited history of observation may not adequately capture the "accident proneness" of the highway segment, as interactions increase over time.

1.2 Research Approach and Description

This dissertation addresses the common modeling problems mentioned above, namely unobserved heterogeneity, correlation of accident counts and the excess zero problem, in highway traffic safety using various model structures derived from "frequentist" and "Bayesian" approaches. The frequentist approach is built on the notion of repeated sampling, which is mainly an abstract notion, and not a notion built on the sample at hand. In this context, the Bayesian approach has potential to lend more insight, especially on the predictive aspect of accident modeling. The intent of this dissertation is hence to provide an illustrative essay of comparisons and contrasts between the frequentist and Bayesian methods in accident contexts. The research also intends to explore the robustness of the estimated parameters under the impact of these modeling issues. Furthermore, this research will shed new light on the major factors contributing to the occurrence of median crossover accidents. Finally, the research aims to shed more light into methods, especially on the Bayesian dimension, in terms of improved accident frequency predictions.

First, the "frequentist" approach is used to estimate models with different model structures using the traditional statistical software such as Limdep and Gauss. Both Limdep and Gauss have an array of optimization routines available for the commonly

occurring list of dataset and minimization problems. The optimization routines are cross-validated between Gauss and Limdep to ensure that common optimization issues such as start values and data scaling do not influence final parameter estimates. A rigorously cross-validated set of parameters from the frequentist approach serves as reliable benchmark for the application of hierarchical Bayes procedures. The highway crash literature has recently evidenced an increased interest in the use of empirical Bayes techniques. Work in the use of hierarchical Bayes (HB) applications is however limited in the field. On the contrary, as an area of application in the area of forecasting as a whole, HB methods have shown significant promise. As a result the Bayesian approach is introduced in this study to address the same critical modeling issues from different perspectives and beliefs. Winbugs is the primary Bayesian software allowing the user to customize estimation algorithms corresponding to prior belief of data and different modeling structures. Both "frequentist" and "Bayesian" approaches contain two processes; namely single-state process and dual-state process in count families. In single-state process all accident frequencies (e.g. accident counts and zero accident) are jointly used to estimate the model. On the other hand the dual-state process separately treats accident counts and zero accident through non-zero accident probability state and zero accident probability state respectively.

Considering the occurrence of median crossover accident, the "frequentist" modeling framework includes the heterogeneity Poisson model with gamma and normally distributed heterogeneity. This type of model assumes that the unobserved effects for each individual segment vary across time and space. One may argue that the differences in heterogeneity in the same roadway section across time are minimal because the geometrics of roadway and roadside remain the same and the traffic volumes across the year are proximate with usual differences equal to the nominal growth rate of 3 to 4 percent. The statistically significant differences are mainly group-specific, in this case, specific to the group of years for each segment. To address group-specific effects, "cluster" heterogeneity models are developed. Cluster heterogeneity models can address a variety of group effects – in addition to the group types mentioned here; they may help address segment clusters where segments are different. This empirical case is not

considered in this dissertation; but the logical extension of the cluster heterogeneity technique is straightforward. Cluster heterogeneity models in this dissertation can be viewed as a taxonomical family of models. In particular, negative multinomial models, random effects Poisson models with gamma and normally distributed heterogeneity and random effects negative binomial models with gamma distributed heterogeneity are considered as part of the taxonomy. In addition to cluster effects in single-state processes, zero-inflated Poisson models (ZIP) with empirical adjustment for standard errors are investigated to explore effects of the repeated excess zero problem on the robustness of estimated parameters.

To gain comprehensive insights into the effects of heterogeneity in accident count models, "Bayesian" analysis is used in comparison with the "frequentist" approach. The Bayesian heterogeneity Poisson (BHP) is the base model and equivalent to the heterogeneity Poisson in the "frequentist" approach. The equivalence arises from the treatment of error structures to capture unobserved heterogeneity. Error structures can be gamma-distribution-based or normal-distribution-based. Regardless of the estimation consequences unique to the normal distribution assumption (i.e., non-closed form issues), the main idea is to derive a marginal distribution by starting off with a conditional Poisson assumption. This primary model is extendable to Bayesian segment-specific effects Poisson structures. Similar to the cluster heterogeneity Poisson models in the "frequentist" approach; this model also assumes that the unobserved effects in the same roadway section are identical regardless of the fact that accident counts are repeated over multiple years of observation. In addition, hierarchical Bayesian heterogeneity Poisson (HBHP) is introduced to the study to explore the robustness of explanatory variables and the accuracy of accident frequency predictions. The hierarchy of the BHP arises from the assumptions underlying the variances of the Bayesian prior which will be discussed in a later section of this dissertation. The dual-state Bayesian structure incorporates a splitting regime to account for excess zeros problem. The splitting regime can be modeled as a logistic function, with the regular count portion being modeled as negative binomial distribution. Given these basic modeling premises, the rest of this dissertation is organized as follows:

Chapter 2 elaborates the relevance of the problem to transportation infrastructure programming and policy. In doing so, the scope of the big-picture motivating the problems discussed in this dissertation is explained.

Chapter 3 presents the empirical setting and serves as a descriptive backdrop of databases used to develop the various models. The vector of regressors used in the development of models is large; so only the key variables are provided along with their descriptive characteristics.

Chapter 4 provides an analytical framework designed for resolving unobserved effects, correlation among accident counts and excess zero problems. The framework provides a comprehensible picture of the scope of the traffic accident frequency model context.

Chapter 5 presents mathematical formulations of all models referred in the research; however, the information presented in this chapter does not intend to be a self-containing text. While the mathematic formulations and descriptions provided are intended to provide a basic understanding of the backgrounds of each modeling approach, the reader is urged to consult the references section for in-depth review of the mathematical constructs.

Chapter 6 presents results from both single-state and dual-state processes embodied in the classical frequentist approach and in the Bayesian approach. There is a vast amount of empirical work to be conducted prior to making definitive statements about "prior" assumptions and critical modeling issues for HB analysis. It is advisable that such work be performed at great length over several dissertations on single-state processes of accident frequency, prior to pursuit of HB issues for multi-state processes. All results are presented in the tables along with interpretations of the regressors. The findings on the impact of critical modeling issues are also summarized in this chapter. At the end of this chapter, the prediction tests of the models in both classical frequentist approach and Bayesian approach and the structural changes in parameter (stability) tests were presented.

Finally, in chapter 7, findings and summaries of the body of the work and recommendations for future research are presented.

Chapter 2
RELEVANCE OF THE PROBLEM TO TRANSPORTATION INFRASTRUCTURE PROGRAMMING AND POLICY

Traffic accidents occur in various forms. The less severe forms such as rear-end accidents result in annualized societal costs of the order of 6,000 to 7,000 dollars (per the National Safety Council). At the other end of the spectrum, high-severity accidents such as incapacitating/disabling and fatal accidents exact a high economic toll on society. Incapacitating accidents on the average range around 300,000 dollars while fatal accidents cost in excess of 3 million dollars. It is essential then that infrastructure policy focus on the prevention and reduction of high-severity accidents. High-severity accidents as a matter of proportion of frequencies, amount to less than five percent of all accidents typically, while low-severity accidents such as property-damage only amount to 60 percent. In this dissertation, I focus on methodologies that can help address the afore-mentioned critical modeling issues, as they relate to high-severity accidents. The intent is to arrive at a methodology that eventually can provide us reliable benefit-cost insights into policy alternatives. Median crossover accidents are generally severe and result in a high cost to society. Such accidents also have a greater potential for creating liability, both because of their severity and because of the inherent link with design deficiencies, i.e. no or weak median barriers. In response to these issues, state and federal departments of transportation are creating systematic processes for determining infrastructure policy related to median crossovers. As an example, the Washington State Department of Transportation (WSDOT) is employing research findings from portions of this dissertation to determine median barrier requirements on state highways. Methodologies are also being developed to address roadway safety design issues related to divided highways in a cost-effective manner. WSDOT is re-examining current median barrier installation guidelines, which are ad hoc and do not make use of current multivariate statistical techniques that can account for effects from roadway geometric factors, roadside characteristics, traffic, and the weather.

Studies on median barrier requirements do not currently benefit from advanced analysis of median crossover accidents. Limited safety analysis techniques have been researched and used in the area of median barrier accidents; see, for example, (Graf and Winegard, 1968; Ross, 1974 and Bronstad, Calcote and Kimball, 1976). The current version of the American Association of State Highways and Transportation Officials (AASHTO) Roadside Design Guide (AASHTO, 1996) suggests the usage of a simple bivariate analysis of average daily traffic (ADT) and median width as a guide in determining median barrier requirements. The relationship is a simple decision rule chart, if ADT is above a certain value and median width below a certain value the chart shows the recommended type of median barrier. Currently, the WSDOT uses a related ADT versus median width relationship when examining median barrier requirements.

The "median treatment study on Washington State Highways" reported by Glad et al., 2002 presents a fairly clear picture of current WSDOT practice. As the report states, "WSDOT guidance for the installation of median barrier (Figure 700-7 in the WSDOT *Design Manual*) is essentially the same as that provided in the AASHTO Roadside Design Guide. AASHTO guidance was developed using a study conducted by the California DOT in 1968. This guidance provides criteria for median barrier installation based on the average daily traffic (ADT) and width of median. The criteria for barrier protection indicates that the designer should "evaluate the need for barrier" on all medians up to 32.8 feet in width when ADT is 20,000, or greater. Barrier is optional for all medians between 32.8 feet and 50 feet or when the median is less than 32.8 feet and the ADT is less than 20,000. AASHTO indicates "barrier not normally considered" for median widths greater than 50 feet."

The report also suggests that there is significant variability in median barrier installation practice at the state level. It states that "the North Carolina Department of Transportation recommends median barrier installation for all new construction, reconstruction, and resurfacing projects with medians 70 feet or less in width. California Department of Transportation has adopted more stringent warrants based on ADT for freeways with medians less than 75' in widths."

The Glad et al., 2002 study also recommends that the 50-foot width requirement for installing a median barrier is optimal from a benefit-cost analysis of observed crash histories on Washington State highways. This finding is based on the comparison of societal costs of median crossovers on sections with and without cable median barrier treatments. The study carefully notes that it did not account for "regression-to-the-mean" effects when considering the impacts of median barrier installation. That is, would median crossovers decreases naturally to a lifetime mean, even without cable barrier installation? If so, how does that affect the true effect of cable barriers? However, limitations exist in the Glad et al. study. The above methodologies do not facilitate accurate predictions of median crossover accidents. The study is strictly historical, and is fairly sensitive to whether or not fatalities occurred in the observed time period. Due to the high cost of fatalities, the absence of a fatal collision can provide for significant improvements in benefit cost ratios for a given type of median barrier treatment. There is much to gain from accurate predictions of median crossover frequencies. Installing median barriers effectively reduces median crossover rates to near zero, although there is always a small chance of median barrier penetration. However, it is not beneficial to install median barriers everywhere on the road network because the frequency of other types of accidents tends to increase in the presence of barriers, while reducing the propensity for a median crossover accident. Median barriers reduce the area that vehicles have to recover from, or escape, an accident in the roadway, and they cause rebound accidents when vehicles strike the barrier and rebound to strike another vehicle traveling in the same direction. Generally median barriers reduce the frequencies of injury accidents, particularly severe accidents. However, contrary examples exist that show that the case is not that simple. For example, a before and after study (Seamons and Smith 1991) found a total increase of roughly 14 percent in all injury (including fatal injury) accidents when median barriers were installed at freeway locations. Median crossover accidents tend to be more serious, with a higher probability of fatalities, whereas barriers might sometimes increase the probability of some types of accidents but most at lesser severity. Median barriers also carry a maintenance cost which is threefold: direct monetary cost, traffic delays, and risk to road crews. A Cost-benefit analysis that considers the whole project in addition to the predictive models is therefore an important

part of developing a full picture to help plan effective measures. Departments of Transportation attempt to strike a balance when scheduling sections for barrier installation and the decision making benefits from an accurate prediction of median crossover frequency. The need for trade off between the economic and safety priorities and a possibility to strike a balance stands in favor of a need for advanced median crossover accident analysis.

With the objective to obtain accurate predictions of median crossover frequencies, it is desired to develop advanced predictive models for median crossover accident frequency. There are some common modeling problems in the estimation of count models. Unobserved heterogeneity is a common occurrence in accident databases, leading to overdispersion. In median crossover accident databases, there is significant number of zero accident counts that suggest possible latent processes at work leading to spurious overdispersion. Another concern is possible correlation among accident counts. This type of correlation between the accident counts for a single section over many time periods (years, usually) causes the coefficient estimates to be inefficient and the estimated standard errors to be biased. Since median accident distributions occur in much lower numbers compared to other accident types (i.e. most of the observed accident counts are zeros) and likely in a more sporadic fashion, a longitudinal history of median accident counts is used to examine fundamental propensities in the long-term for median crossovers. This leads to a possible excess zero problem. The objective of this research is to arrive at such multivariate models that account for forms of unobserved heterogeneity, correlation among accident counts and excess zero counts. This could be achieved by formulation of variations of count models derived from the basic structure of a Poisson model to incorporate the effects of unobserved heterogeneity, accident count correlation and excess zero problem through a frequentist approach and a Bayesian approach. These models will help evaluate the effectiveness of the barrier installation on roadway sections in Washington State in improving the safety in terms of reducing the frequency of median crossovers and accidents of different severities. Also, from a programming standpoint, the efficiency and consistency of modeling median crossovers would be improved. Future research could be devoted to further improving the state-of-

the-practice in Washington State for programming median crossover safety and devising better schemes for reducing median crossover frequency.

Chapter 3
EMPIRICAL SETTING

3.1 Data Assembly

The Washington State consists of 6 regions as presented in figure 3.1 according to regional designation of the Washington State Department of Transportation (WSDOT). The total length of state highways in Washington State highway system is approximately 7,000 center-line miles. In order to pursue the stated objectives, longitudinal data for the period 1990 to 1994 containing crossover accident information on unbarriered medians on the Washington State highway network was used. The longitudinal study is especially useful for median crossovers due to the sporadic nature of median crossovers. The 1990-94 dataset consisted of 275 unbarriered highway sections over the entire Washington State highway network, totaling a length of nearly 670 center-line miles. Mean crossover frequency was 0.24 crashes per year, while per-lane average daily traffic was approximately 7,400 vehicles. Mean median width was 57 feet, with approximately 70 percent of all sections in the 40-foot to 75-foot median width range.

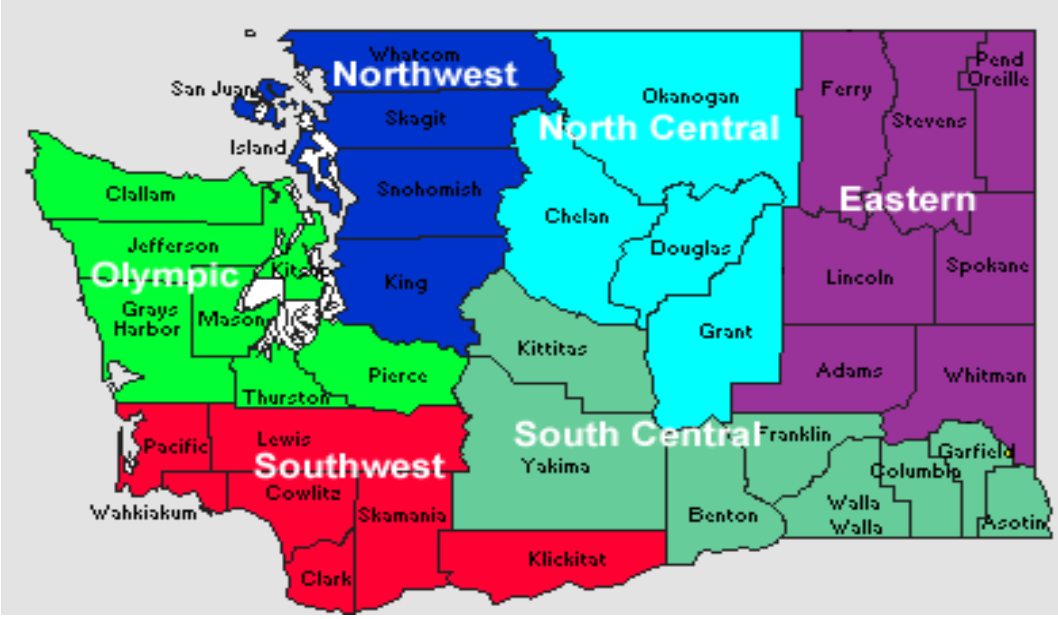


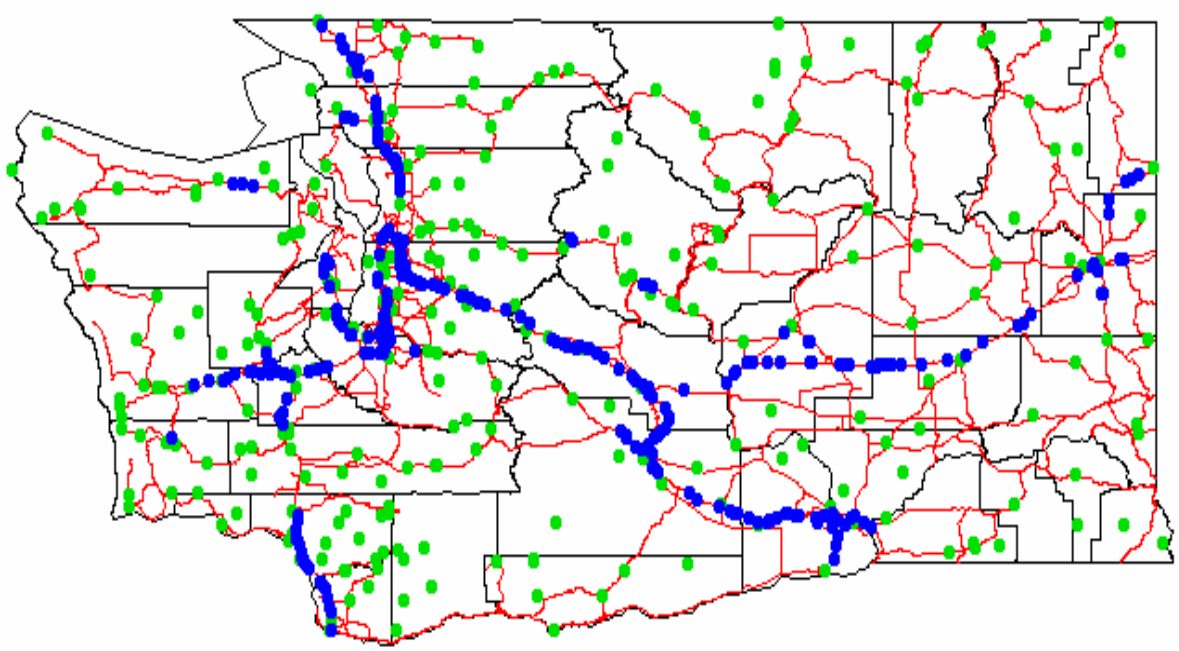
Figure 3.1 The Washington State Department of Transportation Regions
Source: www.wsdot.wa.gov

All accident records responded by the Washington State Patrol on the Washington State Highway system were recorded in the police report. The WSDOT developed the accident database, the Washington State Master Accident Record System (MARS), based on these police records. The annual records in the database contain almost every roadway and roadside incident that occurs during the year. The accident records from this database are the primary source of data for the current median crossover accident study.

