IMPROVING SAFETY ESTIMATION AND PREDICTION USING MULTIVARIATE REGRESSION MODELS IN OBSERVATIONAL ROAD SAFETY STUDIES

by

Anusha Musunuru

A dissertation submitted to the faculty of
The University of Utah
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Civil and Environmental Engineering

The University of Utah

May 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Anusha Musunuru**

has been approved by the following supervisory committee members:

Richard J. Porter	, Chair	**03/02/2017** Date Approved
Xiaoyue Cathy Liu	, Member	**03/08/2017** Date Approved
Ran Wei	, Member	**03/06/2017** Date Approved
Juan Medina	, Member	**03/06/2017** Date Approved
Milan Zlatkovic	, Member	**02/28/2017** Date Approved

and by **Michael E. Barber** , Chair of

the Department of **Civil and Environmental Engineering**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Observational studies are a frequently used "tool" in the field of road safety research because random assignments of safety treatments are not feasible or ethical. Data and modeling issues and challenges often plague observational road safety studies, and impact study results. The objective of this research was to explore a selected number of current data and modeling limitations in observational road safety studies and identify possible solutions.

Three limitations were addressed in this research: (1) a majority of statistical road safety models use average annual daily traffic (AADT) to represent traffic volume and do not explicitly capture differences in traffic volume patterns throughout the day, even though crash risk is known to change by time of day, (2) statistical road safety models that use AADT on the "right-hand side" of the model equation do not explicitly account for the fact that these values for AADT are estimates with estimation errors, leading to potential bias in model estimation results, and (3) the current state-of-the-practice in road safety research often involves "starting over" with each study, choosing a model functional form based on the data fit, and letting the estimation results drive interpretations, without fully utilizing previous study results.

These limitations were addressed by: (1) estimating the daily traffic patterns (by time of day) using geo-spatial interpolation methods, (2) accounting for measurement error in AADT estimates using measurement error models of expected crash frequency, and (3) incorporating prior knowledge on the safety effects of explanatory variables into regression models of expected crash frequency through informative priors in a Bayesian methodological framework. These alternative approaches to address the selected observational road safety study limitations were evaluated using data from rural, two-lane highways in the states of Utah and Washington. The datasets consisted of horizontal curve segments, for which crash data, roadway geometric features, operational characteristics, roadside features, and weather data were obtained.

The results show that the methodological approaches developed in this research will allow road safety researchers and practitioners to accurately evaluate the expected road safety effects. These methods can further be used to increase the accuracy and repeatability of study results, and ultimately expand the current practice of evaluating regression models of expected crash frequency in observational road safety studies.

To my parents, Venkateswara Rao and Revathi Laxmi,

for teaching me the value of education in all its forms

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

ACRONYMS

WHO	World Health Organization
NHTSA	National Highway Traffic Safety Administration
EB	Empirical Bayes
SPF	Safety Performance Function
UDOT	Utah Department of Transportation
NB	Negative Binomial
RHS	Right Hand Side
AADT	Annual Average Daily Traffic
ATR	Automatic Traffic Recorder
OLS	Ordinary Least Squares
RCAL	Regression Calibration
SIMEX	Simulation Extrapolation
MCMC	Markov Chain Monte Carlo
DIC	Deviance Information Criterion
HSIS	Highway Safety Information System
RFIP	Roadside Features Inventory Program
NOAA	National Oceanic and Atmospheric Administration
LiDAR	Light Detection and Ranging
FI	Fatal and Injury
CMF	Crash Modification Factor

ACKNOWLEDGMENTS

I could not have completed this work without the support and guidance I received throughout its development from so many people, and for that I am very grateful. I especially appreciate the unwavering support and mentorship I have received over the years from my advisor, Dr. Richard J. Porter, who taught me so much about Transportation Engineering, Statistics, and the research process. I would like to express sincere thank you to my committee member, Dr. Cathy Liu, for her support, guidance, and encouragement over these years.

Thank you also to the other members of my committee – Dr. Milan Zlatkovic, Dr. Ran Wei, and Dr. Juan Medina – for their guidance and timely feedback on my research. An additional thank you goes to Brian Ray, for being a supportive mentor, who consistently cared about my progress, and contributed immensely to my personal and professional growth.

I gratefully acknowledge the funding sources that made my graduate work possible. My work was supported by the Federal Highway Administration, Mountain Plains Consortium, Utah Department of Transportation, and the Department of Civil & Environmental Engineering at the University of Utah. Thank you also to my fellow graduate students in the Traffic Lab; I truly appreciate all of the encouragement, advice, entertainment, and assistance. I am truly grateful for the time spent with friends here, and back at home, and for all the memories. Thank you for making me realize that there is life beyond work.

Finally, I would like to thank my family - where the basic source of my life energy resides. To my parents, who raised me with a love for science, supported, and encouraged me in all my pursuits; I thank you for your love and patience more than you will ever know. To my sisters, Sameera Musunuru and Pragathi Musunuru - who have always inspired me, and kept me grounded; thank you for always believing in me. To Neel, Ridhima, and Aayush - I love you and I look forward to accompanying you along your own journey of finding your place in the world. Thank you for bringing happiness to our lives. And, to my loving and encouraging husband, Akshay Chitneni, whose support during the final stages of this Ph.D. is very much appreciated. Thank you for everything.

CHAPTER 1

INTRODUCTION

The World Health Organization (WHO) has reported that approximately 1.3 million people are killed in traffic crashes every year, and between 20 and 50 million suffer from nonfatal injuries (*1-2*). These numbers suggest that traffic crashes represent a major public health burden on a global scale. By the year 2030, traffic crashes are projected to reach fifth place among leading causes of death in the world (*1*,*3*). Traffic crashes can have devastating and long-lasting consequences for those involved and their families. The overall economic and societal impact costs of traffic crashes in the United States totaled approximately $242 billion and $836 billion, respectively, in 2010 (*4*). Given the magnitude of these impact costs, there is a need to identify and take effective actions to prevent and mitigate the negative outcomes of traffic crashes and improve road safety.

Recently released U.S. figures from the National Highway Traffic Safety Administration (NHTSA) for the year 2013 show a 3.1 percent decrease in motor vehicle crash fatalities from the previous year and nearly a 25 percent decline in overall highway fatalities since 2004 (*5-6*). Similar to fatality trends, the number of people injured fell from 2,500,000 to 2,313,000 in the years 2006 to 2013 (*7*). Despite increasing traffic on the roadways, traffic fatalities and injuries have stabilized and even diminished in the recent years, specifically since 2006. Several studies have associated this decreasing trend in traffic fatality and injury numbers to be the result of various factors, including changes in roadway design, improved vehicle design, advances in medical treatment, laws to reduce drunk driving and increase seat belt use, better-trained novice drivers, economic trends, and increased funding for safety infrastructure improvements (*8*-*13*). Concerning the possible contribution of road safety research to that progress, it was demonstrated by Elvik et al. (2009) that road safety measures based on the findings of research projects have made major contributions to reducing the number of traffic fatalities and injuries in the best performing countries (*14*).

Road safety research started more than 80 years ago to address the practical information needs identified as a result of increasing numbers of traffic fatalities and injuries (*3*). One of the early documented research studies on accident proneness was carried out in the mid- to late-1920s (*15*-*16*). Over the years, the changing environment of the road transport system, vehicle technology, and the understanding gained about combinations of factors that contribute to crash occurrence resulted in exploring different avenues to road safety research. These include human factors research (e.g., driver simulation, test track, and naturalistic driving studies), behavioral research, vehicle research and testing, vehicle crashworthiness, advanced vehicle technologies, event data recorders, observational road safety studies, and statistical road safety modeling, to name a few (*17*).

Observational studies serve as a major source of knowledge for researchers and other decision makers on the expected road safety effects of highway and traffic engineering decisions. Despite methodological challenges, high-quality observational studies have emerged and played an important role in informing road safety management practices. Using retrospective and observational data, researchers seek to estimate the safety effect of a treatment by observing the same units (e.g., drivers, vehicles of some type, road segments, or intersections, to name a few) before and after the treatment or by comparing units found to have one treatment with other units that have a different treatment (*18*).

Observational Road Safety Studies

Observational studies are needed when random assignments of units to treatment or control groups are not feasible or ethical. The mechanism of treatment or control assignment in observational studies is dictated either by choice of individual or some other factors. Therefore, the treated and control groups may have differed prior to receiving the treatment in ways that are relevant to outcomes of interest (*19-22*). Post-treatment differences may possibly be attributed to the treatment effect, these pretreatment differences, or both.

Pretreatment differences (or selection biases) in observational studies are of two kinds: 1) those that have been accurately measured, called overt biases, and 2) those that have not been measured but are suspected to exist or that are not suspected at all, called hidden biases (*19-20*). Methods that account for overt biases and address uncertainty in hidden biases are proposed in the literature and are intended to support the possibility of making valid causal inferences in observational studies (*23-25*). There are two types of observational studies, before-after study designs and cross-sectional study designs.

Before-After Study Designs

Before-after road safety studies involve a treatment applied to a group of units at some point in time. Treatment effects are estimated by analyzing available data collected for the treated units (and sometimes for another set of untreated units) from before and after the treatment (*26-27*). Before-after studies involve two basic tasks: 1) predicting what would have been the safety of the treated entities in the 'after' period had treatment not been applied, and 2) estimating what the safety of the treated entities in the 'after' period was (*27*).

There are several methods to accomplish these two tasks when evaluating treatment effects using before-after data; however, the empirical Bayes method is considered the current state-of-the-practice (*27-28*). It accounts for regression-to-the-mean bias, changes in traffic volume, and also appropriately evaluates sites with zero crash counts during the analysis period (*26,29-31*). In summary, it can be said that, based on evidence from actual studies and empirical data, the empirical Bayes methodology produces results that are more valid than those produced by more traditional approaches (*32*).

All methods of accomplishing the prediction task of before-after studies (i.e., task 1) consist of two consecutive steps: 1) estimating the expected number of crashes in the before period, and 2) using this and other information to predict how the expected number of crashes would have changed from the before to after period on the treated entities, if the treatment would not have been applied. The predicted result of "what would have been" is then compared with the expected number of crashes in the after period with the treatment (i.e., task 2) to evaluate the safety effect of the treatment (*27*).

A brief description of the empirical Bayes approach to estimating the expected number of crashes in the before period is provided as an example. According to this approach, the best estimate of expected crash frequency on a certain entity (road segment, intersection, etc.) in the "before" period is obtained by combining two sources of information: (i) crash counts for that entity and (ii) information about the expected number of crashes on other similar entities (also known as reference group) (*33*). The expected number of crashes on these similar entities (i.e., reference group) is most commonly estimated using multivariate regression models. These multivariate regression models, also known as safety performance functions (SPFs), commonly provide an estimate of the expected number of crashes per some time period for a road segment or intersection as a function of traffic and road geometric characteristics. Time trends that capture safety effects of other factors not measured or understood over time are also estimated using data from the reference group and can be incorporated directly into the regression model to predict "what would have been" for the treated entities had the treatment not been applied. This approach assumes that the time trends observed at the reference group are the same as what the time trends would have been for the treatment group without the treatment.

Since crashes are count outcomes, and the variance of crash counts is almost always greater than the mean (or overdispersion), negative binomial (NB) regression is commonly used to model the expected number of crashes as a function of traffic, geometric, and other explanatory, right-hand side variables (i.e., the SPF). The parameters of a negative binomial regression model are estimated using maximum likelihood, constructed based on the assumptions that the observations are independent. As an example, a negative binomial regression model of the expected number of crashes on a road segment is written as shown in the following equation:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \gamma_1 t_1 + \gamma_2 t_2 + \cdots + \gamma_k t_k + \varepsilon_i) \quad \text{Equation (1)}$$

where:

μ_i = dependent variable, expected number of crashes for a roadway segment i;

x_{ij} = independent, explanatory variables specifying traffic, geometric, and other characteristics of a roadway segment i;

t_k = yearly indicators for the “unmeasured” independent variables that change with time;

β_j = regression parameters to be estimated that quantify the relationships between the explanatory variables and μ_i;

γ_k = regression coefficient parameters for yearly indicators;

β_0 = intercept;

ε_i = random error term, where exp (ε_i) is a gamma-distributed error term with mean 1 and variance α. In the NB-2 model, the variance in the number of crashes is written as $\mu_i + \alpha\mu_i^2$, with α referred to as the dispersion parameter.

The expected number of crashes for the before period at other similar entities, estimated by SPFs, and the observed number of crashes for the treated entities in the before period are used to estimate the expected number of crashes for the before period at the treated entities. Characteristics of road segments, including traffic volumes and patterns, weather, vehicles, and drivers, change with time. These trends are captured by SPFs both implicitly using the yearly indicators (as shown in the equation above) and explicitly by including the variables that change with time in the model, traffic volume being the most common. The yearly indicator coefficients are estimated along with other regression coefficients. What would have been the expected crash frequency in the after period at the treatment sites without the treatment is then calculated by taking the ratio of predicted values from the SPFs in the after period to the predicted values from the SPFs in the before period, and then multiplying this ratio by the estimate of the expected number of crashes at the treated entities in before period (*27,33*). These "what would have been" estimates are then compared with estimated number of crashes in the after period with treatment to evaluate the safety effect of the treatment. Figure 1 illustrates an observational before-after study evaluation using the Empirical Bayes method.

There are several practical limitations associated with before-after study designs (*26,34*):

- Confounding Factors: There may be several improvements, in addition to the treatment of interest, that are implemented over the years. For example, when a road is rebuilt, several of its attributes are changed, and before-after studies are not practically feasible for the evaluation of many specific types of individual changes. Estimates would reflect the combined effects of all changes in these cases.
- Data and Duration: Several years of before and after period data are generally needed for an adequate sample of sites to handle the method with adequate statistical power. This data collection can be time consuming, and waiting several years after implementation is a practical concern in these types of studies.

Ezra Hauer stated that "….opportunities to do observational before-after studies about, say, the safety effect of change in horizontal curvature, grade, lane width, etc. are few and imperfect" (*18*). For example, there may be a few projects where the degree of horizontal curvature is changed from 10 degrees to 5 degrees, yet there are many horizontal curves with degree of curvature 10 degrees and 5 degrees. In this case, cross-sectional studies are an option. In cross-sectional studies, a safety comparison is made between one group of entities having some common feature (degree of curvature equal to 10 in this case) and a different group of entities not having that feature (degree of curvature not equal to 10 or equal to 5).

In this case, one cannot know the change of safety from before to after of the same entities. Instead, the safety of two different groups of entities is compared. Given the limitations associated with observational before-after studies in terms of limited opportunities to perform them for some treatments of interest and insufficient data, observational cross-sectional studies can be very useful, but they also bring with them additional challenges.

Cross-Sectional Study Designs

Unlike before-after studies, cross-sectional studies do not require data from treated entities both before and after the treatment to evaluate that treatment's safety effects. Instead, safety is compared at sites with and without the treatment of interest (*26*,*29*). In other words, in a before-after study, 'treatment' refers to something that has actually changed from before to after period on the treated entities. In a cross-sectional study, there is no such change; there is a difference between the units evaluated in terms of the treatment (*18*).

Figure 2 illustrates the cross-sectional evaluation of a treatment. From the figure, the difference between the blue triangle and red triangle represents the estimated safety effect of the treatment. In practice, it is very difficult to find sites that are similar enough to each other, except for the treatment of interest, to avoid confounding influences. Hence, cross-sectional studies often use multivariate regression models.

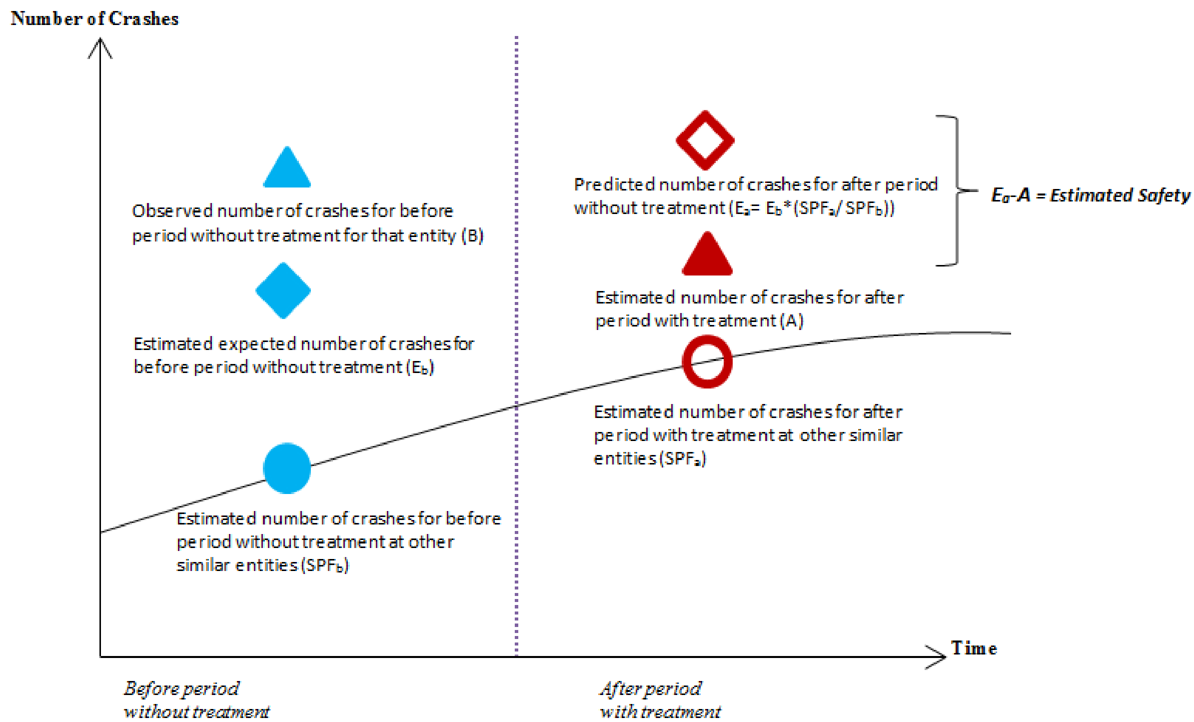


Figure 1 Before-After Study Evaluation Using the Empirical Bayes Method (Adapted from FHWA, 2010)

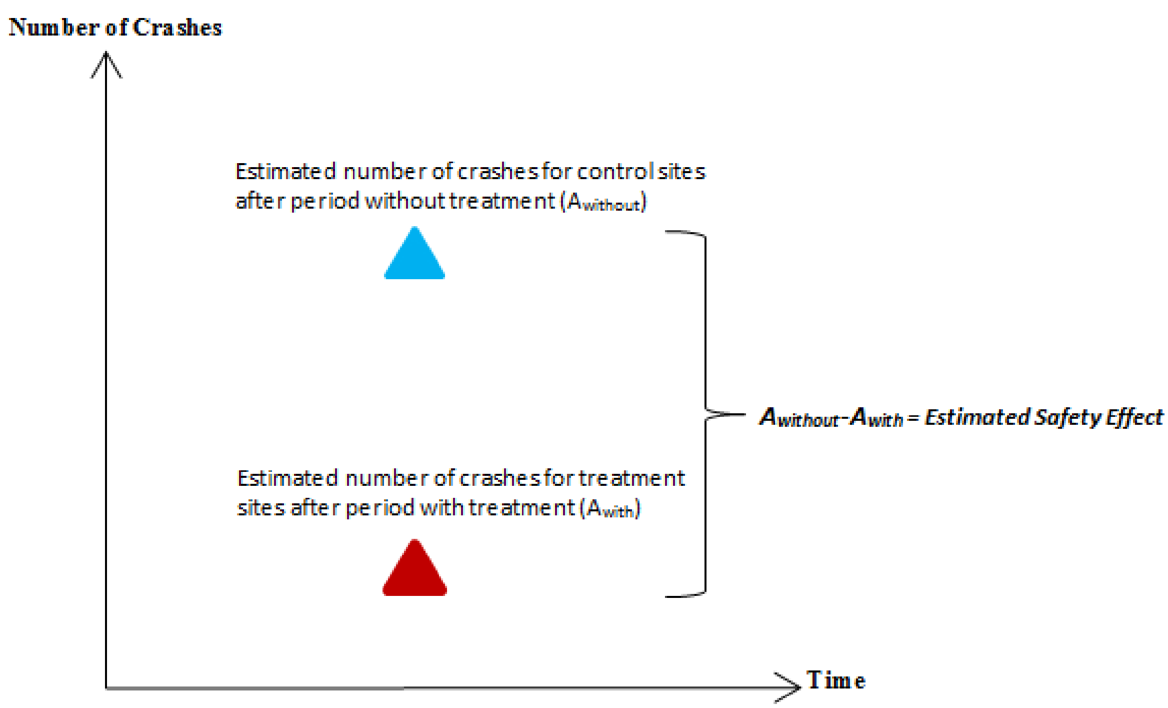


Figure 2 Cross-Sectional Study Evaluation (Adapted from FHWA, 2010)

Treatment effects are derived using the estimated regression coefficients for the variables representing the treatment of interest, while simultaneously accounting for other variables that also impact safety. As with SPFs, NB regression modeling is the most common approach used to model the expected number of crashes as a function of explanatory variables, including the variables that represent the treatment. In this context, the NB regression model is expressed as given in the following equation:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \beta_c z_{ic} + \varepsilon_i) \qquad \text{Equation (2)}$$

where:

μ_i = dependent variable, expected number of crashes for a roadway segment *i*;

x_{ij} = independent, explanatory variables specifying traffic, geometric, and other characteristics of a roadway segment *i*;

z_{ic} = independent variable that corresponds to the treatment of interest;

β_j = regression parameters to be estimated that quantify the relationships between the explanatory variables and μ_i;

β_c = regression coefficient parameter for treatment-related variable;

β_0 = intercept; and

ε_i = random error term, where exp (ε_i) is a gamma-distributed error term with mean 1 and variance α. In the NB-2 model, the variance in the number of crashes is written as $\mu_i + \alpha{\mu_i}^2$, with α referred to as the dispersion parameter.

The above equation is used to estimate the direction and magnitude of the safety effects of the treatment related variable, in addition to estimating the expected number of crashes. The direction of the regression coefficient parameter for treatment-related variable, β_c is used to determine whether the treatment has a positive or negative effect on reducing the expected number of crashes. If $\beta_c < 0$, the treatment is said to have a positive effect on reducing the expected number of crashes, and vice versa.

The magnitude of the regression coefficient parameter (i.e., β_c) for the treatment-related variable is used to estimate the actual change caused by the treatment. For example, if the value of the regression coefficient estimate is β_c (β_c<0) for the treatment related variable (i.e., z_{ic}), the expected number of crashes decreases by β_c times, if the treatment is applied, while all the other variables are constant. However, the regression models in these studies are also used to determine the safety effects of the right-hand side variables on expected number of crashes. For example, if the value of the regression coefficient estimate is β_1 (β_1 >0) for an explanatory variable, x_{i1}, the expected number of crashes increase by β_1 times, if the value of the explanatory variable increases by one unit, given all the other variables are constant.

The major challenge with using cross-sectional studies is that the safety effect estimates may not represent the safety effects at all, but instead be an artifact of factors that were measured but could not be included in the model, factors that could not be measured, factors that were unknown, and/or the choice of functional form for explanatory variables included in the model specification (*35*).