The panel data in this research consists of five years (1990–1994 inclusive) of annual accident counts for 275 roadway sections in Washington State. As a result, the total number of observations in the database is 1,375 observations. The panel data is balanced, with all sections having a full five-year history. This panel data represents all sections (longer than 2,624 ft) without median barriers on divided state highways. The reasons why only sections longer than 2,624 ft are selected are that about 95 percent of shorter sections on divided highways have barriers, and that the shorter sections are more affected by access controls and intersections (Ulfarsson and Shankar, 2003). To ensure adequate representations of all functional classes (e.g. collector, principal arterials and interstate), the data is comprised of 2 collector sections, 84 principal sections and 189 interstate sections. It is noted here that a collector with a low posted speed limit is not the critical location for the occurrence of median crossover accidents.

In addition to median crossover accident information, other components of data extracted from the database included roadway geometrics, median characteristics and traffic volumes. The longer section might introduce the heteroskedasticity in the section; for example; geometric information such as the number of lanes at the beginning of the section is different from that at the end of the section. The geometric and traffic data was aggregated using a weighted average from the section listed in the database. The geometric data contains roadway widths, lane widths, number of lanes, shoulder widths, horizontal curve information, legal speed limit, surfacing type, terrain, median widths and so on. The traffic data includes average annual daily traffic (AADT) and truck volume as percentage of AADT. However, the database did not provide any of the weather information required for the study.

The GIS program ArcView 3.2 was used to match the roadway sections to their weather attributes stored in the historical weather database provided by the Western Regional Climate Center. The mapping criteria involved linking the non-median barrier roadway sections to the nearest corresponding weather stations. Each weather station provided climate data including daily, monthly and annual measurements of temperature, precipitation and snowfall including snow depth, with the records dating back to 1948. The selected roadway sections and all weather stations in Washington State are illustrated in figure 3.2.



*Blue (darker color) = The Selected Roadway Sections, **Green (lighter color) = The Weather Stations

Figure 3.2 Selected Roadway Sections and Weather Stations in Washington State

In this research, 30-year average monthly precipitation depth and 30-year average monthly snow depth are selected as key weather variables. The reason why average values are selected is that the monthly precipitation and monthly snow depth in each year are not significantly different from 30-year average values.

Figure 3.3 shows the distribution of median crossover accident counts. The median crossover accident count ranges from 0 to 7 and the zero median crossover accident count is approximately 83 percent (1,139 counts) of median crossover accident dataset.

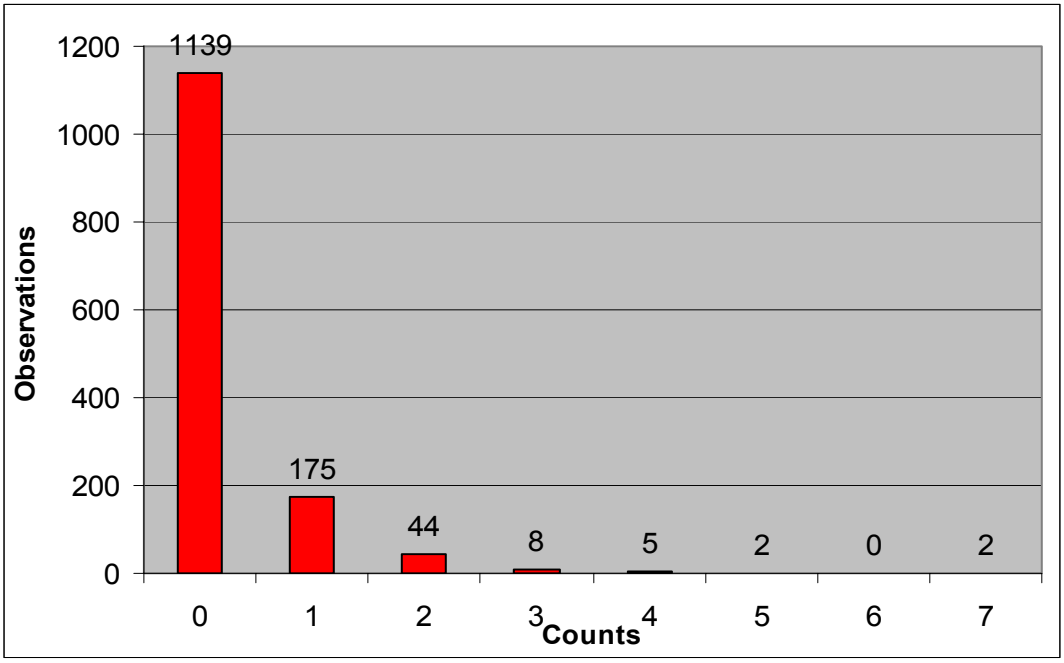


Figure 3.3 The Frequency of Median Crossover Accident Counts in the Dataset

Table 3.1 provides aggregate descriptive statistics of key variables in the median crossover accident dataset for the entire Washington State highway network. The number of median crossover accidents enters into the dataset as dependent variables with traffic variables, roadway geometric variables, median variables, weather variables and interaction variables as explanatory variables.

Table 3.2 presents disaggregate descriptive statistics of major variables in 6 regions defined by WSDOT. This table shows that there are geographical differences (i.e. terrain and weather) among 6 regions. However, this research intends to develop the aggregate median crossover accident models using the data of the entire Washington State highway network.

Table 3.1 Descriptive Statistics of Key Median Crossover Accident Related Variables for the Entire Washington State

Variable	Mean	Std. Error	Min	Max
Dependent Variables				
The number of crossover accidents in section	0.2407	0.6448	0.00	7.00
Traffic Variables				
Average AADT (weighted)	37,354.629	36,974.9664	3,347.00	172,557.00
AADT per lane	7,443.8783	5,828.7465	836.75	28,688.33
Natural logarithm of AADT per lane	8.6306	0.7616	6.73	10.26
Per-lane AADT indicator 1 (1 if per-lane AADT <= 5,000 vehicles, 0 otherwise)	0.4691	0.4992		
Per-lane AADT indicator 2 (1 if per-lane AADT > 5,000 vehicles and <= 10,000 vehicles, 0 otherwise)	0.2873	0.4527		
Per-lane AADT indicator 3 (1 if per-lane AADT > 10,000 vehicles, 0 otherwise)	0.2436	0.4294		
Single truck percentage	4.1960	1.2150	1.90	10.00
Double truck percentage	7.7623	4.6205	0.55	17.80
Truck-train percentage	2.2050	1.5970	0.00	7.00
Total truck percentage	14.1634	6.6821	3.20	32.00
Total truck percentage indicator 1(1 if percentage of total trucks <= 5%, 0 otherwise)	0.0327	0.1780		
Total truck percentage indicator 2 (1 if percentage of total trucks > 5% and < 15%, 0 otherwise)	0.5345	0.4990		
Total truck percentage indicator 3 (1 if percentage of total trucks >=15%, 0 otherwise)	0.4327	0.4956		
Percentage of AADT in the peak hour	11.1158	3.0922	7.30	19.40
Peak hour indicator 1 (1 if percentage of AADT in peak hour <= 9%, 0 otherwise)	0.2691	0.4436		
Peak hour indicator 2 (1 if percentage of AADT in peak hour > 9% and <= 13%, 0 otherwise)	0.5345	0.4990		
Peak hour indicator 3 (1 if percentage of AADT in peak hour > 13%, 0 otherwise)	0.1964	0.3974		
Roadway Geometric Variables				
Collector indicator (1 if the section is collector , 0 otherwise)	0.0073	0.0850		
Principal arterials indicator (1 if the section is principal arterial, 0 otherwise)	0.3055	0.4608		
Interstate indicator (1 if the section is interstate, 0 otherwise)	0.6872	0.4638		
Length of the roadway section in miles	2.4297	2.6899	0.50	19.30
Average number of lanes	4.6036	1.1315	2.00	8.00
Average roadway width	57.4182	15.4683	24.00	121.00
Average speed limit in mph	59.6727	5.5026	35.00	65.00
Speed limit indicator 1 (1 if speed limit < 55 mph, 0 otherwise)	0.0255	0.1576		

Table 3.1 Descriptive Statistics of Key Median Crossover Accident Related Variables for the Entire Washington State (Continued)

Variable	Mean	Std. Error	Min	Max
Speed limit indicator 2 (1 if speed limit >= 55 mph, 0 otherwise)	0.9745	0.1576		
The number of interchanges in section	0.8473	0.8350	0.00	4.00
Minimum horizontal central angle in degrees	13.4311	17.7347	0.00	111.49
Maximum horizontal central angle in degrees	30.2916	23.8828	0.00	111.49
Horizontal curve indicator (1 if the number of horizontal curves >0, 0 otherwise)	0.9091	0.2876		
The number of horizontal curves in section	2.7491	2.8612	0.00	29.00
The number of horizontal curves per mile	1.4400	0.9600	0.00	5.00
The number of horizontal curves per mile indicator 1 (1 if the number of horizontal curves per mile <= 0.5, 0 otherwise)	0.1600	0.3667		
The number of horizontal curves per mile indicator 2 (1 if the number of horizontal curves per mile > 0.5 and <= 2.5, 0 otherwise)	0.7127	0.4527		
The number of horizontal curves per mile indicator 3 (1 if the number of horizontal curves per mile > 2.5, 0 otherwise)	0.1273	0.3334		
The number of grade changes	3.8655	4.0887	0.00	28.00
The number of grade changes per mile	1.8868	1.6935	0.00	20.00
Median Variables				
Minimum median shoulder width in feet	4.4836	1.6833	0.00	10.00
Maximum median shoulder width in feet	5.3055	2.4919	0.00	18.00
Median indicator (1 if median exists in the section, 0 otherwise)	0.9055	0.2927		
Percent medians <= 30 feet	0.0473	0.2123		
Percent medians > 30 feet and <= 40 feet	0.2764	0.4474		
Percent medians > 40 feet and <= 50 feet	0.1164	0.3208		
Percent medians > 50 feet and <= 60 feet	0.0582	0.2342		
Percent medians > 60 feet	0.5018	0.5002		
Weather Variables				
Average monthly precipitation in inches	2.4914	1.8215	0.38	10.98
Average monthly snow depth in inches	1.2625	3.5545	0.00	54.33
Monthly precipitation indicator 1 (1 if average monthly precipitation <= 1.5 inches, 0 otherwise)	0.3891	0.4877		
Monthly precipitation indicator 2 (1 if average monthly precipitation > 1.5 inches and <= 4 inches, 0 otherwise)	0.4356	0.4960		
Monthly precipitation indicator 3 (1 if average monthly precipitation > 4 inches, 0 otherwise)	0.1753	0.3803		
Monthly snow indicator 1 (1 if average monthly snow depth <= 1 inch, 0 otherwise)	0.6975	0.4595		
Monthly snow indicator 2 (1 if average monthly snow depth > 1 inch, 0 otherwise)	0.3025	0.4595		

Table 3.1 Descriptive Statistics of Key Median Crossover Accident Related Variables for the Entire Washington State (Continued)

Variable	Mean	Std. Error	Min	Max
Interaction Variables				
Interaction between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	0.0909	0.2876		
Interaction between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	0.1607	0.3674		
Length of the roadway section on median widths less than or equal to 40 feet	0.6800	1.5024	0.00	12.40
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.2782	0.7845	0.00	5.69
Length of the roadway section on median widths greater than 60 feet	1.4715	2.7502	0.00	19.30

Table 3.2 Descriptive Statistics of Key Median Crossover Accident Related Variables for Each Region in Washington State

Variable	Mean	Std. Error	Min	Max
Northwest Region				
The number of crossover accidents in section	0.3486	0.7598	0.00	7.00
Average AADT (weighted)	75009.0857	47426.5366	19493.00	172557.00
AADT per lane	13220.8109	6594.4235	4478.57	28688.33
Total truck percentage	9.2367	3.6476	3.20	18.20
Length of the roadway section in miles	1.6376	1.2645	0.53	8.20
Average speed limit in mph	58.2857	4.7037	55.00	65.00
Average monthly precipitation in inches	3.1805	1.1731	1.44	10.98
Average monthly snow depth in inches	0.4298	0.7634	0.00	6.75
North Central Region				
The number of crossover accidents in section	0.0857	0.2813	0.00	1.00
Average AADT (weighted)	9305.5714	2629.8591	3347.00	13946.00
AADT per lane	2570.4524	1066.0913	836.75	5604.50
Total truck percentage	17.8538	6.0087	6.13	23.80
Length of the roadway section in miles	2.7452	2.4734	0.51	9.49
Average speed limit in mph	59.7619	7.8257	35.00	65.00
Average monthly precipitation in inches	1.3510	1.8750	0.55	9.94
Average monthly snow depth in inches	4.4963	11.0490	0.00	54.33

Table 3.2 Descriptive Statistics of Key Median Crossover Accident Related Variables for Each Region in Washington State (Continued)

Variable	Mean	Std. Error	Min	Max
Olympic Region				
The number of crossover accidents in section	0.2745	0.6664	0.00	5.00
Average AADT (weighted)	39312.9804	26947.9776	11157.00	163157.00
AADT per lane	9008.1838	4789.9735	2906.50	20394.63
Total truck percentage	9.2331	3.5795	3.20	22.04
Length of the roadway section in miles	1.9833	1.4334	0.50	5.78
Average speed limit in mph	55.2941	2.8774	45.00	65.00
Average monthly precipitation in inches	3.8887	1.2208	0.95	6.74
Average monthly snow depth in inches	0.2499	0.5398	0.00	2.63
Southwest Region				
The number of crossover accidents in section	0.2000	0.4828	0.00	2.00
Average AADT (weighted)	42740.3929	12333.3453	10000.00	79848.00
AADT per lane	7664.8580	1807.8986	2500.00	11814.50
Total truck percentage	19.7614	5.9626	4.50	32.00
Length of the roadway section in miles	1.4900	0.9590	0.52	4.18
Average speed limit in mph	61.4286	6.2699	40.00	65.00
Average monthly precipitation in inches	4.0033	1.1584	2.73	9.27
Average monthly snow depth in inches	0.1405	0.2567	0.00	0.88
South Central Region				
The number of crossover accidents in section	0.1468	0.4259	0.00	3.00
Average AADT (weighted)	15759.3291	7053.0515	3917.00	43014.00
AADT per lane	3780.3719	1683.7351	979.25	10753.50
Total truck percentage	17.9908	5.2692	3.97	27.30
Length of the roadway section in miles	3.0061	3.3063	0.50	18.57
Average speed limit in mph	62.9747	4.0240	55.00	65.00
Average monthly precipitation in inches	1.1679	1.6698	0.38	10.98
Average monthly snow depth in inches	1.6959	2.1683	0.00	9.63
Eastern Region				
The number of crossover accidents in section	0.3385	1.0158	0.00	7.00
Average AADT (weighted)	14607.3077	15792.3142	4805.00	66767.00
AADT per lane	3651.8269	3948.0786	1201.25	16691.75
Total truck percentage	16.4596	7.2779	6.50	27.78
Length of the roadway section in miles	4.4438	4.5982	0.52	19.30
Average speed limit in mph	60.0000	5.0193	55.00	65.00
Average monthly precipitation in inches	1.2092	0.4220	0.61	2.43
Average monthly snow depth in inches	2.7698	1.7081	0.00	6.78

3.2 Descriptive Statistics Discussions

The important details about the descriptive statistics of the variables collected for the 1375 sections are discussed as follows.

Median Crossover Accident Variables

For the entire Washington State, the number of median crossover accident varies from 0 to a high of 7, with a mean of 0.24 as presented in table 3.1. As disaggregate analysis shown in table 3.2, the northwest region, the highest population density region in Washington State, has the highest average median crossover frequencies of 0.35 in the section.

Traffic Variables

From table 3.1, the average AADT ranges from 3,347 to 172,557, with a mean AADT of 37,354.63. From among the traffic variables, it can be seen that the AADT per lane varies from 836.75 to a high of 28,688.33, with a mean of 7,443.88 indicating a high share of roads with high AADT. The natural logarithm of AADT per lane variable was formed to avoid the possible heteroskedasticity due to the high scale of the AADT per lane values. It shows a mean of 8.63 with a minimum and maximum of 6.73 and 10.26 respectively. For per-lane AADT indicators, 645 sections have per-lane AADT less than or equal to 5,000, 395 sections have per-lane AADT ranging from 5,000 to 10,000 and 335 sections have per-lane AADT higher than 10,000.

Truck variables representing as single, double, truck-train and total truck percentages were found to be 4.20, 7.76, 2.21 and 14.16, respectively, on an average. It is interesting to note that the total truck percentage is as high as 32.00 percent in at least one section, while the double truck composition reaches as high as 17.80 percent in some sections. Regarding the percentage of AADT in the peak hour, 11.12 percent is mean of AADT

percentage in peak hour while the highest and the lowest percentage of AADT are 19.40 percent and 7.30 percent respectively.

For traffic statistics in each region, the northwest region has the highest average AADT and average AADT per lane in the section because the northwest region composes of several major cities in Washington State such as Seattle and Bellevue. On the other hand, the north central region shows the lowest average AADT and average AADT per lane in the section because most of roadway sections in this region are located in the mountainous area.

Roadway Geometric Variables

As shown in table 3.1, the sample contains three classes of roads ranging from collectors to interstates. The road type indicators reveal that 68.72 percent of the sections are interstates, 30.55 percent of them are principal arterials, and a very low 0.73 percent of them are collectors.

For roadway geometrics, average length of the roadway sections in miles ranges from 0.50 to 19.30 with a mean length of 2.43. The average number of lanes is 4.60 with a minimum and maximum of 2.00 and 8.00 respectively. From the table, the average roadway width varies from 24.00 to 121.00, with a mean roadway width of 57.42.

The average speed limit for the sections was found to vary from 35 miles per hour to 65 miles per hour, with a mean speed limit of 59.67 miles per hour. The speed limit indicator shows that around 97.45 percent of the sections have posted speed limit greater than and equal to 55 miles per hour. This indicates the presence of a lot of high-speed sections in the database. It is necessarily to note that not all of high speed sections are interstate sections, as shown by the significant presence of principal arterial sections.

On average, 0.84 interchanges were found to be present in a single section. It can also be seen that the central angle in degrees for the horizontal curves varies from 0 to 111.49.

The high mean of 0.9091 for the horizontal curve indicator means that 90.91 percent of the sections have one or more horizontal curves within the section. Also, 2.75 horizontal curves were found to be present in a single section on average.

The per-mile horizontal curve variables reveal that 1.44 horizontal curves were found in a single section. The horizontal curve per mile indicator indicates that around 16.00 percent of the sections are the sections with per-mile horizontal curves less than or equal to 0.5 curves, 71.27 percent of them are the sections with per-mile horizontal curve ranging from 0.5 to 2.5 curves, and 12.73 percent of them are the sections with per-mile horizontal curve greater than 2.5 curves.

Median Variables

Table 3.1 presents that there are a few sections with no median shoulders and sections with an average shoulder width as high as 18 feet. The minimum median shoulder width ranges from 0 to 10 feet while the maximum median shoulder width ranges from 0 to18 feet. There are as high as 1,245 sections out of the 1,375 sections with a median ending or beginning in the section.

For percent medians indicators, 65 sections have median less than or equal to 30 feet in width, 380 sections have median greater than 30 feet and less than or equal to 40 feet in width, 160 sections have median greater than 40 feet and less than or equal to 50 feet in width, 80 section have median greater 50 feet and less than or equal to 60 feet in width, and as high as 690 sections have median wider than 60 feet.