A leading scholar in observational road safety studies has also noted that cross-sectional studies have not proven successful in terms of allowing cause and effect interpretations in road safety because of inconsistent results between the studies (*18*). Cross-sectional study designs usually do not consider a site-selection mechanism to minimize pretreatment differences while identifying treatment and control sites (*29,34*). Potential data and modeling issues that often plague observational study designs are missing information on daily traffic patterns, measurement errors in right-hand side explanatory variables, invalid model assumptions, and uninformed selections of functional form (*29*,*34*,*36*-*37*). These issues can have huge impacts on regression coefficient estimates, and are particularly problematic with cross-sectional studies that seek to use regression models to estimate the effects of right-hand side (i.e., explanatory) variables on expected number of crashes. Also, statistical road safety modeling techniques currently used are not sufficiently structured so that findings from different cross-sectional studies can converge on similar model forms and regression coefficient estimates.

In summary, results obtained from cross-sectional study designs are usually not considered as reliable as those derived from well-designed before-after studies. Even with these drawbacks, there are still many benefits to conducting cross-sectional studies. The opportunities for conducting cross-sectional studies are plentiful and several years of before and after period data are not required. This is why much of what is thought to be known about the safety effects of many road features came from observational cross-sectional studies, in particular, studies that use multivariate regression models to fit the data (*18*).

Problem Statement

As demonstrated in previous sections of this proposal, regression modeling of expected crash frequency plays a key role in observational before-after and cross-sectional road safety studies. The regression models are: 1) used to estimate the expected number of crashes, and 2) in cross-sectional studies, used to estimate the effects of right-hand side variables on expected number of crashes. Various challenges regarding both data and regression modeling approaches exist. Left unaddressed, these challenges will continue to have a significant impact on the accuracy and repeatability of observational road safety study results. Three issues are particularly salient and were addressed in this research: 1) missing information on daily traffic patterns; 2) measurement errors in right-hand side variables; and 3) modeling practices and assumptions, specifically related to letting statistical diagnostics drive the functional forms, letting estimation results drive the interpretations, and ‘starting over’ with each study.

Information on Daily Traffic Patterns

The selection and availability of right-hand side explanatory variables is an important step in the development of regression models of expected crash frequency. Currently, a majority of regression models capture the effect of traffic using average annual daily traffic (AADT) estimates, which represent an estimated average daily volume over the course of a year. The AADT estimate turns out to be a highly influential right-hand side variable in the models, as it represents one part of an exposure term.

However, models that use AADT alone do not explicitly capture what are expected to be significant safety effects of differences in traffic volume patterns throughout the day. For example, previous studies in the literature have shown that there are substantial differences in safety performance during the day and night. Fatal crash rates are reported as four times higher at night than during the day (*38-40*). It is reasonable to believe that the expected number of crashes would vary significantly with the amount of traffic during day versus night on a road segment. However, these types of volume patterns are rarely available over the course of the lengthy study periods needed in road safety research.

Measurement Error in RHS Variables

There are clearly cases when the key explanatory variables in regression models of expected crash frequency suffer from measurement error. Poor measurements of right-hand side variables in regression models introduces bias and inconsistency in regression coefficient estimates, increases error dispersion, and masks the features of the data (*41-42*). Current regression models of expected crash frequency do not explicitly account for the fact that some right-hand side variables used in the safety models are estimates themselves with estimation errors. For example, AADT estimates used in safety models (usually a single value of AADT) have significant uncertainty, because estimating these AADTs frequently involves extrapolating short-term counts over time and space. The effects of these measurement errors on the results of observational road safety studies have not been fully explored.

Model Assumptions, Functional Form, and "Starting Over"

The specifications of regression models of expected crash frequency are often driven solely by statistical significance, i.e., modelers rely on traditional statistical diagnostics to guide the model form (*43*). After a model form has been established, the explanatory variables that impact the expected crash frequency are explored by testing different model specifications. Researchers then let their model estimation results drive the interpretations (*34*). However, each new study typically 'starts over,' poses alternative forms and specifications, and chooses the one that fits the existing data best, without fully utilizing previous study results. Hence, a choice of appropriate model functional form that provides reliable associations between explanatory variables and crash outcomes and which converges on similar model forms and coefficient values is less clear in road safety research (*34*). The effects of including prior knowledge in the form of informative priors in safety models have not been fully explored.

Summary

These three data and modeling issues are problematic for both before-after studies and cross-sectional studies. They are particularly problematic in cross-sectional studies where the sample of data and functional relationship chosen for the regression model can highly influence the result. This research seeks to further explore three key data and modeling limitations and identify possible solutions. Results from this research were expected to improve the accuracy and repeatability of observational study results.

Research Objectives and Scope

The overall objective of this research was to further explore three key limitations in both data and regression modeling approaches used in observational road safety studies and identify possible solutions. Specifically, solutions that seek to develop more informed and complete model specifications using detailed datasets and empirically-derived theory were developed and assessed. Given the framing of the problem, the research questions guiding this work are the following:

1. Can traffic-related measures of exposure be developed at more disaggregate levels than annual average daily traffic in rural areas where the traffic volume data are limited, but traffic patterns during night and day are expected to have a large impacts on safety performance? How does accounting for exposure at more disaggregated levels than annual average daily traffic impact safety effects estimates of selected right-hand side variables as well as model prediction results? In particular:

   a. Is kriging an effective spatial analysis approach for estimation of daily traffic patterns (by time-of-day) in rural areas, where data are limited?
   b. How does including the information on daily traffic patterns, specifically during the night and day, impact safety effects estimates of selected right-hand side variables and model prediction results?
2. How does accounting for measurement errors in right-hand side variables impact safety effects estimates of selected right-hand side variables in regression models of expected crash frequency? In particular:
   a. What are the impacts of ignoring measurement errors on the regression model estimation results?
   b. How do applying functional-type measurement error correction methods, Regression Calibration and Simulation Extrapolation, impact the safety effects estimates of selected right-hand side variables?
3. Does incorporating prior knowledge of the safety effects of explanatory variables into Bayesian model formulations yield different conclusions with respect to model prediction results and safety effects estimates of selected right-hand side variables? In particular:
   a. How can informative priors for explanatory variables be developed using the results of previous observational road safety studies?
   b. How does including informative priors impact safety effects estimates of selected right-hand side variables and model prediction results?

These research questions were explored using data from rural, two-lane highways in the States of Utah and Washington. The datasets consisted of horizontal curves, for which crash data, roadway geometric features, operational characteristics, roadside features, and weather data were obtained. While the methodological framework was applied only to horizontal curves along rural, two-lane highways in this research, it is expected to be applicable to other site and facility types. Findings from this research are expected to lead to methodological advancements in observational road safety studies that increase the accuracy and repeatability of study results.

The significance of this research lies in the application of new methodologies to improve road safety effect estimation and prediction using multivariate regression models. This research provides three major methodological contributions to facilitate the inclusion of new explanatory variables related to daily traffic patterns, correct for measurement errors, and incorporate past information and findings in safety models in addition to the available data. However, with the existing data sources, good quality data are not available that is suitable for application of these sophisticated methodologies, and validate their results completely. In the context of this research, these new methods were incorporated in the regression model development based on certain assumptions (because of the data limitations) and they provide an alternative way to explicitly account for the issues that were left unaddressed in regression modeling of expected crash frequency until date.

For road safety researchers and practitioners, these methods will pave the way for explicitly accounting for the limitations that are conveniently overlooked and not addressed currently in road safety research. The main study findings can be effective in estimating how the incorporation of these new methods affects the modeling results on a general level. However, in a real case study, true value of the data needed to employ and derive useful conclusions is not available. With the increase in the availability of alternative data sources like light detection and ranging (LiDAR) and detailed roadway and roadside inventories, it is hoped that the application of these methods to the real-world case studies will be useful in deriving insights related to determining the specific impacts on safety effects estimates and model prediction results.

CHAPTER 2

LITERATURE REVIEW

This chapter includes a summary of key literature review on alternative approaches that can overcome data and regression modeling issues in observational road safety studies. The literature review done for this research is three-fold. The first section describes the previous research on different measures of exposure that have been used in safety models, and the estimation of daily traffic patterns using different approaches. The second section describes the previous work on the effects of error-prone explanatory variables on parameter estimates, and the measurement error correction methods that are employed to correct for measurement error. The third and last section describes the previous work on Bayesian methodology that has been applied in the context of road safety research.

Measures of Exposure

In regression models of expected crash frequency, measures of exposure are considered key information. Exposure is generally expressed in a form related to amount of travel, but accurately defined as number of opportunities for traffic crashes to occur (*44-46*). The number of opportunities usually refers to amount of travel, which is the number of person miles or vehicle miles of travel performed (*47*). Two basic and different frameworks were suggested in the literature for the measurement of exposure relevant to road safety research. The first is to obtain the measurements of exposure at a site using mechanical traffic counters, human observations, and automatic cameras. The second is to obtain data after the trips are completed, using in-person interviews, telephone interviews, and traffic surveys (*48-50*).

However, with the era of Big Data, new technologies of data collection of various types and various sources are also available. GPS traces of vehicles, GSM traces of mobile phones, and social networking sites are used to derive information on exposure in present days. Choosing an appropriate measure of exposure depends on the intended use of data and population studied, as different exposure measures can produce different results (*51*).

Several types of exposure measures that are relevant to modeling expected number of crashes have been developed in the literature. For most of the trend studies and comparisons of road safety performance on an international and national level, the number of inhabitants or number of vehicles is used as the exposure measure (*46,52*). When calculating the crash rate on a certain part of the road network (i.e., at a segment or an intersection), the most frequently used measures of exposure are AADT, vehicle miles traveled in a given period, road segment length, or number of entering vehicles (*45,46,53,54*). However, some researchers argue that time spent traveling is a better measure of exposure than distance traveled, when comparing the risks of different travel modes (*55*). This is because time-based measures also take into account the variation in traveling patterns, speeds, and environmental factors that are particularly important to consider when comparing risk among different travel modes (*56*).

Depending on the availability of data and scope of the analysis, some of the previous studies in the literature investigated the application of various exposure measures and surrogates for exposure in modeling expected number of crashes. Total population, number of trips, number of registered vehicles or number of licensed drivers, gasoline prices and consumption, volume-to-capacity ratio, and traffic density, which are proxies for exposure, were used in modeling expected number of crashes (*53,57-61*).

Despite the apparent simplicity of the definition and availability of different methods for measuring and including exposure variable in modeling expected crash frequency, there is still a considerable range of ideas in the literature as to which exposure measurement is most desirable to use and how it should be collected in a theoretically satisfactory way (*50,62*). Perhaps the best description of the current state in exposure measurement is that there is no general rule concerning the preferred and reliable measures of exposure and depends on data availability, quality, and objective of the analysis.

Exposure and Road Safety

The importance of exposure variables to modeling expected number of crashes, as mentioned previously, has been recognized in road safety research for over two decades. Road safety researchers have commonly used some combination of segment length (if the unit of interest is a segment) and the number of vehicles that pass by a fixed point (i.e., traffic volume) during a specific time period as measures of exposure (*45,54*). Traffic volume is usually represented as average annual daily traffic in statistical road safety models. AADT represents the average 24-hour traffic volume at a given location over a full 365-day period. AADT estimates are usually based on extrapolating short-term traffic counts conducted during a few days of the year, once every few years, over time and space. Initially, road safety researchers considered a linear relationship between traffic volume and expected crash frequency (*53*). However, a significant body of published work has revealed a nonlinear relationship between crashes and traffic volume (*58,63-67*).

The nonlinear relationship between number of crashes and average annual daily traffic may be due to some factors overshadowed by this aggregated exposure measure. All of these aggregated measures of exposure, be it the most common measures or proxies (surrogates), do not consider temporal traffic variation (*54*). For example, these measures do not consider the possibility that number of crashes during a specific time of day or day of week is related to the prevailing flow rate at that time and the distribution of exposure measure may vary from daytime to nighttime or weekday to weekend.