Weather Variables

Turning to the variables depicting the weather characteristics of the sections presented in table 3.1, the average monthly precipitation in inches, which is the mean of maximums and minimums of monthly precipitation within a single section, was found to vary from 0.38 to 10.98 inches, with an average of around 2.49 inches. It can be seen that around

38.91 percent of the sections, which amounts to 535 sections, have an average monthly precipitation less than 0.38 inches. The high monthly precipitation indicator shows that 241 sections have an average monthly precipitation greater than 4 inches. This shows that the rest of the sections, 599 sections, fall into the category of average monthly precipitation between 1.5 and 4 inches.

The variable depicting the average monthly snow depth is also considered. The mean values for this variable was found to be 1.26. The average monthly snow depth ranged from 0.00 to 54.33 inches. It is necessary to note here that the section with 54.33 inches of snow depth is located in the mountainous area. The monthly snow indicators reveal that 69.75 percent of sections have monthly snow depth less than or equal to 1 inch and 30.25 percent of sections have monthly snow depth higher than 1 inch.

As expected, the average monthly precipitation in the section in the southwest region is the highest among 6 regions in Washington State because most of selected roadway sections and weather stations are located between Pacific Ocean and Cascade Mountains. Most of selected roadway sections and weather stations in the north central region are located at high altitude in the mountainous area. As a result, the average monthly snow depth in the north central region is the highest in comparison with those from other regions in Washington State.

Interaction Variables

A few interaction variables that might possibly explain the median crossover accident occurrences were developed. These variables were created from conditioning the different categories of variables, such as traffic variables, roadway geometrics, median variables and weather variables. The discussions of these variables are provided as follow:

The interaction variable between average monthly precipitation and per-mile horizontal curve shows that 9.09 percent of the sections, which amounts to 125 sections, are the

sections with average monthly precipitation less than or equal to 1.5 inches and with per-mile horizontal curve less than or equal to 0.5 curves. It can also be seen that 16.07 percent of the sections, which amounts to 221 sections, are the section with average monthly precipitation greater than 4 inches and with per-mile horizontal curve higher than 0.5 curves.

The interaction variables between length of roadway section and median width show that average length of roadway with median less than or equal to 40 feet in width is 0.68 miles, average length of roadway with median ranging from 41 to 60 feet in width is 0.27 miles, and average length of roadway with median greater than 60 feet in width is 1.47 miles.

Chapter 4
ANALYTICAL APPRROACH

The analytical framework presented here delves directly into the statistical and econometric treatment of crucial modeling issues raised in earlier chapters. The presented framework is in a form designed to delineate for the reader the logic of thought and the sequence of research inquiries that occurred in this dissertation. Figure 4.1 below details the crucial steps in this framework. Following a discussion of these steps, crucial modeling issues such as unobserved effects, correlation between accident counts and excess zeros in accident count model will be described along with state-of-the-art treatments.

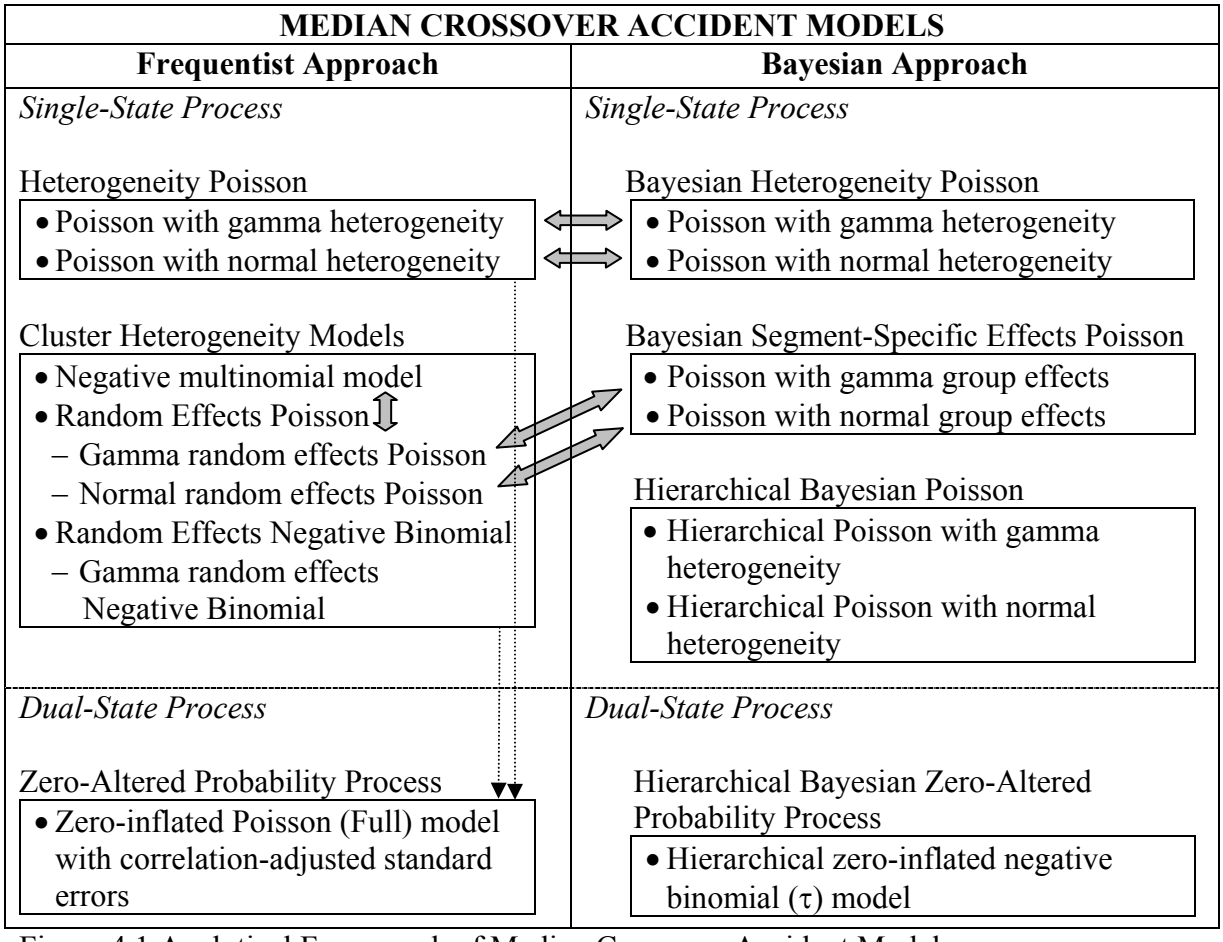


Figure 4.1 Analytical Framework of Median Crossover Accident Models

For ease of discussion, the analytical framework presented in figure 4.1 is separated into frequentist and Bayesian taxonomies as seen in the left and right halves of the figure. Within these taxonomies are single-state and dual-state processes as seen in the top and bottom parts of figure 4.1. A total of 14 different model structures will be examined in response to the above-mentioned taxonomies. They address the most important and commonly recurring modeling issues that arise in the median crossover accidents from different assumptions of models and perspectives of modelers. All models were estimated using the same set of specifications as a benchmark to identify the robustness of the explanatory variables. In other words, this approach allows me to probe for variables that are robust to specification assumptions. By doing so, the intent of this dissertation is to identify "common denominator variables" that would appear to be consistently statistically significant in the various models no matter what modeling issues the estimated models are facing. The following section will describe critical modeling issues along with treatments within the frequentist method followed by a discussion of the Bayesian method.

In frequentist analysis, under the notion of repeated sampling, count data in accident contexts (e.g. zero and positive integer) are often modeled by the Poisson model. The Poisson model is appropriate when the mean and the variance of accident counts are equal. Unobserved heterogeneity that causes a given roadway segment to differ from the average representative segment is a common issue in accident occurrence, leading to the violation of the mean-variance equality assumption of the Poisson model. It should be noted that accidents are modeled at the yearly level. When estimating the probability of accidents at the yearly level, there is no guarantee that the probabilities are evenly distributed within the year. The uneven distribution of probabilities within the interval of observation causes "positive contagion" and along with unobserved heterogeneity, results in an overdispersion problem where variance exceeds the mean (Shankar et al., 1995). The negative binomial (NB) model is suitable for overdispersed accident frequencies, as show in several prior works (see for example Poch and Mannering, 1996; Milton and Mannering, 1996). In general, the gamma distribution is used to capture heterogeneity. Conditioned on the gamma heterogeneity, a Poisson density is first constructed and the

marginal density is derived by integrating out the gamma heterogeneity. Adopting this approach results in the NB model with type II variance-mean relationship, where the variance is a quadratic function of the mean. The alternate approach is to assume a normal heterogeneity effect, whereby a Poisson-normal density is derived. While this appears to capture overdispersion in a quadratic manner, the effect of Poisson-normal overdispersion may be to influence parameter standard errors differently from the gamma-heterogeneity assumption. Regardless of the error assumptions, the upshot of the approaches is they maintain an empirically-validated quadratic variance-mean relationship in accident contexts – on the contrary, one can model the variance-mean relationship as a linear function. However, there is little empirical evidence to support this relationship.

From the standpoint of accurate and reliable forecasts of median crossover frequencies and other accident types associated with medians, multivariate models incorporating the effects of roadway geometrics, median, traffic and weather variables are necessary and also Poisson and NB variant models would be a good starting point for modeling median related accidents (Chayanan et al., 2004 and Shankar et al., 1998).

In this research, multiple years of cross-sectional data on highway accident occurrences including time series information on roadway and roadside geometrics, traffic and weather data are used. For NB models, it is assumed that each observation has individual heterogeneity, however, one may assume identical heterogeneity regardless of time in the same roadway section. Previous research (Hausman et al., 1984) in these matters has explored modeling techniques appropriate for applications to section-specific heterogeneity. Hausman et al., in their seminal 1984 work, explored the applicability of count models as they related to research and development patenting trends in firms. To statistically correlate firm characteristics with longitudinal observations on research and development patent expenditures and returns to investments, the authors suggested the negative binomial and variants as plausible methods in the count data context. The significance of the Hausman et al. work is that it is applicable to modeling overdispersion in count data, while accounting for group-specific effects as well as effects within groups

over time. In this sense, I adapt the framework in Hausman et al. to model overdispersion, section-specific and temporal effects in median crossover data. While the Hausman et al. work was explored in a frequentist context, I adapt the approach to the Bayesian context as well.

To incorporate the random effects into the Poisson model, the Poisson parameter is multiplied with gamma or normally distributed cluster-specific effects. This way, the Poisson parameter becomes a random variable rather than a deterministic function of the exogenous variables. As a result of the random effects assumption, this model is named random effects Poisson (REP) model. A comparison of the results from the negative binomial model and the REP model revealed that the REP model improved the fit to the data tremendously over the NB model, which is a strong indicator for the presence of roadway section-specific effects. Compared to the REP, the NB model underestimated the standard errors of the parameter estimates, therefore overestimating the t-statistic.

In addition to section-specific heterogeneity, the model must handle overdispersion. The overdispersion due to unobserved heterogeneity in the data is allowed by letting each section's Poisson parameter to be randomly distributed, thus estimating the NB model extensions for the data, which results in random effects negative binomial model (RENB). RENB essentially layers a random "location and time" effect on the parent NB by assuming that the overdispersion parameter is randomly distributed across groups. The key advantage of this approach is that the variance-to-mean ratio, which is likely to grow with the expected mean of accidents, is not constrained to be constant across locations, as it is, in the case of the cross-sectional NB model. The RENB model allows for randomly distributed section-specific variation. In the case of accident frequency, it is likely that section-specific effects will be important.

In related research on clusters of data over time, Guo, (1996) considered the negative multinomial model (NM) that accounts for the dependence of the count within a time cluster. Guo started with a conventional Poisson regression model and subjected the multiple counts in the same cluster to a cluster-specific random effect representing the

unobserved effects shared by all the counts of the cluster. A gamma distributed cluster-specific effect in the formulation results in the NM model. It will be shown in a subsequent section that the REP and NM models are equivalent.

Even though the presence of unobserved time-invariant cluster-specific effects allows the REP and NM to be more efficient than NB, there is no strong evidence to show that REP and NM incorporate excess zero accident counts better. It is possible that there are many zero accident counts repeating in a cluster over time and REP and NM specifications may ignore this effect. Thus, zero altered models do still play a significant role in the modeling the median crossovers for the current study. Shankar et al. (1997) suggest that the ZIP structure models are promising and have great flexibility in uncovering processes affecting roadway sections observed with a preponderance of zero accidents. The latent processes that determine the safety behavior of a roadway section are modeled using a suitable count modeling structure that models the accident count probability as a sum of latent and non-latent count probabilities. This flexibility allows highway engineers to better isolate design factors that contribute to accident occurrence and also provides additional insight into variables that determine the relative accident likelihoods of safe versus unsafe roadways. The research by Shankar et al. (1997) revealed that the generic nature of the models offers roadway designers the potential to develop a global family of models for accident frequency prediction that can be embedded in a larger safety management system.

In the dual-state process, the critical modeling issues arise from the combination of cluster-specific effects and excess zero counts. The dual-state process section presents an empirical technique to adjust for correlation effects on standard errors in median crossover accident models. Similar to the classical linear regression model, standard errors in count models are downward biased as observed in empirical analyses on median crossover accidents in this study. To develop a technique for adjusting standard errors upward, the NM specification is used as the base to account for correlation across time. Using the ratio of standard errors from the NM model to the naïve NB model, loading factors which represented the level of inflation in standard errors required to account for

correlation among accident frequencies, are developed. To account for the excess zero problem, a ZIP-Full model of median crossover accident counts is then developed. This excess zero is unique to median crossover accidents, especially due to the fact that only five years of observation are available. Using the loading factors developed on the basis of the NM model, the standard errors of the ZIP-Full model are adjusted to identify key factors affecting median crossover accident frequencies. It is plausible that the empirically adjusted standard errors in the ZIP-Full model are reasonable approximations to efficient estimates that would be theoretically derived.

Bayesian estimation is another econometric technique that is being increasingly used in the field of highway safety. Bayesian analysis produces a density function for the desired information rather than a point estimate given by frequentist or classical methods. The main difference between the Bayesian and frequentist methods is how probability is used. The Bayesian theory views probability as a confirmation of beliefs. Therefore, a numerical probability is the confidence the researcher has in various parameter values. However, the frequentist methods regard probability as the frequency with which an event would appear in repeated sampling.

In the Bayesian domain as shown on the right of figure 4.1., the first model examined is the Bayesian heterogeneity Poisson that accounts for individual heterogeneity of each observation. The heterogeneity structure of Bayesian heterogeneity Poisson model is the same as that in NB model; therefore the estimated parameters and standard errors in both models are comparable. To compare the results of REP, the Bayesian segment-specific effects Poisson was developed on the basis of segment-specific heterogeneity. Similar to the base Bayesian model, Bayesian segment-specific effects Poisson model also incorporate gamma and normally distributed heterogeneity. The last model in the single state process employing Bayesian analysis is the Hierarchical Bayesian Poisson model. Hierarchical Bayesian Poisson is the Bayesian heterogeneity Poisson model except that added layers address uncertainty in the mean and variance of the prior hierarchically. The mean rate of accident frequencies is normally distributed with a mean defined by a vector of explanatory variables and a gamma distributed variance. By doing so, the

structure of Bayesian Poisson become a hierarchy from the product of individual heterogeneity and normal function of the exogenous variables. A multilevel model with carefully chosen priors at various levels for the mean and variance is expected to be more flexible across observations and capture more heterogeneity in the data; hence the hierarchical Bayesian Poisson model would potentially provide improved predictions. The dual-state Bayesian structures involve the incorporation of a splitting regime to account for the possibility of excess zeros in the distribution of counts. Akin to the frequentist case, the splitting regime can be modeled as a logistic function, with the regular count portion being modeled as a negative binomial distribution. One variation in the unobserved portion of the logistic function is the inclusion of normal priors.

The Bayesian approach has one final and major advantage – treatment of distributional assumptions of what the parameters should follow. Through the use of priors, from uniform uninformative priors to more informative priors, subjective prior beliefs about parameters can be tested with the given dataset at hand. In this sense, the Bayesian approach can also be viewed as a method to identify robustness – i.e., what parameters are significant and provide credible estimates under a variety of prior distributions. Variance distributions here greatly improve the model's ability to capture heterogeneity across time and space. For example, one can adopt a two-layered variance prior for capturing heterogeneity in both time and space, with each layer specifically addressing group-specific effects in either time or space. How one characterizes the variance priors matters as well, in addition to the distributional assumptions. If a vector of explanatory variables is shown to be truly related to heterogeneity in either time or space, it would appear then certain roadway geometrics and environmental effects are correlated with unobserved effects in time and space.

To show the promise for improvements in frequency predictions using Bayesian analysis, prediction tests are performed after all models in both approaches are fully developed. Two measures of predictive effectiveness, mean absolute deviation (MAD) and root mean square error (RMSE), are suggested for count data. Percent change of predicted count from observed is not appropriate due to the significant presence of zero counts. In

this research, extra-sample forecast error and within-sample forecast error were tested to ensure the consistency of prediction obtained from classical frequentist approach and Bayesian approach.

Chapter 5
MODELING STRUCTURES

This chapter presents the mathematical derivations of count models used in this dissertation. Also, constraints and proficiencies of each model are discussed briefly.

5.1 Heterogeneity Poisson Models of Median Crossover Accident Frequency

The derivation of the heterogeneity Poisson or Negative Binomial (NB) model begins with the conditionally Poisson probability density function, which is an expression for the probability of the frequency equaling a particular count.

$$f(y_{it} \mid \theta_{it}) = \frac{\exp(\theta_{it}) * \theta^{y_{it}}}{y_{it}!}; \qquad y_{it} = 0,1,2,... \tag{5.1}$$

where y_{it} is the observed frequency of median crossover accidents in roadway section *i* at time *t*, and θ_{it} is the parameter of Possion function. To address potential overdispersion in the data (i.e. the variance being significantly greater than the mean), the NB model introduces an error term, ε_{it}, to relax the Poisson's mean-variance equality constraint. Suppose that the parameter θ_{it} has a random term that enters the conditional mean function multiplicatively, that is,

$$\theta_{it}=\exp(\beta_0+\mathbf{X}_{it}\beta+\varepsilon_{it}) \tag{5.2}$$

$$\theta_{it}=\exp(\beta_0+\mathbf{X}_{it}\beta)*\exp(\varepsilon_{it}) \tag{5.3}$$

$$\theta_{it}=\mu_{it}\nu_{it} \tag{5.4}$$

where $\mu_{it} = \exp(\beta_0+\mathbf{X}_{it}\beta)$ and $\nu_{it} = \exp(\varepsilon_i)$. In these equations, $\mathbf{X}_{it}$ is a vector of geometric, traffic and weather variables for roadway section *i* at time *t* and β is a vector of estimable coefficients. This model can be interpreted as a Poisson model with gamma distributed heterogeneity as a result of the gamma distribution assumption in the error term. The marginal distribution of y_{it} is obtained by integrating out ν_{it},

$$h(y_{it} \mid \mu_{it}) = \int_0^{\infty} f[y_{it} \mid \mu_{it}, \upsilon_{it})] * g(\upsilon_{it}) d\upsilon_{it} \tag{5.5}$$

where, $g(\nu_{it})$ is a mixing distribution. Suppose that the variable ν_{it} has a two-parameter gamma distribution $g(\nu_{it}, \delta, \phi)$

$$g(\mu, \upsilon, \phi) = \frac{\delta^{\phi}}{\Gamma(\delta)} \upsilon^{\delta-1} e^{-\upsilon\phi}, \delta > 0, \phi > 0 \tag{5.6}$$

where $E[\nu]=\delta/\phi$, and $V[\nu]=\delta/\phi^2$. The intercept identification condition is $E[\nu]=1$ which is obtained by setting $\delta=\phi$, which implies a one-parameter gamma family with $V[\nu]=1/\delta=\alpha$

Following the marginal density and computing the first and second moments, I obtain $E[y_{it}] = \exp(\mathbf{X}_{it}\beta)$ and $Var[y_{it}] = E[y_{it}][1 + \alpha E[y_{it}]]$. If α is significantly greater than zero, the accident data is overdispersed and hence the NB model will be hold. It should be noted here that NB distribution will decompose to a Poisson distribution if α is equal to zero. With the normalization of disturbance to identify the mean of the NB distribution (Cameron and Trivedi, 1998), the probability distribution of NB assumption yields closed-form as,

$$P(y_{it}) = \frac{\Gamma(\theta + y_{it})}{\Gamma(y_{it} + 1)\Gamma(\theta)} r_{it}^{y_{it}} (1 - r_{it})^{\theta} \tag{5.7}$$

where $r_{it} = \lambda_{it}/(\lambda_{it} + \theta)$, $\theta = 1/\alpha$, and $\Gamma(.)$ is a gamma function. The estimable parameters, β and the dispersion parameter, α, are estimated through a maximum likelihood function in equation 5.8, see, for example, (Greene, 2003). The objective of this function is to obtain a set of parameters maximizing the joint density or likelihood function over all individuals. The likelihood function to be maximized can be written as,

$$L(\beta,\theta)=\prod_{i=1}^{N}\prod_{t=1}^{T}\frac{\Gamma(\theta+y_{it})}{\Gamma(y_{it}+1)\Gamma(\theta)}r_{it}^{y_{it}}(1-r_{it})^{\theta} \tag{5.8}$$

where N is the total number of sections (275 sections) and T is the total number of years (5 years).

In many otherwise appealing models a closed-form marginal density may not be generated. It is still possible to estimate the parameters of the mixture distribution using computer-intensive methods such as simulated maximum likelihood or simulated method of moments (Gourieroux and Monfort, 1997). Suppose the distribution of a random median crossover accident frequency is conditionally Poisson with normally distributed heterogeneity, $\upsilon_{it} \sim N[0, 1]$, as show in equation 5.9

$$f(y_{it} \mid \mu_{it}, \upsilon_{it}) \sim P[y_i \mid \exp(x_{it}\beta + \sigma\upsilon_{it})] \tag{5.9}$$

The marginal distribution of median crossover accident frequency is obtained by integrating out υ_{it},

$$h(y_{it} \mid \mu_{it}) = \int_{-\infty}^{\infty} f[y_{it} \mid \exp(x_{it}\beta + \sigma\upsilon_{it})] * \Phi(\upsilon_{it}) d\upsilon_{it} \tag{5.10}$$

As shown in equation 5.10, $\Phi(\upsilon_{it})$ follows standard normal density and as a result, there is no closed-form solution for $h(y_{it} \mid \mu_{it})$. An alternative approach is to use simulation. The expression can be approximated by replacing the n-element vector υ by draw from N[0, 1] distribution because the marginal distribution is a mathematical expectation. Then

$$h(y_{it} \mid \mu_{it}) \approx \frac{1}{S}\sum_{s=1}^{S} f[y_{it} \mid \exp(x_{it}\beta + \sigma\upsilon_{it}^{(S)})] \tag{5.11}$$

In case of the absence of observable heterogeneity, a pseudorandom number generator was drawn, S draws, to approximate the term. Simulated maximum likelihood estimates are obtained by maximizing the log-likelihood function:

$$L(\beta,\sigma) = \sum_{s=1}^{N} \ln \frac{1}{S} \sum_{s=1}^{S} f[y_{it} \mid \exp(x_{it}\beta + \sigma\upsilon_{it}^{(S)})] \tag{5.12}$$

The alternative approach based on numerical integration for a Poisson-normal model was used by Hinde, 1982. The integration is approximated by an H-point Gaussian quadrature with quadrature points (nodes) υ_h and weights w_h,

$$h(y_{it} \mid \mu_{it}) \approx \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} w_h [f[y_{it} \mid \mu_{it}, \upsilon_h]] \tag{5.13}$$

With the above-derived marginal density, one can obtain through first and second moments, $E[y_{it}] = \exp(\mathbf{X}_{it}\beta) * \exp(0.5\sigma^2)$ and $Var[y_{it}] = E[y_{it}][1 + [\exp(\sigma^2)-1]E[y_{it}]]$. The dispersion parameter, α, is equivalent to $[\exp(\sigma^2)-1]$ in the variance function and this model will return to a model without heterogeneity or Poisson model if σ is set to zero.

5.2 Cluster Heterogeneity Poisson Models of Median Crossover Accident Frequency

The random effects Poisson (REP) model proposed by Hausman et al. (1984) is the familiar model to accommodate heterogeneity in panel count data. To account for the section-specific variation, REP is preceded as is done for the NB model. A random error term specific to the segment is added to the expression for the mean,

$$\ln \lambda_{it} = \mathbf{X}_{it}\beta + \varepsilon_i \quad ; i=1,\ldots, N; t=1,\ldots,T_i \tag{5.14}$$

where λ_{it} is the mean rate of median crossover accidents for roadway section *i* at time *t*, ε_i is a section-specific (not observation-specific as in the NB model) random error term for the i^{th} section, $\exp(\varepsilon_i)$ is assumed to be independently and identically distributed gamma

distribution with mean 1 and variance $\alpha=1/\theta$. The assumption of mean 1 does not cause loss of generality if equation 5.14 includes an intercept term.

The conditional joint density function of all individual counts for a particular section *i*, given that the individual counts are distributed by equation 5.14 and conditioned on $\exp(\varepsilon_i)$, can now be written as:

$$P(y_{i1}, y_{i2}..., y_{iT} \mid \exp(\varepsilon_i)) = \prod_t p(y_{it} \mid \exp(\varepsilon_i)) \tag{5.15}$$

where t_i denotes the number of time periods observed for section *i* (5 years). This assumes the accident counts in different sections are independent. This is reasonable because these sections are generally not proximate to each other and will therefore only share minimal unobserved effects. The unconditional joint density function for the REP can now be derived by integrating equation 5.15 and by using the assumed gamma distribution of $\exp(\varepsilon_i)$ to give:

$$P(y_{i1}, y_{i2},..., y_{iT}) = \frac{(\prod_t \lambda_{it}^{y_{it}})\Gamma(\theta + \sum_t y_{it})}{(\prod_t y_{it}!)\Gamma(\theta)(\prod_t y_{it})!(\prod_t \lambda_{it})^{\sum_t y_{it}}} u_i^{\theta}(1-u_i)^{\sum_t y_{it}} \tag{5.16}$$

where $\Gamma(.)$ is a value of gamma function and $\lambda_{it}=\exp(\mathbf{X}_{it}\beta)$. Recall that the variance of $\exp(\varepsilon_i)$ is $\alpha=1/\theta$. The degenerate case, when each section has only one observation (i.e. there is no section-specific correlation), yields the NB distribution. The expected value and the variance for REP model are $E[y_{it}] = \exp(\mathbf{X}_{it}\beta)$, $Var[y_{it}] = E[y_{it}][1 + \alpha E[y_{it}]]$ respectively. To estimate estimable parameters, β and dispersion parameter, α, the maximum likelihood approach is applied as presented in equation 5.17

$$L(\beta, \theta) = \prod_{i=1}^{N} \frac{(\prod_t \lambda_{it}^{y_{it}})\Gamma(\theta + \sum_t y_{it})}{(\prod_t y_{it}!)\Gamma(\theta)(\prod_t y_{it})!(\prod_t \lambda_{it})^{\sum_t y_{it}}} u_i^{\theta}(1-u_i)^{\sum_t y_{it}} \tag{5.17}$$

where N is the total number of sections (275 sections). It is noted here that the gamma random effects may be substituted by normal random effects where $\varepsilon_I \sim N[0, \sigma^2]$.

An alternate way to capture cluster heterogeneity in the Poisson distribution is through the negative multinomial (NM) model (Guo, 1996). This model assumes that the unobserved effect in a roadway section is fixed across years whereas the base NB model does not, treating a single roadway section repeated across years as unrelated roadway sections. The NM model suggested by Guo (1996) yields the same estimated parameters and standard errors as the REP model.

For the random effects negative binomial (RENB) model, Hausman et al., 1984 proposed the following approach: the base model is REP shown in equation 5.14. The random term, ε_i is distributed as gamma distribution with parameters (θ_i, θ_i), which produces the negative binomial model with a parameter that varies across groups. Then it is assumed that $\theta_i/(1+\theta_i)$ is distributed as beta (a_n,b_n) which layers the random group effect onto the NB model. The resulting density model is

$$P(y_{i1}, y_{i2}, ..., y_{iT}) = \frac{\Gamma(a_n + b_n)\Gamma(a_n + \sum_{t=1}^{T_i} \lambda_{it})\Gamma(b_n + \sum_{t=1}^{T_i} y_{it})}{\Gamma(a_n)\Gamma(b_n)\Gamma(a_n + b_n + \sum_{t=1}^{T_i} \lambda_{it} + \sum_{t=1}^{T_i} y_{it})} \prod_{t=1}^{T_i} \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \quad (5.18)$$

5.3 Zero-Inflated Poisson Models of Median Crossover Accident Frequency

Let "Z" represents the zero-crash count state of the median crossover accident site, and "Y^*" denote the crash count state for that roadway section. Neither "Z" nor "Y^*" is observed, but only the observed median crossover count "Y" is, such that $Y=Z*Y^*$. Determining the latent components can then be viewed as a mixing distribution problem, with "Z" being modeled as a dichotomous outcome probability and "Y^*" modeled as a count probability.

In vehicular accident contexts, such distributions have been found to be appropriate (Shankar et al., 1997). In particular these studies have highlighted the importance of roadway design deviations as a motivator for partial observability effects. The effect of such deviations has been found to, at the least, cause partial observability, and in certain design situations, overdispersion as well. In the median crossover accident context, design deviations are a significant issue. To highlight this issue, a brief discussion of median barrier warrants is necessary. Median barrier warrants in Washington State, and for that matter in the entire United States, are based on two simplifying factors, namely, average daily traffic (ADT) and median width. No account of interactions between traffic volumes and geometrics, or weather and geometrics is considered in the decision to install a median barrier. Hence, potential interactions arising due to the presence of deviations from preferred geometric design are ignored, leading to the potential for the excess zero median crossover problem. To formally address the excess zero problem, let Y_i be the annual number of accident counts reported for section *i*, and let p_i be the probability that section *i* will exist in the zero-count state over its lifetime. Thus 1- p_i is the probability that section *i* will operate in the non-zero accident state. It is to be noted that in the non-zero count state, the probability of zero accident counts still exists. For my immediate purposes, I assume that this count state follows a Poisson distribution. Given this,

$$Y_i = 0 \text{ with probability } p_i + (1 - p_i)e^{-\lambda_i} \tag{5.19}$$

and,

$$Y_i = k \text{ with probability } (1 - p_i)\left(\frac{e^{-\lambda i}\lambda_i^{k_i}}{k!}\right) \tag{5.20}$$

In equation 5.19 and equation 5.20, the probability of being in the zero-accident state p_i is formulated as a logistic distribution such that $\log(p_i/(1- p_i)) = \mathbf{G_i}\,\boldsymbol{\gamma}$ and λ_i is defined by $\log(\lambda_i) = \mathbf{H_i}\,\boldsymbol{\beta}$, where $\mathbf{G}_i$ and $\mathbf{H}_i$ are covariate vectors, and γ and β are coefficient vectors. The covariates that affect the mean λ_i of the Poisson state may or may not be the same as the covariates that affect the zero-accident state probability (i.e., p_i). Alternatively, vectors $\mathbf{G}_i$ and $\mathbf{H}_i$ may be related to each other by a single parameter τ. This is essentially

a parametric constraint in the sense that the explanatory variables are forced to be the same in both the zero and count states. In such a case, a natural parameterization is $\log(p_i/(1- p_i) = \boldsymbol{\tau}\mathbf{H_i}\ \boldsymbol{\beta}$. It is useful in the accident context to begin with unconstrained parameter vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. This allows different effects to be correlated with the zero state and count state respectively. For example, in the median crossover accident context, it may be useful to consider the impact of median widths alone in the specification of the zero-state probability function, while including design, traffic and environmental interactions as well in the count state function.

Equations 5.19 and 5.20 combined provide the zero-inflated Poisson (ZIP) model. I refer to a model of unconstrained parametric vectors such as the one discussed above as the ZIP-Full model. The maximum likelihood estimation using the gradient/line search approach proposed by Greene (2004) is performed to estimate parameters, β and γ. A likelihood function is given by equation 5.21:

$$L(\beta, \gamma,) = \prod_{i=1}^{n} \left[(1 - p_i) \frac{e^{-\lambda_i} \lambda_i^{y_t}}{y_i!} + Z_i p_i \right] \tag{5.21}$$

where $Z_i = 1$ when $Y_i =$ is observed to be zero and 0 otherwise. An alternate method to estimate ZIP parameters is to employ the expectations maximization technique.

5.3.1 Statistical Validation of the Zero-Inflated Poisson Models of Median Crossover Accident Frequency

In statistically validating the ZIP model, one has to distinguish the base count model (such as the Poisson model for median crossover counts) from the zero-inflated probability model (such as the ZIP). The reasoning behind this test is that the Poisson model does not adequately capture the entire "zero" density, and that the ZIP structure is more reliable. A statistical test for this has been proposed by Vuong (1989). The Vuong test is a t-statistic-based test with reasonable power in count-data applications (Green, 1994). The Vuong statistic (V-statistic) is computed as:

$$V = \frac{\overline{m}\sqrt{N}}{S_m} \tag{5.22}$$

where $\overline{m}$ is the mean with m = log [$f_1(.)$ /$f_2(.)$], (with $f_1(.)$ being the density function of the ZIP distribution and with $f_2(.)$ being the density function of the parent-Poisson distribution), and S_m and N are the standard deviation and sample size respectively.

The advantage of using the Vuong test is that the entire distribution is used for comparison of the means, as opposed to just the excess zero mass. A value greater than 1.96 (the 95 percent confidence level for the t-test) for the V-statistic favors the ZIP-Full model while a value less than -1.96 favors the parent-Poisson (values in between 1.96 and -1.96 mean that the test is indecisive). The intuitive reasoning behind this test is that if the processes are statistically not different, the mean ratio of their densities should equal one. To carry out the test, both the parent and zero-inflated distributions need to be estimated and tested using a t-statistic. Studies (Greene, 1994) have shown that a Vuong statistic has reasonable power and hence is quite reliable. Greene indicates that a significant Vuong statistic in favor of the ZIP model will also favor the ZIP model over the NB model, which would otherwise be the default model for overdispersion from excess zeros.

5.3.2 The Relevance of the Negative Multinomial and Zero-Inflated Poisson Model to Median Crossover Accidents

It is important to keep in mind that the estimated parameters from the ZIP-Full model would be efficient if heterogeneity and dependence of event counts did not exist. Such is not the case in median crossover contexts where the clusters with multiple years of observations are available. In fact, as in the classical linear regression model, one would expect standard errors to be downward biased if correlation of event counts is not accounted for. To correct the variances in a dual-state distribution (such as the ZIP) with correlation of dependent variables, the NM model offers an intuitive approach. The NM captures correlations of accident counts across years without biasing the parameters of the NB distribution. Empirically, this characteristic is of some use in current ZIP-Full

model context. First, as illustrated in figure 5.1, standard errors from a NM model were benchmarked against a naïve NB model first, when repeated observations are the primary source of correlation. The point of importance to note here is that the NM model adjusts the standard errors upward to account for correlation from event counts. This upward adjustment is called the "loading factor." In other words, the standard errors of the NB model are "factored up" by "loading factors" derived from the NM model. As an example, consider the vectors of parameters estimated as "**β**" by a NB model. It is assumed here that **β** has dimensionality k. Consider the NM model vector of parameters $\widetilde{\boldsymbol{\beta}}$ also of dimensionality k, with exactly the same set of regressors as those used for **β**. Then, the loading factor for **β** is a vector μ of with dimensionality k, where $\mu_k = \text{s.e.}(\widetilde{\beta}_k)/\text{s.e.}(\beta_k)$. The "loading factor" for any given parameter β_k varies depending on the effect of correlation in the variance-covariance matrix. The loading factors are then used to adjust upward the standard errors obtained for a naïve ZIP-Full model. The adjusted standard errors are then used to determine parameter significance under correlation of accident counts.

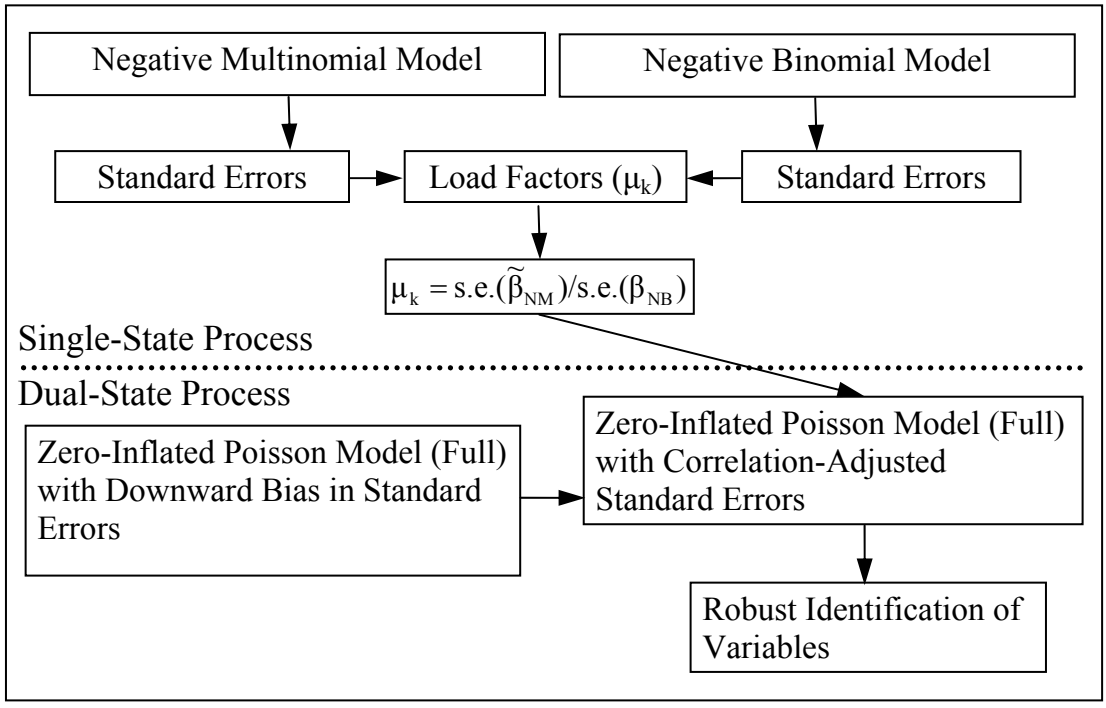


Figure 5.1 Empirical Adjustment for Correlation of Event Counts in Zero-Inflated Poisson (Full) Model

5.4 Bayesian Analysis of Median Crossover Accident Frequency

In classical inference, the sample data y is considered as random whereas the population parameters θ are taken as fixed. In Bayesian analysis, population parameters themselves construct the distributions. Bayesian methods explicitly use probability to quantify uncertainty. After fitting a probability model to data, Bayesian inference summarizes the result by a probability distribution on unobserved quantities such as the parameters of the model, or predictions for new observations. These probability statements condition on the observed value of the data. For more information on Bayesian approaches, the reader should consult Congdon (2003, 2005) and Lancaster (2004) for example.