This discussion is supported by the findings of a study in which the crash rate for a million vehicles is higher at night than at other times of the day (*39*). The differences in safety performance during the day and night are widely recognized by other studies as well, with fatal crash rates reported as much as four times higher at night than during the day (*38-40*). This variation is largely attributed to differences in visibility and the human aspects of driving, such as biological clock influences on driver alertness and sleepiness (*68-69*).

In addition to using aggregated AADT estimates in modeling the expected number of crashes, some of the studies have explored the interaction between traffic flow variables as exposure measures and road safety. Previous literature explored the interaction between hourly traffic volumes and expected crash frequency under different conditions of day, night, and environmental factors (*58,63,70-77*). Some other studies found that, considering variables, such as volume-to-capacity ratio, traffic density, and variation in daily travel patterns, offers richer explanations of crash risk variation than traditional aggregated AADT estimates in statistical road safety models (*39,57-60*).

Estimation of Daily Traffic Patterns

Average annual daily traffic can be determined exactly at areas/sites having permanent automatic traffic recorders (ATRs) in continuous operation that record traffic flows throughout the year. These ATR stations are permanently installed throughout the roadway network. Coverage includes all functional classifications of highways, although coverage significantly decreases for lower classified roadways (*78*). At these permanent ATR stations, loop detectors, weigh-in-motion sensors, and/or other equipment are installed to count the number of vehicles passing through each location, continuously throughout the year (*79-80*). It can be said that the data obtained from ATR stations are the most reliable and comprehensive hourly traffic volume data at that location in a given time period. The number and spatial frequency of ATR sites vary by state and by region, as well as the functional class of the roadway (*81*). On average, one ATR station exists per 400 centerline miles of state-maintained roads and per 1,000 miles of all public roads (*82*). These ATR stations therefore do not cover majority of road segments. Utilizing data from ATRs alone is not practically feasible for road safety research.

For a majority of roadway segments in a region or state where ATR coverage is limited, AADT is estimated by extrapolating short-term traffic counts over time and space. The short-term traffic counters cover a majority of road segments without the permanent counters, and the traffic volume data are collected once every year to few years, with the collection periods ranging from 1 to 7 days (*83*). The average of that sample, along with permanent counts and functional class criteria, are then used to estimate AADT. These AADT estimates have significant uncertainty, and lack information on the daily traffic volume patterns throughout the 24-hour day. In order to estimate traffic volumes at locations where no ATRs or short-term counts are available, most widely used methods in the literature involved ordinary least-squares (OLS) regression with multiple right-hand side variables, time series analysis, neural networks, Gaussian maximum likelihood techniques, and nonparametric regression models (*84-87*).

With advancements in spatial analysis techniques and availability of spatial datasets, researchers have also started to explore methods that utilize the spatial context of traffic data. Recent research suggested the use of spatial interpolation methods for traffic volume estimation at locations where traffic recorder stations are absent (*82,88-91*). One of the most promising spatial interpolation methods is kriging, which in this context uses available data at existing ATR locations and spatial autocorrelation assumption to predict the traffic volumes at locations where volumes are unknown (*92-93*).

Kriging has applications in a wide variety of fields, such as air quality analysis, natural resource analysis, and water studies (*94-96*). Kriging methods are capable of predicting variable values at unmeasured locations, while assessing the errors of these predictions. Most of the previous research on predicting traffic volumes at unmeasured locations relies on the notion that unobserved factors are autocorrelated over space, and the levels of autocorrelation decline with distance (*82,89*). There are much smaller number of studies that accounted for the influence of covariates on traffic volume predictions. These studies considered functional classification of the roadway, facility type, number of lanes, posted speed limit, and socio-demographic characteristics, in addition to the spatial proximity assumption, to predict traffic volumes at unmeasured locations (*90-91,97*). In addition to these covariates, the differences in the total number, spatial frequency, and other known and unknown factors related to traffic counting stations will shape the modeling errors and ultimately have an effect on the spatial prediction of traffic volumes (*81,90*). The published literature concluded that inclusion of covariates resulted in minimizing the mean square prediction error in spatial (i.e., traffic volumes in this context) predictions (*90,98*).

One study that predicted traffic volumes and incorporated characteristics of covariates in spatial interpolation techniques used universal kriging combined with Trans CAD travel demand modeling software to obtain the shortest-path distances (*90*). However, there are many different ways to incorporate characteristics of covariates (also known as secondary information) in spatial analysis. In fields other than transportation, two multivariate geostatistical methods, ordinary cokriging (OCK) and universal kriging (or kriging with an external drift), have been used to incorporate both the spatial proximity and other covariates to make predictions at unmeasured locations (*99-102*). The studies showed that the spatio-temporal kriging with external drift predictions was physically more realistic, and resulted in a much higher estimation precision than other methods in both a statistical and qualitative way (*103-104*).

In summary, there are a few studies that attempted to include covariates in spatial prediction models and they concluded that accounting for the association between the covariates and dependent variable is very important. In the context of this research, these associations are expected to greatly impact traffic volume predictions at unmeasured locations. The next section gives an overview on the implications of including error-prone explanatory variables in regression models of expected crash frequency. The section also includes information on the applications of different types of measurement error correction strategies in the literature.

Measurement Error Models

The quality of data is of central importance for the results of the statistical analysis. In many research areas, the measurement accuracy of a variable is a frequent issue. When the explanatory variables are measured with uncertainty/error, the problem is referred to as measurement error for continuous variables and misclassification for categorical variables (*105*). The implications of ignoring measurement errors in statistical road safety models are substantial, often resulting in unreliable parameter estimates (*106*). Statistical analysis that takes into account the measurement error is much more complicated than the ordinary regression analysis, and most of the statistical packages available do not provide standard programs (or packages) for correcting the effects of measurement error. Hence, the widely used approach is to ignore the measurement error and perform 'naïve analysis' (*107*).

Ignoring measurement error in explanatory variables introduces bias and inconsistency in regression coefficient estimates, and increases error dispersion (*41-42*). Particularly, it attenuates the regression coefficient towards zero in comparison with the result computed from a regression on the same variable measured without error (*107*). In a multivariate regression model, some of the explanatory variables can be measured with and some without error; the parameter estimates of error-free explanatory variables can also be biased if there are measurement errors in other right-hand side variables and the direction of the bias depends on the correlation among the explanatory variables (*108-*

111). In addition to the bias in regression coefficient estimates, measurement error may also mask the features of the data, therefore requiring much larger sample sizes to detect effects (*41-42*). Given these facts, it is surprising that measurement error is completely ignored in statistical road safety modeling.

In statistical road safety modeling, AADT usually turns out to be highly influential on expected number of crashes along with other roadway, roadside, and weather variables that are related to expected crash frequency (sometimes disaggregated by crash type and severity). As mentioned in the previous section, AADT estimates are obtained by extrapolating short-term traffic counts over time and space. This results in significant uncertainty in AADT estimates, which is also known as measurement error. The effects of error-prone explanatory variables on regression coefficient estimates have been recognized and adjustment methods have been applied in regression modeling for over two decades (*112-113*).

A number of methods have been developed to handle different types of measurement errors, study designs, and statistical analysis settings. The most currently available methods are suitable for handling continuous explanatory variables in generalized linear models; however, there have been fewer developments for applications to categorical explanatory variables (*114-116*). In order to appreciate the reliability of the results obtained from employing these methods to correct for measurement error in regression modeling, it is very important that the researchers and practitioners have a sense of the magnitude of the measurement errors associated with error-prone explanatory variables.

As mentioned earlier, many correction strategies have been proposed to deal with measurement error in continuous explanatory variables. These strategies can be broadly grouped into functional and structural type methods (*41*). Functional type approaches regard the error-prone explanatory variables to be unknown, nonrandom constants (or parameters), whereas the structural type approaches regard the error-prone explanatory variables to be random variables (*105-106*). The two functional-type measurement error methods that are widely applied to generalized linear models in previous published studies are Regression Calibration (RCAL) and Simulation Extrapolation (SIMEX) (*108,117-118*). Bayesian methods are an example of structural type approaches, which are implemented using Markov Chain Monte Carlo (MCMC) algorithms to correct for the measurement error in random explanatory variables (*105*).

RCAL is a conceptually straightforward approach to bias reduction and has been successfully applied to a broad range of regression models (*119*). The estimator obtained from this method is fully consistent in linear and log-linear models, and approximately consistent in nonlinear models. The method is appropriate when a linear measurement error with a constant variance applies to the error-prone explanatory variable (*120*). RCAL was first studied and applied in the context of proportional hazards regression (*121*). Later, the method was applied to logistic regression and other generalized linear regression models (*122-124*).

In the case of logistic regression, the method was found to be effective in a number of studies and applications (*124*). However, some studies also found that RCAL can be ineffective in reducing bias in nonlinear models when the effect of error-prone variable on response and the measurement error variance is large (*119*). The quantification of what is meant by ‘large’ is not clear in the literature yet, and no clear distinction has been made on the successful or unsuccessful application of this method to the type of regression models.

SIMEX is a widely applicable, simulation-based method of estimating and correcting for bias in a very broad range of settings, and is the only method that provides a visual display of the effects of measurement error on regression coefficient estimation (*41,117,125*). Similar to RCAL, the estimator obtained from SIMEX is fully consistent in linear models, and approximately consistent in nonlinear models. The method was first proposed by Cook and Stefanski and further developed by Stefanski and Cook (*126*). Later, there were many applications of this method in research fields like biostatistics, epidemiology, and ecology (*127-130*). The published studies found that the SIMEX method is effective in readily estimating and correcting the regression coefficient estimates that are biased due to error-prone explanatory variables in any type of regression models. One study by Fung et al. (1999) evaluated the two measurement error adjustment methods (RCAL, SIMEX) in Poisson regression and found that RCAL performed very well in terms of reducing attenuation bias in regression coefficient estimates and maintaining accurate standard errors of the regression coefficient estimates (*113*).

Measurement error correction methods have been applied to very few studies in transportation engineering literature. Studies that applied measurement error models to error-prone average daily traffic estimates and work zone length during the work zone conditions found significant bias in regression coefficient estimates (*131-132*). In the field of road safety, El-Basyouny and Sayed proposed a lognormal measurement error model applicable to crash prediction models, and compared the regression coefficient estimates obtained from this method with the parameter estimates of models that did not account for measurement error in traffic volume estimates (*133*). The study found that the bias in regression coefficient estimates increases with magnitude of measurement error in traffic volumes when developing a safety performance function (*133*). Despite the existing literature on the likely impacts of measurement error in explanatory variables on regression models of expected crash frequency, the application of measurement error correction strategies in road safety research has caught little attention. However, in recent years, there is an increased interest in the use of epidemiological methods in road safety research (*26*). The same concept of functional type measurement error correction strategies applied to epidemiological studies can be applied to road safety. In road safety, the entity of interest is a roadway segment or intersection, and the error-prone explanatory variable is traffic volume (i.e., AADT) estimates of a given roadway segment or intersection, while modeling expected crash frequency.

Bayesian Framework in Road Safety

Using conventional frequentist approaches, it is virtually impossible to accommodate past information unless the data from past and present are combined, to make it one complete dataset. However, this aggregation of data might ignore the diversity of data structures such as different time of data collection, changes of traffic environment and policies, driver behavioral changes, and other factors (*134*). Hence, a Bayesian framework is important to incorporate the accumulated knowledge from previous observational studies in the analysis.

Bayesian statistics have some advantages when compared to maximum likelihood estimation (in frequentist approach), such as, interesting probabilistic interpretative properties, superiority in dealing with uncertainty and randomness, and the ability to analyze complex data (*135-137*). In road safety studies, Bayesian methods have been widely applied for both identifying hotspot locations, and evaluating the effectiveness of countermeasures (*27,32,138-143*). Within the class of Bayesian formulations, empirical Bayes approach and hierarchical Bayes (also known as full Bayes) approach are commonly used in road safety research.