Bayesian econometrics is the extension of elementary probability, via Bayes' theorem. The posterior distribution combines the likelihood and the prior distributions as well as the marginal distribution, and reflects knowledge about θ being updated after seeing the data. Bayesian inference is expressed as

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)} \tag{5.23}$$

where the term on the left, P(θ|y), is the posterior distribution; the numerator on the right contains the likelihood, P(y|θ), and the prior distribution of parameters P(θ). The denominator on the right, P(y), is called the marginal distribution. It is noted here that P(y) does not involve θ and for the purpose of inference, P(y) can be ignored. Bayesian inference can be often seen as a proportionality relationship defined by:

$$P(\theta|y) \propto P(y|\theta) * P(\theta) \tag{5.24}$$

From the Bayesian perspective, a likelihood, P(y|θ), is the expression for the distribution of the data to be observed given the set of parameters whereas a prior distribution, P(θ), for the parameters reflects knowledge about θ before seeing the data (i.e. on the

subjective view, the priors represent beliefs about θ in the form of a probability distribution).

5.4.1 Bayesian Predictive Distribution

A future observation $\tilde{y}$ may be predicted using a predictive distribution $P(\tilde{y} \mid y)$ based on the posterior distribution as follows:

$$P(\tilde{y} \mid y) = \int P(\tilde{y} \mid \theta) P(\theta \mid y) d\theta \tag{5.25}$$

Importantly, this accounts for the uncertainty in estimating θ. The advantage of Bayesian approaches is directly quantifying uncertainty when predictive models with many parameters are considered. This is especially the case in median crossover accidents where roadway geometrics, traffic factors, median characteristics and environmental conditions play a dominant role. Human factors are neither easily measurable nor maintainable over the long-term as self-sustainable sources of information. Hence, to develop a predictive paradigm that maximizes predictive power, the Bayesian paradigm gives freedom to set up complex models by supplying a conceptually simple method for coping with multiple parameters. Although a realistic model may require many parameters, interest usually focuses on a smaller number of parameters.

5.4.2 Bayesian Heterogeneity Poisson and Bayesian Segment-Specific Effects Poisson Models

Based on Hausman et al., 1984, the Bayesian heterogeneity Poisson specifies the distributions of the observed median crossover accident count vector "**y**" given the individual parameters λ_{it}'s as:

$$\mathbf{y_{it}} | \boldsymbol{\lambda}_{\mathbf{it}} \sim \text{Poisson}(\boldsymbol{\nu}_{\mathbf{it}} \boldsymbol{\lambda}_{\mathbf{it}}) \tag{5.26}$$

where the vector $\boldsymbol{\lambda_{it}}$ is the mean rate of median crossovers in section *i* at time *t*; $\boldsymbol{\lambda_{it}}$ is defined as $\log(\boldsymbol{\lambda_{it}}) = \mathbf{X_{it}}\,\boldsymbol{\beta}$ ($\mathbf{X_{it}}$ is vector of explanatory variables e.g. roadway geometrics, median characteristics and traffic as well as weather variables); $\boldsymbol{\beta}$ is a coefficient vector with a normal prior (mean 0 and variance 0.001) and $\boldsymbol{\nu_{it}}$ is a multiplicative conjugate gamma form for heterogeneity. In this model, the individual Poisson parameters $\boldsymbol{\nu_{it}}$ follow a gamma distribution as shown:

$$\boldsymbol{\nu_{it}} \sim \text{Gamma}(\alpha, \alpha) \tag{5.27}$$

$$\alpha \sim \text{Gamma}(1, 1) \tag{5.28}$$

where α is the overdispersion parameter and is gamma distributed. For this specification, the mean and the variance correspond to the classical NB form with $E[y_{it}] = \exp(\mathbf{X}_{it}\,\boldsymbol{\beta})$ and $\text{Var}[y_{it}] = E[y_{it}][1 + \alpha E[y_{it}]]$. Estimation of the Bayesian model is conducted using a Markov Chain Monte Carlo (MCMC) available in the WinBugs14 program. MCMC is done with 10,000 iterations and over three chains.

Alternatively, the individual model can be formed as an additive (log-normal) form to capture unobserved effects across roadway sections and time. The formulation of the additive form is presented in the following equations:

$$\mathbf{y_{it}}|\boldsymbol{\lambda_{it}} \sim \text{Poisson}(\boldsymbol{\lambda_{it}}) \tag{5.29}$$

$$\log(\boldsymbol{\lambda_{it}}) = \mathbf{X_{it}}\,\boldsymbol{\beta} + \boldsymbol{\varpi_{it}} \tag{5.30}$$

where vector $\boldsymbol{\varpi_{it}}$ is additive normal heterogeneity. The $\boldsymbol{\varpi_{it}}$ is analogous to $\log(\boldsymbol{\nu_{it}})$ of the multiplicative conjugate gamma Poisson. Unlike the gamma distribution, the log-normal is not a conjugate, but remains a popular alternative. The normal heterogeneity is assumed to be distributed with mean 0 and variance σ^2 as follows:

$$\boldsymbol{\varpi_{it}} \sim \text{Normal}(0, \sigma^2) \tag{5.31}$$

The above mentioned structures introduce segment-specific effects through Bayesian analysis.

5.4.3 Bayesian Hierarchical Poisson Models

The greatest practical advantage in Bayesian modeling is the accuracy of the prediction using historical data and appropriate priors. An error term, $\boldsymbol{\nu_{it}}$ and $\boldsymbol{\varpi_{it}}$, is expected to capture the unobserved effects according to non-hierarchical structures described in the previous section. However, in Bayesian hierarchical Poisson models, hierarchical structures are formulated based on the two structures of heterogeneity Poisson in section 5.4.2. Equations 5.32 to 5.36 present the hierarchical Poisson with gamma distributed heterogeneity while the hierarchical Poisson with normally distributed heterogeneity is described in equations 5.37 to 5.41.

$$\mathbf{y_{it}}|\boldsymbol{\lambda_{it}} \sim \text{Poisson}(\exp(\boldsymbol{\nu_{it}}\boldsymbol{\lambda_{it}})) \quad (5.32)$$

$$\boldsymbol{\nu_{it}} \sim \text{Gamma}(\alpha,\alpha) \quad (5.33)$$

$$\alpha \sim \text{Gamma}(1,1) \quad (5.34)$$

$$\boldsymbol{\lambda_{it}} = \text{N}(\mathbf{X_{it}}\,\boldsymbol{\beta}, \sigma^2) \quad (5.35)$$

$$\sigma^2 \sim \text{Gamma}(0.01,0.01) \quad (5.36)$$

$$\mathbf{y_{it}}|\boldsymbol{\lambda_{it}} \sim \text{Poisson}(\exp(\exp(\boldsymbol{\varpi_{it}})\boldsymbol{\lambda_{it}})) \quad (5.37)$$

$$\boldsymbol{\varpi_{it}} \sim \text{Normal}(0, \sigma_0^2) \quad (5.38)$$

$$\sigma 0^2 \sim \text{Gamma}(0.01,0.01) \quad (5.39)$$

$$\boldsymbol{\lambda_{it}} = \text{N}(\mathbf{X_{it}}\,\boldsymbol{\beta}, \sigma^2) \quad (5.40)$$

$$\sigma^2 \sim \text{Gamma}(0.01,0.01) \quad (5.41)$$

The mean rate vector $\boldsymbol{\lambda_{it}}$ is normally distributed with mean $\mathbf{X_{it}}\,\boldsymbol{\beta}$ and with varying precision σ^2 in both structures. With this hierarchy, the mean rate of accident counts $\boldsymbol{\lambda_{it}}$ is distributed normally with a mean defined by a vector of explanatory variables and

estimable parameters and precision parameter σ^2. Hence it is expected to arrive at a more reliable mean rate; leading to more accurate predictions.

5.4.4 Hierarchical Bayesian Zero-Inflated Negative Binomial Model

Median crossover accidents are rare events when viewed at the annual level. Compared to the commonly occurring accident types such as rear-end or sideswipe accidents, the frequency distribution of median crossover accidents contains zeros in excees of what a conventional Poisson model can incorporate. Traditional applications of single-state process models such as Bayesian heterogeneity Poisson and Bayesian segment-specific effects Poisson cannot account for this excess zero density. A Hierarchical Bayesian zero-inflated negative binomial model is promising due to its ability to add "zero density" through a splitting regime. Furthermore, it has great flexibility in uncovering the process affecting median crossover frequencies. The fundamental advantage of the zero-inflated hierarchical Bayesian model would be in its ability to offer better predictions in terms of "safe" versus "unsafe" median crossover roadway sections. The Hierarchical Bayesian zero-inflated negative binomial model is formulated below.

$$\ell(\beta; y, \lambda, p) \alpha \prod_{i=1}^{n} \prod_{t=1}^{T} \left[(1 - p_i t) \left(\frac{e^{-\lambda_{it}} \lambda_{it}^{y_{it}}}{y_{it}!} \right) + z_{it} p_{it} \right] \quad (5.42)$$

$\mathbf{y_{it}}|\boldsymbol{\lambda_{it}} \sim \text{Poisson}(\boldsymbol{\nu_{it}}\boldsymbol{\lambda_{it}})$ (5.43)

$\boldsymbol{\nu_{it}} \sim \text{Gamma}(\alpha,\alpha)$ (5.44)

$\alpha \sim \text{Gamma}(1,1)$ (5.45)

$\log(\mathbf{p_{it}}/(1-\mathbf{p_{it}})) = \tau^*(\mathbf{X_{it}}\boldsymbol{\beta}+\boldsymbol{\varpi_{it}})$ (5.46)

$\boldsymbol{\varpi_{it}} \sim \text{Normal}(0, \sigma^2)$ (5.47)

$\tau \sim \text{Normal}(0, 0.001)$ (5.48)

Equation 5.42 represents the likelihood density relationship for the hierarchical Bayesian zero-inflated negative binomial model. In the equation, the variable z_i is an indicator

variable used to combine the likelihoods from the splitting regime with the regular Poisson or count regime. A value of one for z_i would indicate that the location in question was observed to be in the zero accident state, while a value of zero would indicate a non-zero observed accident count. A non-zero accident outcome does not rule out the "zero probability" portion of the density function however. The regular Poisson density shown in the first part of equation 5.42 accounts for this component. Equations 5.43 to 5.45 show model structures for the non-zero accident probability state. In this state, accident frequencies are modeled as a negative binomial model or a Poisson model with gamma distributed heterogeneity described in equation 5.44. It is also noted here that $\log(\boldsymbol{\lambda_{it}}) = \mathbf{X_{it}}\,\boldsymbol{\beta}$ where $\mathbf{X_{it}}$ is vector of explanatory variables and $\boldsymbol{\beta}$ is a coefficient vector with a normal prior (mean 0 and variance 0.001).

Equations 5.46 to 5.48 provide model structures of the zero accident probability state as a logistic function that relates the odds-ratio of the zero accident state to a vector of variables used to describe the non-zero accident state. The relationship is developed using a scalar parameter named τ. As shown in the equation, the scalar influences the unobserved effect as well. The unobserved effects in the logistic function are captured by an additive normal prior $\boldsymbol{\varpi_{it}}$ with the mean of zero and variance of σ^2. Equation 5.48 represents the prior density of the tau (τ) parameter and this parameter is defined as a normal prior with mean zero and with varying precision 0.001. Whether τ is multiplicative with the unobserved effect, or just with the vector of variables is inconsequential in terms of the estimated parameters.

Chapter 6
MODELING RESULTS

This chapter presents the results from the model estimations. The organization of the models shown in this chapter follows the analytical frameworks presented in chapter 4. First, the single-state process models will be discussed and followed by the dual-state process models. In each process, frequentist and Bayesian results are discussed respectively.

6.1 Analyses and Results of Single-State Process Models

6.1.1 Frequentist Models of Median Crossover Accident Frequency

In the development of models of both single-state and multi-state processes using frequentist analysis, Limdep 8.0 and Gauss 3.2.38 software are used. Depending on the nature of the problem, standardized or customized estimation algorithms were employed. In order to compare the robustness of variables with different modeling issues and the accuracy of predictions, the same set of specifications is used.

In all single-state models, factors positively correlating with median crossover accidents included the interaction variables between length of median and three categories of median widths as well as the number of interchanges in the section. The three categories of the median width variable interacted with the length of the section variable were 1) less than or equal to 40-foot median width, 2) between 40 to 60-foot median width, and 3) greater than 60-foot median width. Different ranges of the median width were experimented with but the result showed that these three categories, when interrelated with the section length, had the greatest impact on the crossover accident likelihood. The magnitudes of the coefficients of the length variables suggest that for a given length, the likelihood of median crossover accidents increases the greatest on sections with median widths between 40 and 60 feet wide in comparison to the other width categories. On the

other hand, sections with median width wider than 60 feet have the least contribution to the likelihood of median crossover frequency.

Three factors negatively correlate with median crossover accident frequencies. These include the traffic volume indicator variable (if average annual daily traffic was less than 5,000), the interaction variable between the number of horizontal curves less than or equal to 0.5 per mile in the section and the average monthly precipitation being less than or equal 1.5 inches, and the interaction variable between the number of horizontal curves greater than 0.5 per mile in the section and the average monthly precipitation being greater than 4 inches.

The weather effect appearing in the form of the interaction variables played a significant role in median crossover likelihood. In a section where the average number of horizontal curves was less than or equal to 0.5 per mile, the median crossover counts were expected to decrease if the average monthly precipitation was less than or equal to 1.5 inches. Crossover accident frequency also decreases when average monthly precipitation exceeds 4 inches on sections with greater than 0.5 horizontal curves per mile. Both effects point to the range of interactions between precipitation and horizontal curves on median crossover frequencies.

The base model in this dissertation is the NB model or equivalently, the Poisson model with gamma distributed heterogeneity. Alternatively, one can test heterogeneity as a normally distributed segment-specific effect. The model results are presented in table 6.1 with parameters compared for gamma and normal heterogeneities. In table 6.1, it is noted that the estimated parameters in both models are identical but the standard errors somewhat different; however, both models still gain high level of confidence (exceeding 95 percent). It is noted here that the dispersion parameter in Poisson-normal model is not statistically different from zero. Both models have the same restricted log-likelihood (all estimated parameter are constrained to zero) at -1462.3480. Also, the log-likelihood values for the converged models are -727.9237.

Table 6.1 Parameter Comparisons between Poisson-Gamma and Poisson-Normal Heterogeneity Models of Median Crossover Accident Frequency

Variable	Poisson with Gamma Heterogeneity		Poisson with Normal Heterogeneity	
	β* (σ**)	t-stat	β (σ)	t-stat
Constant	-1.8990 (0.1326)	-14.3180	-1.8992 (0.2308)	-8.2300
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-0.9455 (0.1725)	-5.4810	-0.9456 (0.1850)	-5.1130
Length of the roadway section on median widths less than or equal to 40 feet	0.3177 (0.0433)	7.3340	0.3177 (0.0372)	8.5330
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.4324 (0.0619)	6.9810	0.4325 (0.0573)	7.5530
Length of the roadway section on median widths greater than 60 feet	0.1122 (0.0259)	4.3250	0.1122 (0.0300)	3.7370
The number of interchanges in the section	0.2398 (0.0815)	2.9420	0.2398 (0.0847)	2.8310
Interaction 1 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-1.0116 (0.3495)	-2.8940	-1.0120 (0.3570)	-2.8350
Interaction 2 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.5329 (0.2166)	-2.4600	-0.5323 (0.2239)	-2.378
Dispersion parameter (α)	0.8295 (0.2037)	4.0720	0.8307 (0.5991)	1.386
ϖ	-	-	$0.21*10^{-8}$ ($0.94*10^{8}$)	0.0000
Restricted log-likelihood (All parameters = 0, α = 0)	-1,462.3480		-1,462.3480	
Log-likelihood at convergence	-727.9237		-727.9237	
Adjusted ρ^2	0.4993		0.4989	
Number of observations	1,375[Ψ]			

* Estimated Coefficient, ** Standard Error

[Ψ] 5 years of data for 275 separate non-median barrier sections

Table 6.2 shows the results of random effects Poisson (REP) models of median crossover accident frequency. In comparison with NB models, the coefficient values of REP are similar to those of NB but not identical. Compared to the REP model, the NB model underestimated the standard errors of the parameter estimates, therefore overestimating the t-statistic. Overestimating the t-statistic potentially results in the identification of irrelevant variables as significant effects. From a policy standpoint, extra variables, which should not exist in the model if t-statistic is not overestimated, imply unnecessary data collection needs required for programming and prioritization. Despite the fact that

REP models have higher standard errors; leading to lower t-statistic, the high levels of confidence in REP models still remain. It is noted that the NB model is the special case of the REP model if there is one observation in the roadway section. Without the section-specific effects, all estimated parameters and standard errors would be exactly the same as those in the NB model. The log-likelihood at convergence in the REP models improves in comparison to the NB; it is concluded that the REP model is a superior alternate to the NB when modeling longitudinal median crossover accident frequencies. Note the log-likelihoods are comparable across the single-state process models in frequentist method because they share the same specifications and the same dataset.

Table 6.2 Parameter Comparisons between Gamma Random Effects Poisson and Normal Random Effects Poisson Models of Median Crossover Accident Frequency

Variable	Gamma Random Effects Poisson		Normal Random Effects Poisson	
	β* (σ**)	t-stat	β (σ)	t-stat
Constant	-1.9991 (0.1617)	-12.3600	-2.1284 (0.1729)	-12.3060
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-0.9474 (0.2076)	-4.5630	-0.9454 (0.2055)	-4.6010
Length of the roadway section on median widths less than or equal to 40 feet	0.3560 (0.0453)	7.8530	0.3090 (0.0359)	8.6190
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.4448 (0.0652)	6.8240	0.4171 (0.0640)	6.5210
Length of the roadway section on median widths greater than 60 feet	0.1198 (0.0387)	3.0920	0.1155 (0.0374)	3.0860
The number of interchanges in the section	0.2693 (0.1037)	2.5960	0.2494 (0.0986)	2.5310
Interaction 1 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-1.0295 (0.4344)	-2.3700	-0.9169 (0.4545)	-2.0170
Interaction 2 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.4625 (0.2357)	-1.9630	-0.4376 (0.2333)	-1.8760
Dispersion parameter (α)	0.4289 (0.1421)	-3.0170	-	-
ϖ	-	-	0.6344 (0.1212)	5.234
Restricted log-likelihood (All parameters = 0, α = 0)	-1,462.3480		-1,462.3480	
Log-likelihood at convergence	-723.1721		-724.8458	
Adjusted ρ^2	0.5026		0.5014	
Number of observations	1,375$^{\Psi}$			

* Estimated Coefficient, ** Standard Error. $^{\Psi}$ 5 years of data for 275 separate non-median barrier sections

The negative multinomial (NM) model results of median crossover accident frequency are presented in table 6.3. Comparing the results from the NM and REP models with gamma distributed heterogeneity it is noted that both models present the same estimated parameters and log-likelihood but the standard errors of the NM model are slightly higher than those in the REP model. Theoretically, the NM model will arrive at the same coefficients and level of confidence as well as log-likelihood as those from the REP model. (Difference in standard errors here is attributed to estimation algorithms). Guo (1996) discusses the NM model for the treatment of correlation of counts. On the other hand, Hausman et al. (1984) developed the REP model to specifically capture section-specific heterogeneity. The equality of REP and NM in these findings establishes that correlation of accident counts in the same roadway section reflects section-specific heterogeneity and vice versa.

Table 6.3 Negative Multinomial Model of Median Crossover Accident Frequency

Variable	β* (σ**)	t-stat
Constant	-1.9990 (0.1987)	-10.0612
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-0.9476 (0.2110)	-4.4907
Length of the roadway section on median widths less than or equal to 40 feet	0.3560 (0.1029)	3.4588
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.4448 (0.0951)	4.6765
Length of the roadway section on median widths greater than 60 feet	0.1198 (0.0389)	3.0769
The number of interchanges in the section	0.2692 (0.1185)	2.2723
Interaction 1 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-1.0294 (0.4788)	-2.1500
Interaction 2 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.4625 (0.2575)	-1.7958
The inversion of dispersion parameter ($\theta = 1/\alpha$)	2.3318 (3.1036)	0.7513
Restricted log-likelihood (All parameters = 0, $\alpha = 0$)	-1,462.3480	
Log-likelihood at convergence	-723.1721	
Adjusted ρ^2	0.5026	
Number of observations	1,375[Ψ]	

* Estimated Coefficient, ** Standard Error

[Ψ] 5 years of data for 275 separate non-median barrier sections

Table 6.4 presents the result of the random effects negative binomial (RENB) model. Recall that the RENB model is the extension of REP model with the extra assumption that the overdisperison parameter is randomly distributed across groups. The RENB shows that except the constant, all estimated parameters and standard errors are close to those from the REP with gamma distributed heterogeneity. The overdisperison parameter in terms of (a_n, b_n) is not significantly different from zero because the level of confidence for a_n is less than 90 percent (i.e. t-statistic = 0.9150). This result suggests that majority of the overdispersion may be captured by the REP model's treatment of section-specific heterogeneity and thereby rendering the overdisperison factor insignificant. In this case, the RENB degenerates to REP and therefore both models yield the same coefficients and standard errors. Note, some part of the heterogeneity (not significant to be captured by the overdispersion parameter and the section-specific error term) was subsumed in the constant in the REP model and as a result the REP model's constant (-1.9991) is changed to 1.3033 in RENB model. Also, the RENB model has a slightly higher log-likelihood than REP model and the RENB model is marginally better in overall fit.

From a single-state process model standpoint, all explanatory variables are "common variables" because all variables exist in every model with high level of confidence (exceeding 90 percent) despite the impact of unobserved heterogeneity and correlation among median crossover accident frequencies.

Table 6.4 Gamma Random Effects Negative Binomial Model of Median Crossover Accident Frequency

Variable	β* (σ**)	t-stat
Constant	1.3033 (1.1581)	1.1250
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-0.9403 (0.2068)	-4.5460
Length of the roadway section on median widths less than or equal to 40 feet	0.3453 (0.0450)	7.6760
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.4354 (0.0666)	6.5370
Length of the roadway section on median widths greater than 60 feet	0.1184 (0.0385)	3.0720
The number of interchanges in the section	0.2612 (0.1028)	2.5410
Interaction 1 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-1.0010 (0.4345)	-2.3040
Interaction 2 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.4623 (0.2408)	-1.9200
a_n	69.4037 (75.8869)	0.9150
b_n	2.5757 (0.9589)	2.6860
Restricted log-likelihood (All parameters = 0, $\alpha = 0$)	-1,462.3480	
Log-likelihood at convergence	-722.8471	
Adjusted ρ^2	0.4910	
Number of observations	1,375$^{\Psi}$	

* Estimated Coefficient, ** Standard Error

$^{\Psi}$ 5 years of data for 275 separate non-median barrier sections

6.1.2 Bayesian Approach Models of Median Crossover Accident Frequency

Table 6.5 presents estimation results for the Bayesian Poisson with gamma and normally distributed heterogeneity of median crossover accident frequency whereas Bayes estimates of Poisson with gamma and normal group effects are shown in table 6.6. Also, the model results of hierarchical Bayes estimates with gamma and normally distributed heterogeneity are introduced in table 6.7. As shown in all results of Bayesian analysis, the specifications are the same as those in frequentist approach. These findings show that Bayesian heterogeneity Poisson models agree closely with classical NB models and Bayesian group effects Poisson models also suggest the same results as those of the REP

models in terms of estimated parameter and confidence levels. It is noted here that normally distributed prior, N[0, 0.001], for all variables helps establish the equality between the classical analysis and Bayesian analysis; it is for all practical purposes a uniform prior and not informative.

The useful results from the Bayesian analysis are the credibility intervals associated with coefficient values and predictive power. As shown in tables 6.5 to 6.7, the trend in parameter signs and magnitudes as their credibility levels increase from 2.5 percent to 97.5 percent can be noted. The mean value is the value of the parameter typically reported in most studies, including frequentist studies. As expected, some variation exists between the 2.5^{th} percentile credibility and 97.5^{th} percentile credibility estimates of the coefficients in every model. All variables including constants maintained their "sign" as credibility levels increase. The trend in estimated parameters' credibility levels suggests that all variables remain fairly robust in response to common modeling issues such as heterogeneity and correlations among accident counts.

Classical likelihood estimates are provided in the tables for comparison with Bayesian results for single-state process models. Hierarchical Bayesian Poisson presents the highest adjusted ρ^2, 0.634, among gamma distributed heterogeneity models whereas in the group of normally distributed heterogeneity models, the Bayesian Poisson gives the greatest adjusted ρ^2 at 0.637.

Table 6.5 Parameter Comparisons between Bayesian Poisson-Gamma and Bayesian Poisson-Normal Models of Median Crossover Accident Frequency

Variable	Bayesian Poisson with Gamma Heterogeneity				Bayesian Poisson with Normal Heterogeneity			
	β* σ ** t***	Credibility Percentiles of Coefficient			β* σ ** t***	Credibility Percentiles of Coefficient		
		2.5 %	50 %	97.5 %		2.5 %	50 %	97.5 %
Constant	-1.897 0.139 -13.638	-2.165	-1.894	-1.613	-2.157 0.153 -14.144	-2.457	-2.154	-1.847
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-0.953 0.182 -5.242	-1.324	-0.950	-0.609	-0.953 0.167 -5.711	-1.304	-0.945	-0.631
Length of the roadway section on median widths less than or equal to 40 feet	0.323 0.043 7.565	0.230	0.322	0.406	0.283 0.034 8.372	0.218	0.283	0.354
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.431 0.064 6.781	0.305	0.431	0.555	0.414 0.059 7.041	0.299	0.413	0.529
Length of the roadway section on median widths greater than 60 feet	0.112 0.026 4.305	0.063	0.112	0.162	0.106 0.025 4.244	0.057	0.106	0.153
The number of interchanges in the section	0.234 0.083 2.823	0.057	0.234	0.384	0.226 0.085 2.650	0.054	0.229	0.402
Interaction 1 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-1.050 0.355 -2.962	-1.786	-1.046	-0.382	-0.964 0.347 -2.774	-1.676	-0.961	-0.283
Interaction 2 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.558 0.222 -2.515	-1.000	-0.550	-0.136	-0.554 0.218 -2.549	-0.967	-0.547	-0.134
The inversion of dispersion parameter ($\theta = 1/\alpha$)	1.206 0.288 4.192	0.769	1.159	1.934	-	-	-	-
σ	-	-	-	-	1.283 0.164 7.809	1.022	1.254	1.658
Restricted log-likelihood (All parameters = 0)	-1,375.000				-1,789.000			
Log-likelihood at convergence	-626.700				-646.000			
Adjusted ρ^2	0.542				0.637			

* Estimated Coefficient, ** Standard Error, *** t-statistic

Table 6.6 Parameter Comparisons between Bayesian Poisson-Gamma with Group Effects and Bayesian Poisson-Normal with Group Effects Models of Median Crossover Accident Frequency

Variable	Bayesian Poisson with Gamma Group Effects				Bayesian Poisson with Normal Group Effects			
	β* σ ** t***	Credibility Percentiles of Coefficient			β* σ ** t***	Credibility Percentiles of Coefficient		
		2.5 %	50 %	97.5 %		2.5 %	50 %	97.5 %
Constant	-1.981 0.183 -10.831	-2.357	-1.983	-1.615	-2.176 0.165 -13.188	-2.503	-2.176	-1.849
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-0.974 0.200 -4.877	-1.383	-0.979	-0.578	-0.983 0.184 -5.337	-1.344	-0.984	-0.634
Length of the roadway section on median widths less than or equal to 40 feet	0.357 0.051 7.065	0.262	0.357	0.458	0.326 0.044 7.432	0.242	0.326	0.413
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.451 0.079 5.744	0.301	0.446	0.611	0.424 0.079 5.359	0.261	0.425	0.575
Length of the roadway section on median widths greater than 60 feet	0.122 0.033 3.725	0.057	0.123	0.183	0.120 0.030 4.027	0.062	0.120	0.181
The number of interchanges in the section	0.253 0.113 2.240	0.023	0.245	0.477	0.257 0.101 2.556	0.065	0.255	0.456
Interaction 1 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-1.032 0.398 -2.596	-1.845	-1.004	-0.306	-0.961 0.396 -2.425	-1.733	-0.963	-0.190
Interaction 2 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.469 0.225 -2.082	-0.908	-0.460	-0.034	-0.452 0.231 -1.956	-0.898	-0.447	-0.002
The inversion of dispersion parameter ($\theta = 1/\alpha$)	2.097 0.537 3.906	1.249	2.029	3.364	-	-	-	-
σ	-	-	-	-	1.509 0.203 7.423	1.166	1.486	1.972
Restricted log-likelihood (All parameters = 0)	-1,434.000				-1,699.000			
Log-likelihood at convergence	-675.500				-679.500			
Adjusted ρ^2	0.526				0.598			

* Estimated Coefficient, ** Standard Error, *** t-statistic

Table 6.7 Parameter Comparisons between Hierarchical Bayesian Poisson with Gamma and Normally Distributed Heterogeneity of Median Crossover Accident Frequency

Variable	Hierarchical Bayesian Poisson with Gamma Heterogeneity				Hierarchical Bayesian Poisson with Normal Heterogeneity			
	β* σ ** t***	Credibility Percentiles of Coefficient			β* σ ** t***	Credibility Percentiles of Coefficient		
		2.5 %	50 %	97.5 %		2.5 %	50 %	97.5 %
Constant	-2.860 0.364 -7.864	-3.641	-2.811	-2.242	-2.187 0.176 -12.412	-2.537	-2.186	-1.855
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-1.662 0.397 -4.189	-2.593	-1.616	-0.980	-1.002 0.199 -5.030	-1.415	-0.994	-0.624
Length of the roadway section on median widths less than or equal to 40 feet	0.539 0.101 5.329	0.362	0.537	0.752	0.302 0.041 7.336	0.223	0.301	0.382
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.651 0.112 5.832	0.437	0.648	0.883	0.430 0.065 6.585	0.306	0.429	0.561
Length of the roadway section on median widths greater than 60 feet	0.178 0.046 3.871	0.093	0.176	0.275	0.112 0.028 4.061	0.057	0.113	0.164
The number of interchanges in the section	0.318 0.120 2.652	0.090	0.317	0.562	0.230 0.087 2.647	0.057	0.231	0.394
Interaction 1 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-2.451 1.081 -2.267	-5.280	-2.182	-0.925	-1.122 0.375 -2.995	-1.877	-1.116	-0.388
Interaction 2 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-1.155 0.431 -2.682	-2.090	-1.119	-0.429	-0.601 0.237 -2.533	-1.074	-0.594	-0.151
The inversion of dispersion parameter ($\theta = 1/\alpha$)	2.664 0.790 3.373	1.477	2.529	4.356	-	-	-	-
σ	-	-	-	-	4.109 0.900 4.569	2.531	4.058	5.896
Restricted log-likelihood (All parameters = 0)	-1,675.000				-1,531.000			
Log-likelihood at convergence	-609.600				-639.500			
Adjusted ρ^2	0.634				0.580			

* Estimated Coefficient, ** Standard Error, *** t-statistic

6.2 Analyses and Results of Dual-State Process Models

6.2.1 Development of Loading Factors for Standard Error Adjustment in Frequentist Zero-Inflated Poisson

Table 6.8 presents a comparative analysis of standard errors for significant parameters in the NB and NM models of median crossover accidents. It is to be noted here that only significant effects appearing in the NM model are used to develop the loading factors. It may be that variables significant in the NB may not be significant in the NM due to inflation in the standard error. As is noted in table 6.8, the loading factor varies by parameter, from a minimum adjustment of 1.1889 for the precipitation-horizontal curves interaction variable to 2.3759 for the length variable for medians narrower than or equal to 40 feet. The standard error adjustment for the "alpha" parameter (overdispersion effect) is ignored in my discussion, since the alpha parameter does not appear in the ZIP model of median crossover accidents. The ZIP model is discussed next, considering the loading factors developed in table 6.8.

6.2.2 An Empirically Adjusted Correlation among Accident Count in Frequentist Zero-Inflated Poisson Model of Median Crossover Accident Frequency

A ZIP-Full model of median crossover accidents was estimated and loading factors from table 6.8 were used to adjust for correlation of accident counts in the five-year median crossover dataset. Median crossover accident counts were the dependent variable in this estimation. The ZIP-Full model was estimated and validated by the Vuong statistic. The result revealed a Vuong statistic of 3.1618 suggesting that the ZIP-Full as the favorable model compared to the basic Poisson model.

Table 6.8 Load Factors Developed from NB and NM Models for Adjusting Standard Errors in Zero-Inflated Poisson (Full) Model of Median Crossover Accident Frequency

Variable	NB	NM	Load Factor[ξ]
	β* (σ**)	β (σ)	
Constant	-1.8990 (0.1326)	-1.9990 (0.1987)	1.4980
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-0.9455 (0.1725)	-0.9476 (0.2110)	1.2233
Length of the roadway section on median widths less than or equal to 40 feet	0.3177 (0.0433)	0.3560 (0.1029)	2.3759
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.4324 (0.0619)	0.4448 (0.0951)	1.5357
Length of the roadway section on median widths greater than 60 feet	0.1122 (0.0259)	0.1198 (0.0389)	1.5010
The number of interchanges in the section	0.2398 (0.0815)	0.2692 (0.1185)	1.4538
Interaction 1 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-1.0116 (0.3495)	-1.0294 (0.4788)	1.3698
Interaction 2 between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.5329 (0.2166)	-0.4625 (0.2575)	1.1889
α	0.8295 (0.2037)	-	N.A.
θ (1/α)	-	2.3318 (3.1036)	N.A.
Restricted log-likelihood (All parameters = 0, α=0)	-1,462.3480	-1,462.3480	
Log-likelihood at convergence	-727.9237	-723.1721	
Adjusted ρ^2	0.4993	0.5026	
Number of observations	1,375[Ψ]		

***** Estimated Coefficient, ****** Standard Error

[ξ] The load factor was the proportion of the S.E. of NM to the S.E. of NB

[Ψ] 5 years of data for 275 separate non-median barrier sections

Table 6.9 Zero-Inflated Poisson (Full) Model of Median Crossover Accident Frequency with Correlation Adjusted Standard Errors

Variable	β* (σ**)	t-stat	Load Factor	Adjusted Standard Error	Adjusted t-stat
Non-zero accident Poisson probability state					
Constant	-1.0012 (0.1583)	-6.3250	1.4980	0.2371	-4.2224
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-1.0077 (0.1754)	-5.7440	1.2232	0.2146	-4.6960
Length of section where medians are less than 40 feet wide	0.1499 (0.0263)	5.7020	2.3759	0.0625	2.4000
Length of section where medians are between 40 feet and 60 feet wide	0.3954 (0.0604)	6.5500	1.5357	0.0927	4.2652
Length of section where medians are wider than 60 feet	0.0495 (0.0251)	1.9740	1.5010	0.0377	1.3154
Number of interchanges in section	0.1834 (0.0711)	2.5800	1.4538	0.1034	1.7747
Interaction 1 between average monthly precipitation indicator and the number of horizontal curves per mile indicator (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-0.8648 (0.3235)	-2.6740	1.3698	0.4431	-1.9517
Interaction 2 between average monthly precipitation indicator and the number of horizontal curves per mile indicator (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.4366 (0.2185)	-1.9990	1.1889	0.2598	-1.6810
Zero accident probability state as logistic function					
Constant	1.1049 (0.2950)	3.7460	1.4980	0.4419	2.5004
Length of section where medians are less than 40 feet wide	-1.4137 (0.4414)	-3.2020	2.3759	1.0488	-1.3479
Length of section where medians are between 40 feet and 60 feet wide	-0.3691 (0.1663)	-2.2200	1.5357	0.2554	-1.4455
Length of section where medians are wider than 60 feet	-0.5098 (0.1692)	-3.0130	1.5010	0.2539	-2.0076
Restricted log-likelihood (constant only)	-889.7221[Ψ]				
Log-likelihood at convergence	-718.6493				
Vuong statistic	3.1618				

* Estimated Coefficient, ** Standard Error

[Ψ] This restricted log-likelihood was obtained from the restricted log-likelihood computed by the Poisson model.

In estimating the significance of key variables in the ZIP-Full model, standard errors estimated using the likelihood function shown in equation 5.14 were adjusted through multiplication with the "loading factors" developed in table 6.8. By so doing, I empirically account for correlation present in the dataset. Table 6.9 shows the empirically adjusted correlation ZIP-Full model of median crossover accidents.

In non-zero accident Poisson probability state, the specifications are the same as those in single-state process models. However, in zero accident probability state, there are interaction variables between length and three categories of median widths. The three categories of the median width interacted with the length of the section were 1) less than or equal to 40-foot median width, 2) between 40 to 60-foot median width, and 3) greater than 60-foot median width. These three variables are subset of specifications in non-zero accident Poisson probability state.

The results show that ***before accounting for accident count correlation***, all the variables included in the ZIP-Full model were highly significant and thus have a high level of confidence (exceeding 95 percent). However, ***after accounting for accident count correlation***, the level of significance for all variables decreased. The length variable for medians wider than 60 feet had the lowest t-statistic of 1.3154.

It is also noted that the parameters in non-zero accident state always negatively correlate with the same parameters in zero accident state (e.g. opposite signs). For example, the length variable for medians narrower than 40 feet positively correlates with median crossover accidents, while it negatively correlates with the probability of a section being in a zero count state in its lifetime.

As a final thought, it should be noted that the impact of median widths on the probability of a median crossover accident occurring in a section's lifetime (zero versus non-zero state) follows a trend that is different from their impacts on positive crossover frequencies (the Poisson state.). In the zero state, for a given length, median widths in excess of 60 feet have the highest odds ratio of being in a non-zero accident state, while

the count state effects suggests they will have the least number of positive crossover counts.