In principle, priors formally represent available information. But in practice, noninformative and improper priors are often used in many studies (*144*). These noninformative priors do not make use of real prior information and often times result in a posterior inference that is not quite credible. The use of informative priors in Bayesian road safety analysis is rare, perhaps because it is difficult to express prior knowledge as a probability distribution or because informative priors are perceived as overly subjective, or a fear that they could reduce model accuracy (*145-147*).

A prior not only affects the precision of estimates, but also the location of posterior, therefore affecting predictive accuracy (*145*). The ability to incorporate informative priors in Bayesian analysis has been exploited in a little more detail in other disciplines such as ecology and water resources. Past research in ecological modeling suggests that including prior information in a Bayesian model will increase the precision of relevant parameter estimates and posterior predictive distributions (*148*). However, the effect on model accuracy is not known and can only be done through model validation.

As mentioned earlier, most of the current studies that modeled expected number of crashes using Bayesian inference methods have incorporated noninformative priors (or vague priors), which ignored the merit of Bayesian inference methods. Recently, researchers started to incorporate prior information from previous studies and experts' judgment, but very few studies focused on developing prior distributions of independent variables, using parameter estimates of variables obtained from previous studies (*149-150*). The road safety studies that have constructed informative priors based on historical data used power prior technique; method of moments, maximum likelihood estimation, maximum entropy estimation, starting from a noninformative pre-prior and fitting a prior based on confidence/credible interval matching (*151-154*).

The findings from previous studies in the literature stated that the informative priors improved the model goodness-of-fit compared to the noninformative priors (*151*). This research specifically tends to use Bayesian formulations incorporating informative priors in the analysis based on regression coefficient parameters estimated from previous rigorous and well-defined observational studies, and accumulated knowledge in the field of road safety.

Models of Expected Crash Frequency on Horizontal Curves

In addition to the improvement that an appropriate exposure measure can bring to regression modeling of expected crash frequency, proper selection of explanatory variables is also needed when modeling expected number of crashes (*53*). Extensive research has been performed to examine the relationship between traffic crashes, traffic flow characteristics, roadway features, and geometric attributes. Many studies looked at horizontal curve features that affect crash performance and experience on rural, two-lane roads (*155-161*). The results were based on an analysis of horizontal curves, with corresponding crash, traffic, roadway, and geometry-related variables. Safety models developed with these variables revealed significantly more curve crashes for sharper curves, narrower curve widths, lack of spiral transitions, and increased superelevation deficiency. The coordination and safety effects of interaction between horizontal and vertical alignments on rural, two-lane roads were investigated by Bella (2015), and Bauer et al. (2013), respectively (*162-163*). The results indicated that the expected crash frequency increases with curve length and increasing percent grade, with all other factors being the same.

A significant amount of published literature exists that looked at the relationship between detailed cross-sectional characteristics (lane width, and shoulder width) and expected crash frequency (*164-169*). The results showed that cross-section of the roadway can have a significant impact on safety of rural, two-lane roads; earlier studies concluded that the expected crash frequency decreased with an increase in lane width. However, a recent study showed that after a minimum required lane width for the vehicle, any additional width beyond this minimum allows the driver to increase speed, which increases the expected crash frequency (*164*). Fitzpatrick et al. (2008) investigated how driveway density on rural two-lane highways impacted crash rates, using accident modification factors (*170*). The results showed that the probability of crash occurrence increases as number of driveways on rural two-lane roads increase. Many studies have also found that the risk of traffic crashes increase with high speeds and high traffic volumes on rural roads (*171-174*). There are relatively few studies that investigated the effect of weather on crashes in rural two-lane roads (*175-177*).

Summary of the Literature Review

In spite of the steady progress in the methodological advancements in multivariate regression modeling of expected crash frequency, there are still many fundamental data and modeling issues in observational studies that have not been completely addressed. The three key data and modeling limitations that will be addressed in this research are: (1) a majority of statistical road safety models use AADT to represent traffic volume and do not explicitly capture differences in traffic volume patterns throughout the day, even though crash risk is known to change by time-of-day; (2) statistical road safety models that use AADT on the "right-hand side" of the model equation do not explicitly account for the fact that these values for AADT are estimates with estimation errors, leading to potential bias in model estimation results; and (3) the current state-of-the-practice in road safety research often involves "starting over" with each study, choosing a model functional form based on the data fit, and letting the estimation results drive interpretations, without fully utilizing previous study results.

Each of these issues may substantially influence the estimated safety effects of explanatory variables as well as model prediction results, and ultimately influence the inferences drawn from the analysis of data. As road safety researchers continue to conduct more studies that help in exploring and understanding the relationships between explanatory variables and safety outcomes in detail, these issues will need to be addressed. This research will seek to provide key insights into these key data and modeling issues, and explore possible solutions. From the reviewed literature, the alternative approaches that will help in addressing these issues are identified.

The daily traffic patterns (by time of day) will be estimated using the kriging approaches that have been applied in other fields such as air quality analysis, natural resource analysis, and water studies. The functional-type measurement error approaches, Regression Calibration and Simulation Extrapolation, will be useful in accounting for the measurement error in AADT estimates while modeling the expected number of crashes. The prior knowledge on the safety effects of explanatory variables will be incorporated into the regression models of expected crash frequency using Bayesian methodological framework. More details on the specific methods and the data that will be used to address the three research questions are provided in the following sections.

CHAPTER 3

METHODOLOGY

This section describes methods and approaches in detail that are used for addressing the selected observational road safety study data and modeling issues discussed in the previous sections. The methods described in this research are four-fold. The first section discusses the methods for estimating daily traffic patterns (by time of day). The traffic pattern estimates were then ultimately used in safety models. The second section describes the functional-type measurement error correction approaches that were employed to correct for measurement error in explanatory variables. The third section describes the basic safety modeling framework that was used to assess the potential modeling impacts of findings from the first two efforts on traffic pattern estimates and measurement error corrections. The fourth and final section discusses the Bayesian methodological framework that was used to incorporate informative priors while modeling expected crash frequency.

Estimation of Daily Traffic Patterns

For over 40 years, the research on the estimation of AADT was mostly based on the combined use of short-term traffic counts and factor approaches (i.e., using seasonal and monthly adjustment factors for different periods of the year). The fact that these AADT estimates have significant uncertainty, and lack information on the daily traffic volume patterns throughout the 24-hour day, constantly presents a challenge when developing regression models of expected crash frequency. One of the most promising approaches for addressing this issue in the literature is the use of spatial interpolation methods, i.e., kriging. Kriging is one of the optimal interpolation methods used to estimate unknown values based on regression against observed values of neighboring data. This method is frequently used in the field of geostatistics to interpolate the unknown values. In the context of this research, kriging uses available data at existing traffic recorder stations to estimate traffic volumes at locations where volumes are unknown. In this part of the research, kriging was used to determine whether traffic-related measures of exposure can be developed at more disaggregate levels than annual average daily traffic in rural areas where the traffic volume data are limited, but traffic patterns during night and day are expected to have a large impacts on safety performance.

The first step in the process of estimating daily traffic patterns was to determine how to disaggregate traffic volume data throughout the year. The continuous traffic volume data that were used in this research was obtained from all ATR stations in Utah (corresponding to all facility types, and area types) for years 2009 through 2013, the study period. The ATR data were available through the Utah Department of Transportation (UDOT) website.

Data were collected at 15-minute intervals and stored at 1-hour intervals for 365 days of the year. Area type, and other location information, were provided with each ATR station. There were some instances of incomplete traffic volume data for a variety of reasons (e.g., ATR is turned off, out of service), but the missing data did not impact the ability to conduct the spatial analysis outlined in this section. Information provided by these counters can be used to study the temporal variations of traffic volume, such as volume by time-of-day, day-of-week, and seasonal traffic patterns. As already noted in the previous chapters, differences in safety performance during night and day are widely recognized, with fatal crash rates reported as much as four times higher at night than during the day. This has been attributed to differences in visibility and other driver-related factors such as biological clock influences on driver alertness and sleepiness (*68*).

Hence, disaggregating traffic volume data by time of day, with time of day characterized as "day" and "night," was a logical starting point to test the spatial interpolation methods in rural settings. The spatial modeling methodology used to estimate day and night traffic volumes at locations without ATRs consisted of the following three steps:

1. *Data preparation and transformation*: categorize and perform exploratory data analysis and visual normality tests for traffic volume data collected during day and night at locations with ATRs.
2. *Variogram modeling and kriging interpolation*: fit and select appropriate semivariogram models, perform kriging interpolations, and estimate kriging standard error.
3. *Performance assessment of the kriging method*: employ K-fold cross-validation that involves removing 20 percent of the observed (i.e., ATR-measured) sample from the dataset, and then predicting the values at those locations using information from the remaining ATR stations.

The following subsections describe each of these steps in detail. The first step involves the preparation of data for spatial interpolation; the second one involves accounting for spatial autocorrelation and kriging maps, while the third one involves cross-validation.

Data Preparation and Transformation

For this study, as mentioned earlier, the continuous traffic volume data were obtained from ATRs throughout the state for 5 years from 2009 through 2013. Data were collected from 100 ATR stations throughout the state, with majority of the ATRs concentrated in the urban areas along Interstate 15 and Interstate 80. The hourly traffic volume data obtained from ATR stations in Utah were disaggregated into day and night traffic volumes based on sunrise and sunset times.

The time period from 1.5 hours after sunset to 1.5 hours before the sunrise was classified as night, with the remaining hours classified as day (*178-179*). Lengths of day and night in Utah vary greatly throughout the year because of the high latitude of the United States. In the middle of June, nights are approximately 6 hours long whereas in December, nights are about 11 hours long. To take into account this significant variation in day and night durations, the entire dataset was divided into four seasons corresponding to selected months of the year. Day and night traffic volumes were then determined across the entire study period. Table 1 shows the day and night times for the state of Utah as a function of season.

The method of kriging was first developed by Georges Matheron and relies on the notion that unobserved factors are autocorrelated over space, and the levels of autocorrelation decline with distance (*92*). Kriging assumes that data exhibit stationarity, indicating that the correlation (covariance or semivariogram) between any two locations depends solely on the distance between them, not on their exact locations (*180*). Kriging generally leads to an optimum estimator and yields best results when the data are normally distributed. Thus, the inconsistency present in the observed data should be identified and fixed prior to model development and analysis.