6.2.3 Hierarchical Bayesian Zero-Inflated Negative Binomial Model of Median Crossover Accident Frequency

A hierarchical Bayesian zero-inflated negative binomial (ZINB), (τ) model of median crossover accidents was estimated by the MCMC algorithm using the Winbugs software by assuming that unobserved effects in non-zero accident probability state are captured by a gamma density prior and that heterogeneity in the zero accident probability state is normally distributed.

Table 6.10 shows the results of the hierarchical Bayesian ZINB (τ) model of median crossover accident frequencies. The findings show that all variables, except the tau (τ) parameter, were highly significant and thus have a high level of confidence (exceeding 95 percent). A low level of confidence for the tau (τ) parameter leads to the conclusion that it is not significantly different from zero. A zero value of τ suggests the estimated lifetime probability of roadway sections being in the zero probability state to equal 0.5. (This is a fairly restrictive assumption to make, since some sections may deviate from a lifetime probability of zero.) An alternate way to describe the zero state would be to express τ as a section-specific parameter with an assumed distributional characteristic.

Factors positively correlating with median crossover accidents included the interaction variables between length and the three categories of median widths, as well as the number of interchanges in the section. The three categories of median widths interacted with the length of the section were a) less than or equal to 40 foot median width, b) 40 to 60 foot median width, and c) greater than 60 foot median width. Different ranges of the median width were experimented but the result showed that these three categories, when interrelated with the section length, had the greatest impact on the crossover accident likelihood.

The magnitudes of the coefficients of the length and median width interacted variables provided insights into median width effects. All other factors accounted for; the likelihood of median crossover accidents is greatest on sections with median width between 40 and 60 feet wide in comparison to the other width categories. One would argue that less than 40 foot median widths would have the greatest impact; however, it is to be noted here that the sample size was relatively slow, and in addition, sections that are left unbarriered with less than 40 foot median widths are inherently likely to have very low crossover likelihoods.

Three factors negatively correlating with median crossover accidents included the traffic volume indicator (1 if average annual daily traffic was less than 5,000 vehicles), and the two interaction variables between number of horizontal curves per mile and average monthly precipitation indicators. The first curve-precipitation variable captures the interaction between number of horizontal curves per mile less than or equal to 0.5 in the section and average monthly precipitation being less than or equal to 1.5 inches. The second curve-precipitation variable captures the interaction between number of horizontal curves per mile greater than 0.5 in the section and average monthly precipitation being greater than 4.0 inches.

The weather effect appearing in the form of the interaction variables played a significant role in median crossover likelihood. In a section where the average number of curves was less than or equal to 0.5 per mile, the median crossover counts were expected to decrease if the average monthly precipitation was less than or equal to 1.5 inches. Likewise, if the average monthly precipitation was greater than 4.0 inches, median crossover accidents are expected to decrease with higher magnitude on sections with horizontal curves per mile being greater than 0.5. The interaction between number of horizontal curves being less than 0.5 and average monthly precipitation in the 1.6 to 3.99 inches was insignificant.

Different structures of heterogeneity and prior densities for heterogeneity were experimented with. However, among the class of hierarchical Bayesian ZINB (τ) models,

the hierarchical Bayesian ZINB (τ) model with gamma distributed heterogeneity in the non-zero accident probability state and normally distributed heterogeneity in the zero accident probability state is the most appropriate model in terms of providing the best predictions of median crossover accident frequencies.

Table 6.10 Hierarchical Bayesian Zero-Inflated Negative Binomial (τ) Model of Median Crossover Accident Frequency

Variable	β* (σ**)	t-stat	Credibility Percentiles of Coefficient		
			2.5 %	50 %	97.5 %
Zero accident probability state as logistic function and Non-zero accident probability state as negative binomial function (vectors of regressors constrained to be the same)					
Constant	-0.937 (0.118)	-7.957	-1.178	-0.938	-0.716
Per-lane AADT indicator (1 if per-lane AADT <= 5000 vehicles, 0 otherwise)	-0.734 (0.155)	-4.734	-1.049	-0.731	-0.449
Length of section where medians are less than 40 feet wide	0.204 (0.033)	6.272	0.138	0.203	0.267
Length of section where medians are between 40 feet and 60 feet wide	0.338 (0.055)	6.151	0.233	0.341	0.438
Length of section where medians are wider than 60 feet	0.083 (0.023)	3.604	0.038	0.083	0.130
Number of interchanges in section	0.165 (0.067)	2.455	0.042	0.161	0.306
Interaction 1 between average monthly precipitation indicator and the number of horizontal curves per mile indicator (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	-0.676 (0.300)	-2.253	-1.265	-0.671	-0.066
Interaction 2 between average monthly precipitation indicator and the number of horizontal curves per mile indicator (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	-0.404 (0.198)	-2.038	-0.808	-0.397	-0.024
τ	-17.430 (13.580)	-1.284	-50.260	-12.680	-1.741
Restricted log-likelihood (All parameter = 0)	-1,375.000[Ψ]				
Log-likelihood at convergence	-510.600				

* Estimated Coefficient, ** Standard Error

[Ψ] This restricted log-likelihood was obtained from the restricted log-likelihood computed by Bayesian Poisson-Gamma model.

6.3 Prediction and Temporal Transferability Test

6.3.1 Prediction Test

To assess the predictive abilities of the various models, model predictions are presented in this section. In the general case of accident count models, two measures of predictive effectiveness are used: a) mean absolute deviation (MAD) and b) root mean square error (RMSE). MAD is the average of the absolute values of the prediction errors. It is appropriate when the cost of forecast errors is proportional to the absolute size of the forecast error. This criterion is also called MAE (mean absolute error). RMSE is the square root of the average of the squared values of the prediction errors and is appropriate to situations in which the cost of an error increases as the square of that error. In this research, percent change of predicted count from observed count is not appropriate due to the significant presence of zero counts in the database. MAD and RMSE using in this research are formulated as follows:

$$MAD = (\sum_{i=1}^{275} | O_i - P_i |) / 275 \tag{6.1}$$

$$RMSE = \sqrt{(\sum_{i=1}^{275} (O_i - P_i)^2) / 275} \tag{6.2}$$

where O_i and P_i are observed and predicted median crossover accident counts in year 1994 from section 1 to 275. If the predicted count is close to observed count, MAD and RMSE will be minimized. In other words, the lesser the MAD and RMSE, the more accurate the prediction provided by the model. In prediction tests, all models were developed using 4 years of data (1990-1993) and then tested against the extra-sample observations or the predicted crossover frequencies in the fifth year (1994). In addition, all models were developed using 5 years of data (1990-1994) and then the predictions were tested against within-sample observations.

The predictions of classical frequentist method models are presented in table 6.11 while table 6.12 shows the predictions from Bayesian approach models. In the classical approach, Poisson-gamma and Poisson-normal heterogeneity models arrive at the same MAD and RMSE of 0.36 and 1.07 respectively in extra-sample observation predictions and likewise NM and REP with gamma distributed heterogeneity also provide the same MAD and RMSE of 0.40 and 1.80 respectively in extra-sample observation predictions. In single-state process of classical method models, REP with normally distributed heterogeneity has the lowest MAD and RMSE and hence it is model providing more reliable prediction in comparison with others in single-state process. The dual-state ZIP-Full yields the best prediction among the frequentist approach models with the lowest MAD and RMSE. It is also noted here that the random effects negative binomial model provides unreasonable predictions because of the fact that overdispersion parameter in this model induces additional noise.

As expected, the models in Bayesian analysis arrive at better predictions in comparison with classical models with the same model and prior heterogeneity structures. For example, in extra-sample observation predictions, Bayesian gamma heterogeneity Poisson shows a MAD of 0.24 compared to MAD of 0.36 in the classical Poisson with gamma distributed heterogeneity. In addition, the Bayesian segment-specific effects Poisson with gamma prior density arrives at lower errors in forecasting both extra-sample and sub-sample observations in comparison with those from the classical REP with gamma distributed heterogeneity. Interestingly, both models in the hierarchical Bayesian Poisson structure provide the same overall predictive accuracy as non hierarchical structures in the Bayesian analysis (e.g. Bayesian heterogeneity Poisson). It might be concluded that the hierarchical structure seems to not add informative value related to the excess zero problem in median crossover accident frequency databases. Similar to the finding on predictive accuracy of the ZIP model in classical approach, the Bayesian ZINB (τ) model provides better prediction in comparison with other models in Bayesian analysis.

Table 6.11 The Prediction of Classical Frequentist Approach Models

Model	Errors in Forecasting Observations	
	Extra-Sample 1 years (1994)	Sub-Sample 5 years (1990-1994)
Single-State Process		
Heterogeneity Poisson		
• Gamma distributed heterogeneity	0.36* (1.07**)	0.39 (1.13)
• Normally distributed heterogeneity	0.36 (1.07)	0.39 (1.13)
Cluster Heterogeneity Poisson		
• Negative multinomial model	0.40 (1.80)	0.43 (1.84)
• Random effects Poisson		
– Gamma distributed heterogeneity	0.40 (1.80)	0.43 (1.84)
– Normally distributed heterogeneity	0.32 (0.80)	0.35 (0.88)
• Random effects negative binomial		
– Gamma distributed heterogeneity	8.29 (46.04)	8.29 (46.07)
Dual-State Process		
Zero-Altered Probability Process		
• Zero-inflated Poisson (Full) model with correlation-adjusted standard errors	0.30 (0.47)	0.33 (0.56)

* Mean Absolute Deviation (MAD)

** Root Mean Square Error (RMSE)

Table 6.12 The Prediction of Bayesian Approach Models

Model	Errors in Forecasting Observations	
	Extra-Sample 1 years (1994)	Sub-Sample 5 years (1990-1994)
Single-State Process		
Bayesian Heterogeneity Poisson		
• Gamma distributed heterogeneity	0.24* (0.36**)	0.30 (0.41)
• Normally distributed heterogeneity	0.24 (0.36)	0.30 (0.41)
Bayesian Segment-Specific Effects Poisson		
• Gamma distributed heterogeneity	0.28 (0.43)	0.33 (0.51)
• Normally distributed heterogeneity	0.28 (0.43)	0.33 (0.51)
Hierarchical Bayesian Poisson		
• Gamma distributed heterogeneity	0.24 (0.36)	0.30 (0.41)
• Normally distributed heterogeneity	0.24 (0.36)	0.30 (0.41)
Dual-State Process		
Bayesian Zero-Altered Probability Process		
• Zero-inflated negative binomial (τ) model	0.21 (0.31)	0.22 (0.38)

* Mean Absolute Deviation (MAD)

** Root Mean Square Error (RMSE)

As shown in both table 6.11 and 6.12, MAD and RMSE appear to be highly correlated and one would argued that either forecasting method is redundant. The fact that highly correlated forecasting errors does not necessarily mean its low quality of forecasting methods. It should be keep in mind that the significant presence of zero counts in the

database causes the occurrence of highly correlated forecasting errors. Kennedy (1998) suggested that the best forecasting method, overall, is a "combined" forecast, formed as a weighted average of a variety of forecasts, each generated by a different technique. However, in this research, two forecasting methods, MAD and RMSE, are formed separately and hence the recommendation for the future research is to incorporated two forecasting methods to ensure the accuracy of median crossover accident forecast.

6.3.2 Structural Change in Parameters (Stability) Test

In this section, I present briefly my assessment of potential structural changes in parameters within-sample. It must be noted that the original parameters are estimated using a longitudinal sample of five consecutive years. The structure of the models is to take advantage of the longitudinal nature of the dataset through fixed and random effects. However, the notion of structural change provides some idea on the stability of parameters to the extent that they may help identify the minimum length of the panel for the assessment of median crossover counts. A likelihood based transferability test is performed to test whether or not their estimated coefficients are transferable. Transferability tests involve the computation of likelihoods of sub-samples in order to compare with the likelihood of the overall sample. Two points are of note here: a) the likelihood test involves the computation of all involved likelihoods, not just the restricted or unrestricted likelihoods, and b) the likelihood test only suggests that the parameter space is transferable or not. It does not say specifically which parameters are transferable and which ones aren't. Careful experimentation with parameter restrictions is required to conduct parameter specific assessments of structural change and to ensure that predictions made with the model have some validity in that the estimated parameters are stable over time. In testing temporal transferability, log-likelihood statistic tests can be applied as presented in the following equation.

$$\text{Log-Likelihood Statistic} = -2[\text{Log-L}(\beta_{(5\text{ years})}) - \text{Log-L}(\beta_{(3\text{ years})}) - \text{Log-L}(\beta_{(2\text{ years})})] \qquad (6.2)$$

where Log-L($\beta_{(5\ years)}$) is the log-likelihood at convergence of the model estimated with 5 years of data (1990-1994), Log-L($\beta_{(3\ years)}$) is the log-likelihood at convergence of the sub model estimated with 3 years of data (1990-1992) and Log-L($\beta_{(2\ years)}$) is the log-likelihood at convergence of the sub model estimated with 2 years of data (1993-1994). In this test, same set of specifications are used to calculate log-likelihood. Log-likelihood statistic, LL, is chi-square, χ^2, distributed with the degrees of freedom equal to the number of estimated parameters in the model including constant. This log-likelihood statistic is tested against the null hypothesis that all estimated parameters are transferable.

As shown in table 6.13, the test results show that the estimated coefficients of every model in the classical frequentist sense are transferable because all log-likelihood statistics are less than chi-square, χ^2, statistic at 95 percent level of confidence. On the other hand, in the Bayesian context, the non-hierarchical Bayesian heterogeneity Poisson models pass the transferability test as presented in table 6.14. The log-likelihood statistic for the Bayesian segment-specific effects Poisson with gamma distributed heterogeneity slightly exceeds the cutting point while the model with normally distributed heterogeneity in the Bayesian segment–specific Poisson has log-likelihood statistic less than cutting point. Interestingly, the hierarchical Bayesian Poisson with gamma prior density of heterogeneity generates the highest log-likelihood statistic. These findings can be explained by the fact that the hierarchical structures in accident mean rate and normal prior density of accident mean rate introduce variability in estimated parameters over time. It is also noted here that log-likelihoods at convergence for the hierarchical Bayesian Poisson with normally distributed heterogeneity in single-state process and for Bayesian zero-altered model in dual-state process could not be calculated due to lack of sufficient panel data.

Table 6.13 Temporal Transferability Test of Classical Frequentist Approach Models

Model	Log-likelihood at convergence			LL statistic	χ^2 statistic*
	5 years (1990-1994)	3 years (1990-1992)	2 years (1993-1994)		
Single-State Process					
Heterogeneity Poisson					
• Gamma distributed heterogeneity	-727.92	-454.06	-269.79	8.14	16.92
• Normally distributed heterogeneity	-727.92	-454.06	-269.71	8.30	16.92
Cluster Heterogeneity Poisson					
• Random effects Poisson					
– Gamma distributed heterogeneity	-723.17	-454.26	-268.77	0.29	16.92
– Normally distributed heterogeneity	-724.85	-454.94	-268.32	3.17	16.92
• Negative multinomial model	-723.17	-454.28	-268.83	0.13	16.92
• Random effects negative binomial					
– Gamma distributed heterogeneity	-722.85	-453.92	-268.79	0.26	16.92
Dual-State Process					
Zero-Altered Probability Process					
• Zero-inflated Poisson (Full) model with correlation-adjusted standard errors	-718.65	-450.97	-260.94	13.48	19.68

* At 95 percent level of confidence

Table 6.14 Temporal Transferability Test of Bayesian Approach Models

Model	Log-likelihood at convergence			LL statistic	χ^2 statistic*
	5 years (1990-1994)	3 years (1990-1992)	2 years (1993-1994)		
Single-State Process					
Bayesian Heterogeneity Poisson					
• Gamma distributed heterogeneity	-626.70	-386.80	-235.50	8.80	16.92
• Normally distributed heterogeneity	-646.00	-400.70	-245.10	0.40	16.92
Bayesian Segment-Specific Effects Poisson					
• Gamma distributed heterogeneity	-675.50	-418.50	-246.70	20.60	16.92
• Normally distributed heterogeneity	-679.50	-422.20	-249.90	14.80	16.92
Hierarchical Bayesian Poisson					
• Gamma distributed heterogeneity	-609.60	-374.40	-211.20	48.00	16.92
• Normally distributed heterogeneity	-639.50	NA	NA	NA	16.92
Dual-State Process					
Bayesian Zero-Altered Probability Process	-510.60	NA	NA	NA	27.59
• Zero-inflated negative binomial (τ) model					

* At 95 percent level of confidence

Chapter 7
CONCLUSIONS AND RECOMMENDATIONS

7.1 Model Conclusions and Recommendations

In conclusion, a suite of plausible modeling structures for the assessment of median crossover accident frequencies was developed in detail. The models considered ranged from frequentist models involving overdispersed accident data in a longitudinal sense, including fixed and random effects structures. In addition, zero-inflated models were also examined in the frequentist sense. Bayesian counterparts to these structures were also developed and tested for a variety of priors, including uninformative priors, normal priors and gamma priors for heterogeneity. It can be deduced from the results in the previous chapter that the behavior of Bayesian structures is relatively unstable when hierarchy is introduced. While hierarchy on the one hand, helps provide flexibility for better predictions, it also introduces model convergence issues and thus related stability concerns. What can be deduced from the results on hierarchical models is that panels need to be longer than what would be required for classical model development. The trade-offs between assessments of parameter uncertainty, predictions and model data requirements and convergence appears to be a prominent area of research for the years to come.

Results from the previous chapter on normal heterogeneity are worthy of further testing. While the gamma heterogeneity treatment is well known in the literature and consistent with analytical tractability requirements, much is unknown about the pros and cons of normal heterogeneity treatments. One would expect the normal heterogeneity structure to be more unstable due to the fact that it is not a closed-form structure, with estimations being mainly approximations using quadratures with finite set of points. However, compared to the gamma heterogeneity predictions, the normal heterogeneity structure appears to hold valid in the median crossover context. It is fair to say that for the foreseeable future, gamma heterogeneity, while showing marginally poorer predictive accuracy, is a reasonable method to incorporate the effect of various unknowns: known

unknowns including measurable and immeasurable ones, as well as unknown unknowns. One of the main purposes of this dissertation was to develop fairly parsimonious models while developing sufficient predictive power. To this extent, commonly available variables such as median widths, precipitation data from nationally available sources and roadway geometric data, such as horizontal curvature, number of interchanges, as well as traffic volume, were used in the development of models. As such, one can view the model structures (setting aside the mathematical component) as "naïve" models or benchmark models for future research to be calibrated against in terms of predictive ability. In the truest sense, a naïve model would involve a time series model with just lagged dependent variables. However, I did not test such models here, due to the lack of support for the length of lag. Further research needs to be conducted to establish lengths of lags prior to using time series models as benchmarks.

The models shown in this research demonstrate common variables regardless of critical modeling issues namely unobserved effects or heterogeneity, correlation due to shared unobserved effects through accident counts and excess zero problems. However, the data used in this research is somewhat limited in its coverage of geographic, environmental and geometric effects. The question raised from this research is that "do the common variables still robust in the models if the models are developed based upon data with geography, environment and geometric design different from those in Washington State?". The possible answer is all common variables would remain high level of significance in the model. However, the level of significance of constant or intercept in the model could varies because unobserved effects as major modeling issue that do not exist in Washington State, for example weather in the dessert area, are subsumed in the constant. It is fair to say that the common variables developed from this research provide the reasonable range to any kind of geography, environment and geometric design in the median crossover accident context. As future research, one can use common variables developed by this study as naïve specifications to explore median crossover accident frequency. Furthermore, a more thorough examination of crossover accidents on national level where geographic, environmental and geometric condition vary significant may reveal other unique insights into factors contributing to median crossover accident.