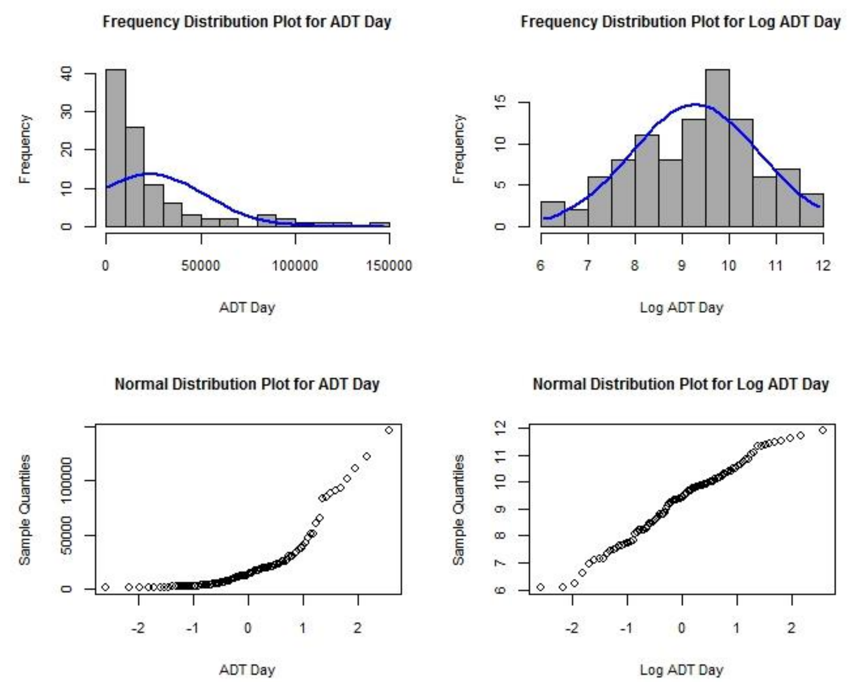
Table 1 Day and Night Times by the Season of the Year

Season/Month of the Year	Day Time	Night Time
Spring: March – May	5 am – 9:59 pm	10 pm – 4:59 am
Summer: June - August	4 am – 9:59 pm	10 pm – 3:59 am
Fall: September - November	6 am – 7:59 pm	8 pm – 5:59 am
Winter: December - February	6 am – 6:59 pm	7 pm – 5:59 am

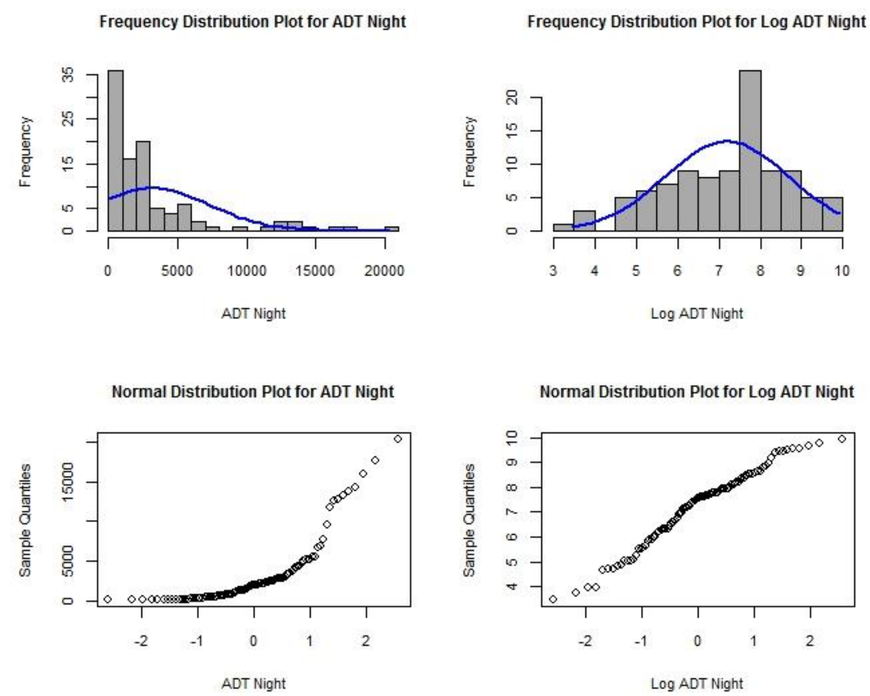
This includes detecting and removing outliers, performing normality tests using the observed data, and applying data transformations for non-normally distributed datasets. A log-transformation is very common and often used for data that have skewed or nonnormal distributions. To meet the assumption of data normality, the distribution in the histogram should be bell-shaped and the normal probability plot (normal Q-Q plot) should be a straight line at a 45 degree angle between the values and quantiles. In this research, a log-transformation was applied to day and night traffic volume data when the datasets did not satisfy the normal distribution assumptions. Figure 3 shows the frequency distribution plots and normal distribution plots for nontransformed and transformed data for day and night traffic volumes obtained from ATR stations in Utah.

Variogram Modeling and Kriging Interpolation

In addition to the traffic volume data that were obtained from all ATR stations across all facility types and area types in Utah, other measurable characteristics around the location of interest were also used in estimation of daily traffic patterns. These characteristics include socio-demographic characteristics, and other location characteristics of ATR stations that were assumed to have an impact on the traffic volume estimates in that area. For each ATR location, information on the number of lanes and functional classification was obtained from UDOT in the form of ArcGIS shapefiles.



Frequency Distribution and Normal Distribution Plots for Average Daily Traffic during the Day for Nontransformed and Transformed Data



Frequency Distribution and Normal Distribution Plots for Average Daily Traffic during the Night for Nontransformed and Transformed Data

Figure 3 Frequency Distribution and Normal Distribution Plots for Nontransformed and Transformed Data for Day and Night Traffic Volumes

Socio-economic data for the entire state at a census block level were obtained from the United States Census Bureau for the year 2010. This dataset included information on population counts, and household unit counts at the census block level. Location information available for the ATR stations was used to import the data to ArcGIS and to link the ATR stations to the corresponding additional information, i.e., number of lanes, functional classification, and socio-economic characteristics using the ArcGIS “spatial join” function. Descriptive statistics for the data points corresponding to the ATR stations are provided in Table 2. The traffic volumes to be predicted by kriging were assumed to depend on several observable factors (or covariates), which are linked to the location of interest, creating a trend estimate, $\mu(s)$. The spatial variables can be defined as shown in the following equation (*82*):

$$Z_i(s) = \mu_i(s) + \epsilon_i(s) \qquad \text{Equation (3)}$$

where:

$Z_i(s)$ is the variable of interest at site i (in this case, day or night traffic volumes);

s is the location of site i, determined by coordinates (x,y);

$\mu_i(s)$ is the deterministic trend (or drift); and

$\epsilon_i(s)$ is the random error component.

Based on the characteristics of variable of interest, $Z_i(s)$, there are three types of kriging: 1) if information on explanatory variables is lacking, is ordinary kriging; 2) if information on explanatory variables is available, is universal kriging; 3) if trend is known, is simple kriging. The universal kriging technique, where trends depend on explanatory variables and unknown regression coefficients, was employed in this research to model disaggregate traffic volumes at individual sites ($Z_i(s)$).

Table 2 Descriptive Statistics of Variables Linked to ATR Stations

Variable	Description	Mean	Std. Dev.	Min	Max
Total = 100 Observations					
ADT_Day	Average daily traffic during the day (veh/day)	22824	29191	428	146,411
Ln_Day	Log average daily traffic during the day	9.27	1.35	6.06	11.89
ADT_Ngt	Average daily traffic during the night (veh/day)	3148	4189	31	20358
Ln_Ngt	Log average daily traffic during the night	7.19	1.49	3.44	9.92
Num_Lane	Number of Lanes	3.02	1.28	1	6
Num_Lane3	Indicator Variable 1=number of lanes greater than 3 0=otherwise	0.61	0.49	0	1
Func_Clas1	Indicator Variable 1=Interstate 0=otherwise	0.27	0.47	0	1
Func_Clas2	Indicator Variable 1=Other Freeway/Expressway 0=otherwise	0.05	0.22	0	1
Func_Clas3	Indicator Variable 1=Other Principal Arterial 0=otherwise	0.42	0.49	0	1
Func_Clas4	Indicator Variable 1=Minor Arterial 0=otherwise	0.15	0.36	0	1
Func_Clas5	Indicator Variable 1=Major Collector 0=otherwise	0.06	0.24	0	1
Func_Clas6	Indicator Variable 1=Minor Collector 0=otherwise	0.01	0.10	0	1
Func_Clas7	Indicator Variable 1=Local 0=otherwise	0.03	0.17	0	1
Pop	Population in the census block with the ATR	46.74	92.37	0	399
Hou	Number of Housing units in census block with the ATR	22.50	44.56	0	204

In this approach, $\mu(s)$ can be a deterministic function of any form. A simple assumption is to use a linear function where $\mu(s) = X\beta$, with X containing explanatory variables characterizing each traffic recorder station surrounding the road segment of interest. In addition to these variables, $Z(s)$ is also influenced by "unmeasured" variables that influence traffic volumes at other, nearby locations due to spatial autocorrelation. These influences can be defined as $Z(s + h)$, where h represents the distance between the two sites. In other words, this means that the random terms $\epsilon_i(s)$ are spatially correlated. Such spatial autocorrelation can be quantified by the semivariogram $\gamma(h)$, defined as:

$$\gamma(h) = \frac{1}{2} var[Z(s + h) - Z(s)] \qquad \text{Equation (4)}$$

where:

$var[Z(s + h) - Z(s)]$ is the variance of the differences between corresponding traffic volumes at sites s and s+h.

Semivariogram analysis consists of the experimental semivariogram calculated from the data and the theoretical semivariogram model fitted to the data. In other words, the formula mentioned above is used to calculate the experimental semivariogram. With the experimental semivariogram concept defined, it was then necessary to select an appropriate curve, or semivariogram model that best fits the relationship between $\gamma(h)$ and h for a given dataset.

As mentioned earlier, the semivariogram model in kriging does not depend on the actual values, but their distribution. With the construction of a perfect semivariogram model, the predictions will represent a better value distribution with minimum estimation errors. There are several commonly used theoretical semivariogram models, including Exponential, Gaussian, Spherical, and Matern M.Stein's parameterization. These models rely on three parameters that describe the function shape and quantify the level of spatial autocorrelation in the data.

The nugget (c_0) is the intercept of the semivariogram with the vertical axis. The nugget effect reflects the discontinuity at the semivariogram's origin, as caused by factors such as sampling error, or inaccuracy in the instruments used for measurement, and short scale variability. The range (a) determines the threshold distance at which $\gamma(h)$ stabilizes, observations separated by a distance larger than the range are considered as spatially independent. The sill $(c_0 + c_1)$ is the maximum $\gamma(h)$ value; the higher the sill value, the higher the prediction variances (*82*).

Figure 4 illustrates these parameters and concepts. The semivariogram properties: the nugget, range, and sill, can provide insights on which model fits the existing data best (*181-182*). The semivariogram model parameters were estimated using iterative generalized least squares regression technique and all the available software packages use Euclidean distances between sites to define h, which can easily be determined using the site locations. In addition to the problem of stationarity, directionality of data is also found to be important (*182*). The spatial correlation of the set of measurements, in this case, traffic volume data, is often found to vary with direction, which means the data are anisotropic. However, in practice, it is very difficult to establish anisotropy from the data alone, particularly when there are fewer observations (or measurements). The open source software, R Studio, along with its geostatistical packages, was used to plot, fit, and verify the accuracy of alternative semivariogram models. All semivariogram models were considered to be isotropic for this analysis, meaning that the spatial autocorrelation structure was assumed to be the same moving outward in all directions from a site.

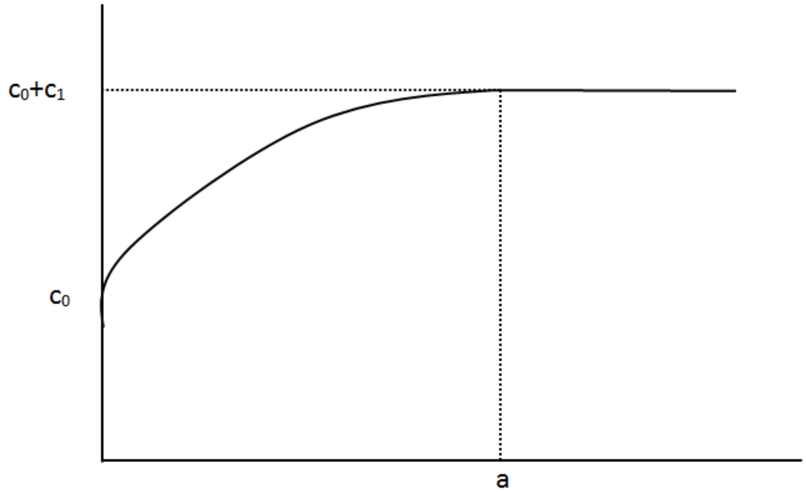


Figure 4 Illustration of Semivariogram Model (*82*).

This means that, in this study, the spatial autocorrelation depended only on the distance from the site, not on the direction. Given the isotropic semivariogram model assumption, spatial autocorrelation as a function of distance could then be estimated while taking into account other selected covariates characterizing each site. The resulting model could be used to predict day and night traffic volumes at locations without ATR stations.

Four different covariate specifications were tested when modeling the log-transformed day and night traffic volumes: 1) a model with no covariates, 2) a model with one covariate (number of lanes or number of housing units), 3) a model with two covariates (two indicator variables for functional classification), and 4) a model with multiple covariates (number of lanes, population or number of housing units, and two indicator variables for functional classification). In addition to population and housing variables, other economic variables such as average income and employment status were also investigated in the covariate specifications. However, these variables did not have as strong of a relationship in the semivariogram models as population and housing variables.

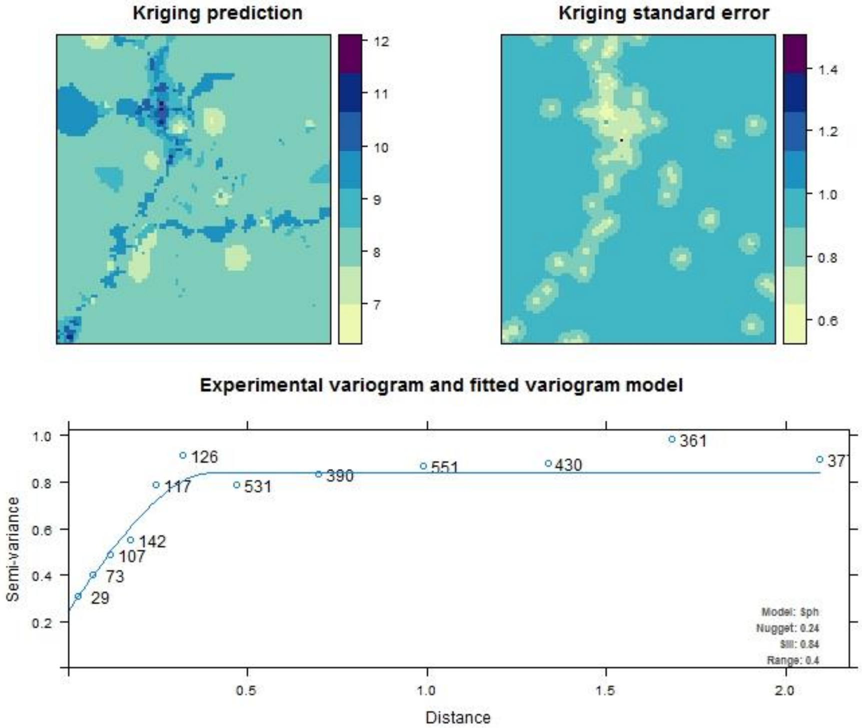
Four semivariogram models (Exponential, Gaussian, Spherical, Matern M.Stein’s parameterization) for each of the four different covariate specifications were estimated for both the transformed day and transformed night traffic volumes (i.e., thirty-two models). The semivariogram parameter estimates for the “best fitting” models for each of the covariate specifications are shown in Table 3. The idea behind developing and showing four semivariogram models with different number of covariates was to provide the reader with additional information so that the methodology can be easily understood and repeated or adapted to future studies. Matern M.Stein’s semivariogram model provided the best fit for the first three model specifications for both day and night traffic volumes. For the fourth specification, the Spherical model fit the data best for day traffic volumes and the Gaussian model fit the data best for night traffic volumes. The fourth specification was finally used for further analysis (i.e., kriging prediction) in this research. More discussion on the reasons for selecting this model specification is provided in the section below.

From Table 3, results indicate that the nugget values for Model 4 (with multiple covariates) for both day and night traffic volumes were less than for the other models. This shows that there is less variance in the observed data (due to measurement errors or sampling errors) for this particular covariate specification for both day and night traffic volumes. Based on the model fit, and semivariogram parameter estimates, Model 4 was selected as the final semivariogram model for all further analysis in the research, used to predict day and night traffic volumes. In other words, this semivariogram model specification with multiple covariates was used to develop kriging interpolation maps and estimate the standard error of predictions.

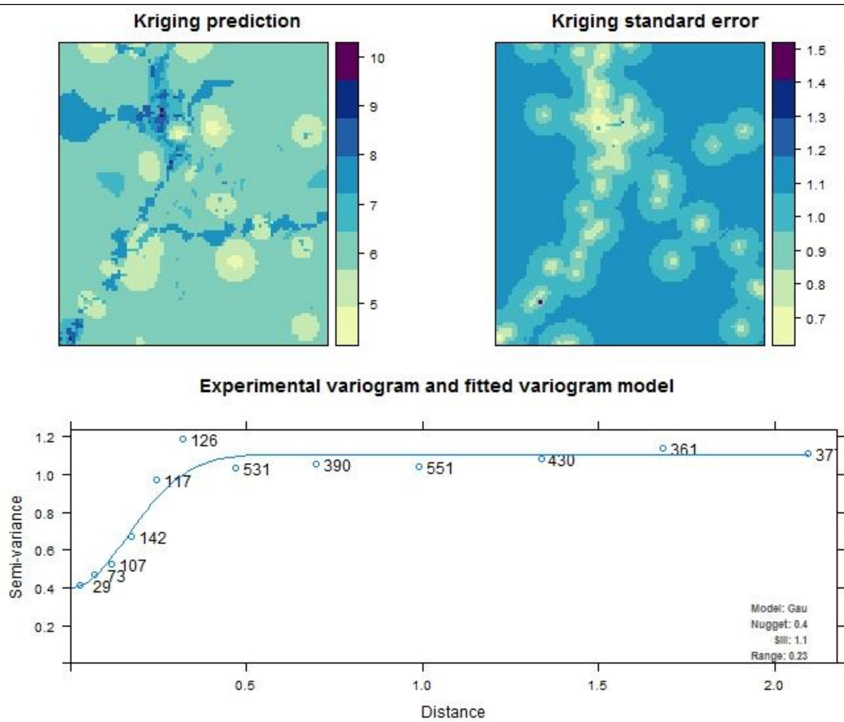
Table 3 Semivariogram Parameter Estimates for Models with Different Sets of Covariates for Transformed Day and Night Traffic Volume Data

Model type with/without covariates	**Semivariogram Model**	**Nugget (c_0)**	**Sill (c_0+c_1)**	**Range (a)**
	Day Traffic Volumes			
Model1 (No covariate)	Matern M. Stein (kappa=10)	0.60	2	0.30
Model 2 (Num_Lane3)	Matern M. Stein (kappa=10)	0.72	1.50	0.30
Model 3 (Func_Clas1, Func_Clas2)	Matern M. Stein (kappa=10)	0.40	1.60	0.30
Model 4 (Num_lan3, Pop, Func_Clas1, Func_Clas2)	Spherical	0.24	0.84	0.40
	Night Traffic Volumes			
Model1 (No covariate)	Matern M. Stein (kappa=10)	0.70	2.40	0.30
Model 2 (Hou)	Matern M. Stein (kappa=10)	0.75	2.30	0.24
Model 3 (Func_Clas1, Func_Clas2)	Matern M. Stein (kappa=10)	0.28	1.90	0.25
Model 4 (Num_lan3, Hou, Func_Clas1, Func_Clas2)	Gaussian	0.40	1.10	0.23

The final semivariogram models and kriging interpolation maps for day and night traffic volume estimates, developed using Model 4, are presented in Figure 5. From the kriging predictions, it can be seen that the predicted values range from 6 to 12 for log-transformed day traffic volumes and 4 to 10.5 for log-transformed night traffic volumes. The higher predictions of traffic volumes are represented in dark blue on the map, and are found in the Salt Lake City urban area. The smaller values are represented in light yellow, and fall to the south of Salt Lake City in the rural areas of the State. The standard error estimates range from 0.5 to 1.5 for predicted (i.e., expected) log-transformed day traffic volumes and 0.6 to 1.5 for predicted log-transformed night traffic volumes. The log-transformed values for traffic volumes obtained through kriging interpolation maps were then back-transformed to get the actual traffic volume estimates.



Kriging Interpolation Maps with Num_Lane3, Pop, Func_Clas1, and Func_Clas2 as Covariates for Transformed Day Traffic Volume Data



Kriging Interpolation Maps with Num_Lane3, Hou, Func_Clas1, and Func_Clas2 as Covariates for Transformed Night Traffic Volume Data

Figure 5 Kriging Interpolation Maps for Transformed Day and Night Traffic Volumes Based on Selected Model 4 (Specification with Multiple Covariates)

Performance Assessment of Kriging Results

A final step was implemented to determine how the kriging methods performed in predicting disaggregate traffic volumes at locations where the volumes were unknown. The performance assessments of the selected semivariogram models and kriging interpolation maps (identified in the previous section) were executed using a validation procedure known as K-fold cross-validation. In the K-fold cross-validation procedure implemented in R Studio, a set of measured points (20 percent of data points in this case) in the spatial domain were removed.

Traffic volumes at these locations were then estimated using the selected semivariogram model and kriging interpolation map as though the measurements were not available. The process is repeated with multiple “20 percent subsamples” until all of the available data points are removed at least once in a subsample. The resulting K-fold cross-validation statistics served as diagnostics to demonstrate whether the performance of the selected model was acceptable. The statistics were also used to check whether the prediction was unbiased and as close as possible to the measured values. The variability of the predictions was also assessed.

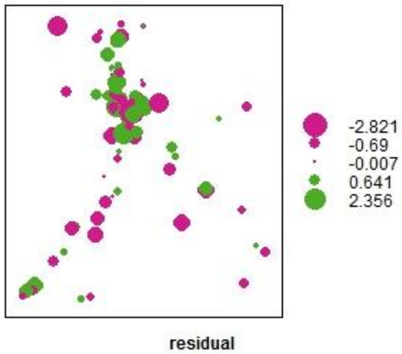
In particular, the cross-validation error, defined as the difference between the kriging estimate of traffic volume and the measured value, is calculated. This gives the ‘map of errors’ and scatter plots of estimated data from the corresponding measured values. The cross-validation diagnostics for the selected model (i.e., Model 4) for day and night traffic volume predictions are provided in Table 4. The cross-validation statistics show that the correlation between observed and predicted values is positive and close to one for both day and night volumes, which means that the predictions matched observed values from a directional and order of magnitude perspective.

Table 4 K-fold Cross-Validation Diagnostics for Transformed Day and Night Traffic Volumes Based on Selected Model 4 (Specification with Multiple Covariates)

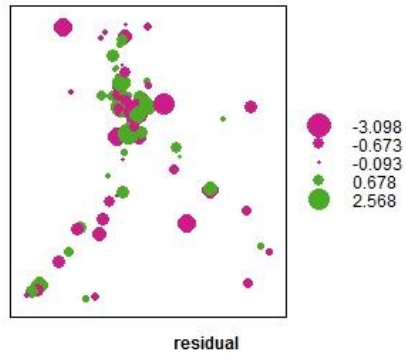
K-fold Cross-Validation Diagnostics for Model 4		Day	Night
mean_error	The mean of cross-validation residual	-0.0271	-0.0366
me_mean	Mean error divided by the mean of observed values	-0.0029	-0.0050
MSE	Mean squared error	0.6674	0.8238
MSNE	Mean squared normalized error	1.1230	1.1340
cor_obspred	Correlation between observed and predicted values	0.7963	0.7918
cor_predres	Correlation between predicted and residual values	-0.0187	-0.0020
RMSE	Root mean squared error of residual	0.8169	0.9076
RMSE_sd	RMSE divided by the standard deviation of observed values	0.6023	0.6083
URMSE	Unbiased root mean squared error of residual	0.8165	0.9069
iqr	Interquartile range of residuals	0.8149	0.9214

The diagnostics also show that the correlation between predicted and residual values is close to zero, and the mean of the cross-validation residuals is very small, and close to zero. The bubble plots from the cross-validation procedure applied to Model 4 for both day and night traffic volume predictions are shown in Figure 6. These bubble plots show that most of the higher positive and higher negative values of residuals are concentrated in the Salt Lake City urban area. The smaller positive and negative values of residuals in the bubble plots fall to the south of Salt Lake City in the rural areas of the State of Utah.