7.2 Institutional and Policy Conclusions and Recommendations

From an agency standpoint, the models developed in this dissertation are portable. The predictions can be performed in spreadsheet format so state decision makers and analysts can use post-processed model information to estimate median crossover probabilities on the network. Since the entire network was used in the development of the models, the portability of the findings of this dissertation is complete. From a design policy standpoint, several variables were found to be significant in their correlation with median crossover accidents. Most importantly, the magnitude of the 40 to 60 foot median width variable is suggestive. Up until recently, the WSDOT used less than 30 foot median widths as basic median barrier requirement widths. Recently, some empirical work conducted by WSDOT staff suggested that the barrier width requirement be extended to 50 feet. While such a change in design policy was not established on the basis of statistical models, this magnitude of the 40 to 60 foot median width variable appears to support that change from a frequency standpoint. However, more analysis is required in terms of benefit cost as well as severity of crossovers to establish complete support for a move to a 50-foot or higher median barrier width requirement. Furthermore, this issue is complicated by the lack of geographical consistency in median barrier width requirements. In this sense, the contribution of this dissertation is in providing important direction in terms of establishing data and modeling protocols for the development of sound design policy.

References

American Association of State Highway and Transportation Officials (1996) *Roadside Design Guide*.

Bronstad, M.E., Calcote, L.R., and Kimball, C.E. (1976) Concrete Median Barrier (Research. Report, FHWA-RD-73-3) Federal Highway Administration, U.S. Department of Transportation.

Cameron, A.C. and Trivedi (1998) P.K. *Regression Analysis of Count Data*. First Edition, Cambridge University Press, New York.

Congdon, P. (2005) *Bayesian Models for Categorical Data*. First Edition, John Wiley&Sons Ltd., England.

Congdon, P. (2003) *Applied Bayesian Modelling*. First Edition, John Wiley&Sons Ltd., England.

Chayanan, S., Shankar, V.N., Sittikariya, S., Ulfarsson, G.F., Shyu, M.B. and Juvva, N.K. (2004) Median crossover accident analyses and the effectiveness of median barriers. (Final Research Report, WA-RD 580.1) Washington State Department of Transportation, Washington.

ESRI (2002) *ArcView GIS Version 3.2*, Environmental Systems Research Institute, California.

Glad, R.C., Albin, R., Macintosh, D. and Olson, D. (2002) Median treatment study on Washington State Highway. (Final Research Report, WA-RD 516.1) Washington State Department of Transportation, Washington.

Graf, V.D. and Winegard, N.C. (1968) Median Barrier Warrants. Traffic Department of the State of California, California.

Greene, W.H. (2004) *LIMDEP Version 8.0, User's Manual*. Econometric Software Inc., Bellport, New York.

Greene, W.H. (2003) *Econometric Analysis*. Fifth Edition, Prentice Hall, New Jersey.

Greene, W.H. (1994) Accounting for excess zeros and sample selection in Poisson and negative binomial regression models (Working papers EC-94-10) Stern School of Business, New York University, New York.

Gourieroux, C. and Monfort, A. (1997) *Simulation Based Econometric Methods*, Oxford University Press, Oxford.

Guo, G. (1996) Negative multinomial regression models for clustered event counts. *Sociological Methodology* **26**, 113–132.

Hausman, J., Hall, B. and Griliches, Z. (1984) Econometric-models for count data with an application to the patents R and D relationship. *Econometrica* **52(4)**, 909-938.

Hinde, J. (1982) Compound Poisson Regression Models. *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*, Springer-Verlag, New York.

Kennedy, P. (1998) *A Guide to Econometrics*. Fourth Edition, The MIT Press, Cambridge, Massachusetts.

Lancaster, T. (2004) *An Introduction to Modern Bayesian Econometrics*. First Edition, Blackwell Publishing Ltd, England.

Milton, J.C. and Mannering, F.L. (1996) The relationship between highway geometrics, traffic related elements and motor vehicle accidents. (Final Research Report, WA-RD 403.1) Washington State Department of Transportation, Washington.

Poch, M. and Mannering, F.L. (1996) Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering* **122(3)**, 105-113.

Ross Jr., H.E. (1974) Impact Performance and Selection Criterion for the Texas Median Barriers. (Final Research Report, 140-8) Texas Transportation Institute, Texas A&M University.

Seamons, L. L., and Smith, R.N. (1991). Past and Current Median Barrier Practice in California. (Final Research Report, CALTRANS-TE-90-2) California Department of Transportation, California

Shankar, V.N., Chayanan, S., Sittikariya, S., Shyu, M.B., Juvva, N.K. and Milton J.C. (2004) The marginal impacts of design, traffic, weather, and related interactions on roadside crashes. *Transportation Research Record* **1897**, 156-163.

Shankar, V.N., Albin, R.B., Milton, J.C. and Mannering, F.L. (1998) Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. *Transportation Research Record* **1635**, 44–48.

Shankar, V.N., Milton, J.C. and Mannering, F.L. (1997) Modeling statewide accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis and Prevention* **29(6)**, 829–837.

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

Ulfarsson, G. F. and Shankar V.N. (2003) An accident count model based on multi-year cross-sectional roadway data with serial correlation. *Transportation Research Record* 1840, 193-197.

Voung, Q. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57(2)**, 307-333.

Western Regional Climate Center (2002) *Historical Climate Information*, Desert Research Institute, Retrieved from http://www.wrcc.dri.edu/summary/climsmwa.html.

Appendix

Descriptive Statistics of Key Median Crossover Accident Related Variables

Dependent Variables

Variable	Mean	Std. Error	Min	Max
The number of crossover accidents in section	0.2407	0.6448	0.00	7.00

Traffic Variables

Variable	Mean	Std. Error	Min	Max
Average AADT (weighted)	37,354.6291	36,974.9664	3,347.00	172,557.00
AADT per lane	7,443.8783	5,828.7465	836.75	28,688.33
Natural logarithm of AADT per lane	8.6306	0.7616	6.73	10.26
Per-lane AADT indicator 1 (1 if per-lane AADT <= 5,000 vehicles, 0 otherwise)	0.4691	0.4992		
Per-lane AADT indicator 2 (1 if per-lane AADT > 5,000 vehicles and <= 10,000 vehicles, 0 otherwise)	0.2873	0.4527		
Per-lane AADT indicator 3 (1 if per-lane AADT > 10,000 vehicles, 0 otherwise)	0.2436	0.4294		
Single truck percentage	4.1960	1.2150	1.90	10.00
Double truck percentage	7.7623	4.6205	0.55	17.80
Truck-train percentage	2.2050	1.5970	0.00	7.00
Total truck percentage	14.1634	6.6821	3.20	32.00
Total truck percentage indicator 1(1 if percentage of total trucks <= 5%, 0 otherwise)	0.0327	0.1780		
Total truck percentage indicator 2 (1 if percentage of total trucks > 5% and < 15%, 0 otherwise)	0.5345	0.4990		
Total truck percentage indicator 3 (1 if percentage of total trucks >=15%, 0 otherwise)	0.4327	0.4956		
Percentage of AADT in the peak hour	11.1158	3.0922	7.30	19.40
Peak hour indicator 1 (1 if percentage of AADT in peak hour <= 9%, 0 otherwise)	0.2691	0.4436		
Peak hour indicator 2 (1 if percentage of AADT in peak hour > 9% and <= 13%, 0 otherwise)	0.5345	0.4990		
Peak hour indicator 3 (1 if percentage of AADT in peak hour > 13%, 0 otherwise)	0.1964	0.3974		

Roadway Geometric Variables

Variable	Mean	Std. Error	Min	Max
Collector indicator (1 if the section is collector , 0 otherwise)	0.0073	0.0850		
Principal arterials indicator (1 if the section is principal arterial, 0 otherwise)	0.3055	0.4608		
Interstate indicator (1 if the section is interstate, 0 otherwise)	0.6872	0.4638		
Length of the roadway section in miles	2.4297	2.6899	0.50	19.30
Average number of lanes	4.6036	1.1315	2.00	8.00
Average roadway width	57.4182	15.4683	24.00	121.00
Average speed limit in mph	59.6727	5.5026	35.00	65.00
Speed limit indicator 1 (1 if speed limit < 55 mph, 0 otherwise)	0.0255	0.1576		
Speed limit indicator 2 (1 if speed limit >= 55 mph, 0 otherwise)	0.9745	0.1576		
The number of interchanges in section	0.8473	0.8350	0.00	4.00
Tangent length in miles	0.8421	1.2083	00.00	10.20
Minimum horizontal curve length in feet	1,047.9055	983.0674	0.00	6,438.00
Minimum horizontal central angle in degrees	13.4311	17.7347	0.00	111.49
Maximum horizontal central angle in degrees	30.2916	23.8828	0.00	111.49
The number of horizontal curves in section	2.7491	2.8612	0.00	29.00
Horizontal curve indicator (1 if the number of horizontal curves >0, 0 otherwise)	0.9091	0.2876		
The number of horizontal curves per mile	1.4400	0.9600	0.00	5.00
The number of horizontal curves per mile indicator 1 (1 if the number of horizontal curves per mile <= 0.5, 0 otherwise)	0.1600	0.3667		
The number of horizontal curves per mile indicator 2 (1 if the number of horizontal curves per mile > 0.5 and <= 2.5, 0 otherwise)	0.7127	0.4527		
The number of horizontal curves per mile indicator 3 (1 if the number of horizontal curves per mile > 2.5, 0 otherwise)	0.1273	0.3334		
Minimum radius of horizontal curve in feet	4,267.2364	4,875.0795	0.00	38,400.00
Absolute value of minimum grade in percents	0.6015	0.9174	0.00	5.00
Absolute value of maximum grade in percents	2.7956	1.2795	0.20	6.72
The number of grade changes	3.8655	4.0887	0.00	28.00
The number of grade changes per mile	1.8868	1.6935	0.00	20.00
The number of grade changes per mile indicator 1 (1 if the number of grade changes per mile <= 1, 0 otherwise)	0.2255	0.4180		
The number of grade changes per mile indicator 2 (1 if the number of grade changes per mile >1 and <= 2, 0 otherwise)	0.4291	0.4951		
The number of grade changes per mile indicator 3 (1 if the number of grade changes per mile >2 and <=3, 0 otherwise)	0.2000	0.4001		
The number of grade changes per mile indicator 4 (1 if the number of grade changes per mile >3, 0 otherwise)	0.1455	0.3527		
Average yearly pavement friction	46.8181	5.6273	20.00	61.50
Pavement friction indicator 1 (1 if average yearly pavement friction <= 30, 0 otherwise)	0.0015	0.0381		
Pavement friction indicator 2 (1 if average yearly pavement friction > 30 and <50, 0 otherwise)	0.6844	0.4649		
Pavement friction indicator 3 (1 if average yearly pavement friction .>= 50, 0 otherwise)	0.3142	0.4644		

Median Variables

Variable	Mean	Std. Error	Min	Max
Median indicator (1 if median exists in the section, 0 otherwise)	0.9055	0.2927		
Percent medians <= 30 feet	0.0473	0.2123		
Percent medians > 30 feet and <= 40 feet	0.2764	0.4474		
Percent medians > 40 feet and <= 50 feet	0.1164	0.3208		
Percent medians > 50 feet and <= 60 feet	0.0582	0.2342		
Percent medians > 60 feet	0.5018	0.5002		
Minimum median shoulder width in feet	4.4836	1.6833	0.00	10.00
Maximum median shoulder width in feet	5.3055	2.4919	0.00	18.00
Shape of median indicator 1 (1 if shape of median is flat, 0 otherwise)	0.6436	0.4791		
Shape of median indicator 2 (1 if shape of median is fixed slope, 0 otherwise)	0.1673	0.3734		
Shape of median indicator 3 (1 if shape of median is concave, 0 otherwise)	0.1127	0.3164		
Shape of median indicator 4 (1 if shape of median is convex, 0 otherwise)	0.0727	0.2598		
Slope of median indicator 1 (1 if slope of median is flat, 0 otherwise)	0.6436	0.4791		
Slope of median indicator 2 (1 if slope of median is slight, 0 otherwise)	0.1964	0.3974		
Slope of median indicator 3 (1 if slope of median is medium, 0 otherwise)	0.1273	0.3334		
Slope of median indicator 4 (1 if slope of median is steep, 0 otherwise)	0.0291	0.1681		
Type of median indicator 1 (1 if type of median is paved, 0 otherwise)	0.0436	0.2044		
Type of median indicator 2 (1 if type of median is sand or gravel, 0 otherwise)	0.0582	0.2342		
Type of median indicator 3 (1 if type of median is low grass, 0 otherwise)	0.8800	0.3251		
Type of median indicator 4 (1 if type of median is high grass (> 3 feet), 0 otherwise)	0.0145	0.1198		

Weather Variables

Variable	Mean	Std. Error	Min	Max
Average monthly precipitation in inches	2.4914	1.8215	0.38	10.98
Average monthly snow depth in inches	1.2625	3.5545	0.00	54.33
Monthly precipitation indicator 1 (1 if average monthly precipitation <= 1.5 inches, 0 otherwise)	0.3891	0.4877		
Monthly precipitation indicator 2 (1 if average monthly precipitation > 1.5 inches and <= 4 inches, 0 otherwise)	0.4356	0.4960		
Monthly precipitation indicator 3 (1 if average monthly precipitation > 4 inches, 0 otherwise)	0.1753	0.3803		
Monthly snow indicator 1 (1 if average monthly snow depth <= 1 inches, 0 otherwise)	0.6975	0.4595		
Monthly snow indicator 2 (1 if average monthly snow depth > 1 inches, 0 otherwise)	0.3025	0.4595		

Traffic Interaction Variables

Variable	Mean	Std. Error	Min	Max
The product of percentage of single trucks and Average AADT (weighted)	139,838.4250	114,909.5980	17,234.80	538,418.10
The product of percentage of double trucks and Average AADT (weighted)	219,727.2520	170,662.5530	15,517.64	750,522.20
The product of percentage of total trucks and Average AADT (weighted)	416,459.5270	307,086.2220	54,890.80	1,403,150.20

Design Interaction Variables

Variable	Mean	Std. Error	Min	Max
Interaction between horizontal curve and posted speed > 45 mph (1 it the number of horizontal curves > 0 and posted speed >= 45 mph, 0 otherwise)	0.9018	0.2977		
Interaction between horizontal curve and posted speed > 50 mph (1 it the number of horizontal curves > 0 and posted speed >= 50 mph, 0 otherwise)	0.8982	0.3025		
Interaction between horizontal curve and posted speed > 55 mph (1 it the number of horizontal curves > 0 and posted speed >= 55 mph, 0 otherwise)	0.8909	0.3119		
Interaction between horizontal curve and intersection (1 if the number of horizontal curves > 0 and the number of interchanges > 0, 0 otherwise)	0.5673	0.4956		
Interaction between the number of horizontal curves per mile and pavement friction (1 if the number of horizontal curves per mile > 2.5 and pavement friction >= 50, 0 otherwise)	0.0356	0.1854		
Interaction between the number of horizontal curves per mile and pavement friction (1 if the number of horizontal curves per mile <= 0.5and pavement friction <= 30, 0 otherwise)	0.0007	0.0270		

Traffic and Design Interaction Variables

Variable	Mean	Std. Error	Min	Max
Interaction between total truck percentage and posted speed (1 if truck percentage >= 15% and posted speed > 55 mph, 0 otherwise)	0.3964	0.4893		
Interaction between total truck percentage and the number of curves per mile (1 it truck percentage <= 5% and the number of curves per mile <= 0.5, 0 otherwise)	0.0036	0.0602		
Interaction between total truck percentage and the number of curves per mile (1 it truck percentage >= 15% and the number of curves per mile >2.5, 0 otherwise)	0.0364	0.1873		
Interaction between total truck percentage and pavement friction (1 if truck percentage >= 15% and pavement friction <= 30, 0 otherwise)	0.0000	0.0000		
Interaction between total truck percentage and pavement friction (1 if truck percentage <= 5% and pavement friction >= 50, 0 otherwise)	0.0073	0.0850		

Weather and Traffic Interaction Variables

Variable	Mean	Std. Error	Min	Max
Interaction between average monthly precipitation and total truck percentage (1 if average monthly precipitation > 2.5 inches and total truck percentage > 15%, 0 otherwise)	0.1287	0.3350		
Interaction between average monthly snow depth and total truck percentage (1 if average monthly snow depth > 1.0 inches and total truck percentage > 15%, 0 otherwise)	0.1658	0.3721		

Weather and Design Interaction Variables

Variable	Mean	Std. Error	Min	Max
Interaction between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation <= 1.5 inches and the number of horizontal curves per mile <= 0.5, 0 otherwise)	0.0909	0.2876		
Interaction between average monthly precipitation and the number of horizontal curves per mile (1 if average monthly precipitation > 4.0 inches and the number of horizontal curves per mile > 0.5, 0 otherwise)	0.1607	0.3674		
Interaction between average monthly precipitation and the number of horizontal curves (1 if average monthly precipitation > 0 and the number of curves > 0, 0 otherwise)	0.9091	0.2876		
Interaction between average monthly snow depth and the number of horizontal curves (1 if average monthly snow depth > 0 and the number of curves > 0, 0 otherwise)	0.5927	0.4915		
Interaction between average monthly precipitation and pavement friction (1 if average monthly precipitation > 2.5 inches and pavement friction <= 30, 0 otherwise)	0.0015	0.0381		
Interaction between average monthly precipitation and posted speed (1 if average monthly precipitation > 2.5 inches and posted speed > 55 mph, 0 otherwise)	0.1724	0.3778		

Design and Median Interaction Variables

Variable	Mean	Std. Error	Min	Max
Length of the roadway section on median widths less than or equal to 40 feet	0.6800	1.5024	0.00	12.40
Length of the roadway section on median widths greater than or equal to 41 feet and less than or equal to 60 feet	0.2782	0.7845	0.00	5.69
Length of the roadway section on median widths greater than 60 feet	1.4715	2.7502	0.00	19.30

Vita

Sittipan Sittikariya received his Bachelor's degree in Civil Engineering from Chulalongkorn University, Thailand, in 1999 and he also earned a Master's degree in Transportation Engineering from the same university in 2001. In 2002, he joined the University of Washington, Seattle, as a PhD student in transportation engineering program and later on in 2004 he transferred to the Pennsylvania State University. He served as a lead researcher in the Transportation Infrastructure Modeling Group (TIMG) and as a teaching assistant on several transportation engineering classes in undergraduate and graduate levels offered at two major universities in the U.S., namely the Pennsylvania State University and the University of Washington. In 2006, he earned a Doctor of Philosophy at the Pennsylvania State University in Transportation Engineering. He has published peer-reviewed journal articles and conference and symposium proceedings in the area of transportation safety, pedestrian safety, infrastructure investment planning and design policy as well as travel demand modeling. In the United State, he has significant exposure to real-world transportation work experience through projects he has worked for private consulting firms, transportation institutes, department of transportation and universities. In addition, he has gained international work experience in transportation area from Japan and Thailand. Currently, he served as a lead transportation engineer in DKS Associates in Seattle.