Based on the semivariogram model diagnostics, kriging model predictions, and cross-validation statistics, the final models with multiple covariates (i.e., Model 4) perform reasonably well in predicting day and night traffic volumes throughout the State. As mentioned earlier, the back-transformed values of day and night traffic volume predictions (obtained from Model 4) were then used in statistical road safety regression models of expected number of crashes at rural, two-lane horizontal curves.



Cross-Validation Bubble Plot for Transformed Day Traffic Volume Data



Cross-Validation Bubble Plot for Transformed Night Traffic Volume Data

Figure 6 Cross-Validation Bubble Plots for Transformed Day and Night Traffic Volumes Based on Selected Model 4 (Specification with Multiple Covariates)

This analysis resulted in more information on the daily traffic patterns in an area, which aids in increasing the knowledge on how the expected number of crashes change as the traffic patterns change. The major disadvantage to cross-validation error estimation is that no error information was available for locations where there are no stations (i.e., where no observed traffic volume data are present). Hence, this cross-validation study provided estimations of errors for the traffic volumes where the ATR stations were present. There is still room for improvement in the predictions of day and night traffic volumes, and future directions in improving these kriging model predictions are discussed in the last chapter of this dissertation, 'Conclusions and Recommendations'.

Measurement Error Correction Approaches

This section presents a methodology for modeling expected number of crashes while accounting for measurement error in the AADT estimates. Acknowledging that both functional and structural-type measurement error correction strategies can be used to correct for measurement error, the methodology in this research focuses on functional-type measurement error approaches. Functional type approaches do not require any distributional assumptions for the unobserved (or true) explanatory variable, whereas the structural approaches do require a parametric distribution for the unobserved (or true) explanatory variable. In other words, if the unobserved variable, i.e., true value of AADT, is regarded as an unknown constant, then functional approaches should be applied. Since measurement error correction methods are scarcely applied in the field of road safety, functional type approaches are a good starting point for exploring and understanding the effects of measurement error in regression modeling of expected crash frequency, with structural approaches as logical follow-ons for future research work.

The statistical concepts and theory behind applying these functional-type approaches are discussed in this section. The most common practice in statistical research, as well as applied research, is to implement only one measurement error correction approach and to contrast the results before and after the measurement error correction is applied (*114*). Since these approaches heavily rely on assumptions, this study explores and implements two different functional-type approaches and compares the results between the two approaches.

In modeling the expected number of crashes, significant uncertainty (or measurement error) in right-hand side variables, specifically AADT estimates, has been a long-term concern. The impact of this measurement error on the ability to quantify the relationship between all explanatory variables and expected number of crashes is likely substantial and could result in significantly biased regression coefficient estimates. The fundamental prerequisite to adjust for measurement error is to clearly define the structure of the measurement error. The basic type of measurement error, i.e., *classical measurement error*, where the true value is measured with additive error, is assumed to be present in the AADT estimates in this research. In detail, the true value (X_i) of the AADT estimates (also referred to as latent variable) is unobservable, and let W_i denote the observed values of the AADT estimates, sometimes referred to as a surrogate variable. The classical measurement error model can be expressed as given in the following equation:

$$W_i = X_i + \varepsilon_i \qquad \text{Equation (5)}$$

where:

W_i = observed value of the AADT estimates for the roadway segment *i*;

X_i = unobserved (or true) value of the AADT estimates at the roadway segment *i*; and

ε_i = measurement error, an independent variable with mean zero and usually a constant variance.

Regression Calibration and Simulation Extrapolation are two functional-type approaches to measurement error analysis that are widely applied in epidemiological studies that can likely be applied to road safety research. They were applied in this study to determine the impacts of ignoring measurement errors on statistical road safety modeling results and to determine the impacts of making measurement error corrections on safety effects estimates of selected right-hand side variables. The basis and methodology of Regression Calibration and Simulation Extrapolation are described in the following sections.

Regression Calibration

Regression Calibration (RCAL) is a standard method for correcting bias in regression coefficient estimates due to measurement error. It has become popular because of its simplicity in application to any type of regression model (*122,124,183-184*). The basis for the RCAL algorithm in measurement error analysis is the construction of the calibration model for generation of estimated covariate values for the unknown true covariates (*120*). This is accomplished using replication, validation, instrumental data, or measurement error variance in the place of unknown true covariates. The unobserved covariates are represented by their predicted values, and then a standard analysis is conducted to obtain the parameter estimates. Finally, the resulting standard errors are adjusted to account for the estimation of the unknown covariates, using either bootstrap or sandwich methods (*41,120*). In summary, RCAL method is implemented by substituting the unobserved X_i with its expectation given the surrogate W_i and then performing the standard analysis (*107*).

The true covariate (or explanatory variable) subjected to measurement error is X_i, observed values for X_i are represented by W_i, covariates measured exactly are Z (x_{ij}), and the response variable is Y_i ($E[Crashes_i]$). The RCAL method replaces X_i by the regression of X_i given W_i and Z as an approximation, i.e., E (X_i|Z, W_i), and then performs a standard analysis. The regression of Y_i on E (X_i|Z, W_i) and Z then gives unbiased estimates (*185*).

Modeling and estimating the regression of X_i on (Z, W_i) requires additional data in the form of internal/external replicate observations, instrumental variables, or validation data. However, in the context of traffic volumes (i.e., AADT estimates), a calibration function cannot be constructed using replicated or validation data because the short-term counts collected in the field are not representatives of the true annual average daily traffic volume throughout the year. Also, collecting traffic volume data for the typical safety study duration (3-6 years) in rural areas is likely impractical. Hence, measurement error variance ($\sigma_u{}^2$) in Log AADT estimates was used for modeling and estimating the regression of X_i on (Z, W_i). In summary, regression calibration estimation in this study consisted of the following primary steps:

1. Use the observed values of X_i, i.e., W_i, the measurement error variance ($\sigma_u{}^2$), and the error-free variables Z , to estimate the regression of X_i;
2. The estimate of X_i obtained from the above regression is X_i*; and
3. Run a standard analysis, i.e., regress Y_i on (Z, X_i*) to obtain regression coefficient estimates that account for measurement error.

Simulation Extrapolation

SIMEX is another general measurement error correction method that shares the simplicity, generality, and approximate-inference characteristics of the RCAL method, and is suitable for problems with additive measurement error. This method is a simulation-based method of estimating and reducing bias due to measurement error (*41,125*). SIMEX consists of a ‘simulation step’ and an ‘extrapolation step’, and is particularly useful for complex models with simple measurement error structures (*114*). The estimates are obtained by adding additional measurement error to the data in a resampling-like stage, and establishing a trend (or a simple bivariate plot) of measurement error-induced bias versus the added measurement error variance via a simulation study. Once the trend is established, the final estimates are obtained by extrapolating this trend back to the case of no measurement error (*125,186*). In summary, SIMEX is a self-contained simulation method, and its computational cost is high. SIMEX estimation in this study consisted of the following steps (*187-188*):

1. In the first step, which is also known as the simulation step, additional

measurement errors are generated and added to the error-prone variable, W_i. If the existing measurement error variance of W_i is ${\sigma_u}^2$, and the additional measurement error variance added is $\boldsymbol{\lambda}{\sigma_u}^2$, then the total measurement error variance is $(1+\boldsymbol{\lambda}){\sigma_u}^2$.

2. In the second step, regression coefficient estimates for error-prone explanatory variable (W_i) are obtained for this increased measurement error and the process is repeated 1000 times (this is where Monte Carlo simulation comes into place). This is done using an algorithm that would have been used if there were no measurement error (naïve estimation).
3. After these two steps are repeated 1000 times, the average value of the estimate for each value of $\boldsymbol{\lambda}$ is calculated. The average values of the regression coefficient estimates are plotted against the variance of the additional measurement error values and the resulting graphical displays containing an extrapolated function with error-contaminated regression coefficient estimates is obtained. By default, STATA software fits a quadratic model to model the trend.
4. The fourth and final step involves extrapolation to the ideal case of no measurement error, where the value of $\boldsymbol{\lambda}$ = -1 (i.e., total measurement error variance equals zero). This results in a SIMEX regression coefficient estimate for all the explanatory variables, when there is no measurement error.

As mentioned earlier, an estimate of the measurement error variance was required before the two measurement error correction methods were applied. This estimate is usually derived by considering deviations from a 'gold standard' value, through collecting additional data. A gold standard value or the feasibility of collecting additional traffic volume data in rural areas is not likely to be available in road safety studies. Hence, a good starting point was to use previous research and findings on the measurement error variance in AADT estimates. Ezra Hauer identified one of the studies that estimated the percent coefficient of variation, using number of count days, and AADT estimates as shown in the following equation (*27*):

$$v = 1 + \frac{7.7}{(number\ of\ count\ days)} + \frac{1650}{AADT^{0.82}} \qquad \text{Equation (6)}$$

where:

v = coefficient of variation, i.e., ratio of standard deviation and the mean; and

AADT = annual average daily traffic.

The general formula for coefficient of variation in lognormal distribution is as follows:

$$v = \sqrt{e^{SD^2} - 1} \qquad \text{Equation (7)}$$

where:

v = coefficient of variation of W_i; and

SD = standard deviation of log (W_i), where $W_i$is the value of the observed AADT estimates for a roadway segment *i*.

Using Equations 6 and 7, the coefficient of variation is determined using the number of count days and the observed AADT estimates (i.e., W_i), and the standard deviation of Log AADT estimates is back calculated. The squared value of the standard deviation was the measurement error variance, ${\sigma_u}^2$ that was used to apply measurement error correction methods in this research. However, the information on the number of count days for AADT estimates of the roadway segments was not available to calculate the coefficient of variation objectively (i.e., using the real data). Since the values of the variables for the estimation of measurement error variance were unclear, a sensitivity analysis is conducted using the count day values from 2 to 7 days, along with the available AADT estimates to calculate the measurement error variance.

This resulted in several coefficients of variation values, and back calculating those values resulted in the measurement error variance values that fell in between the values of 0.05 and 0.20. Hence, this research evaluated the impacts of making measurement error corrections on safety effects estimates of explanatory variables by using the measurement error variance values (i.e., in AADT estimates) of 0.05 to 0.20, with increments of 0.05, resulting in four different values in this part of the analysis. The hypothesis is that the application of these two measurement error correction approaches, RCAL and SIMEX, will correct the bias in all the explanatory variables, caused by error-prone AADT estimates in regression models of expected crash frequency.

A convenient feature of the two measurement error approaches is that standard software can be used for estimation. The analysis was implemented using the ‘rcal’ function and ‘simex’ function in STATA software, respectively, and the known measurement error variance was defined as a matrix that was used to generate the parameter estimates and standard errors accounting for measurement error in AADT estimates. In other words, the measurement error variable (i.e., in AADT estimates) with mean zero and constant measurement error variance (four different values) was employed in this research along with RCAL and SIMEX methods to correct for the bias due to measurement error.

The inference for standard error estimation for both measurement error correction approaches in this research was done using the bootstrap method. Bootstrapping technique refers to any test that relies on random sampling with replacement. The standard errors are assigned to the sample parameter estimates and this technique results in an estimation of the sampling distribution of the standard error of the parameter estimates. In this research, the standard error estimates were obtained from 1000 bootstrap simulations of the data. It is worth noting that, specifically for SIMEX, the bootstrapping was particularly time consuming, particularly given the simulation step generating 1000 Monte Carlo simulations for the measurement error variance before estimating the standard errors via the bootstrap method.

The major difference between the two approaches is that the RCAL approach attempts to estimate the unknown true covariate and then run a standard analysis using the approximant in the place of unknown true covariate. SIMEX simulates data in order to see the effect of measurement error on the regression coefficient estimates, and then extrapolates the trend back to the results where the covariate has no measurement error. The parameter estimates and standard errors obtained from both measurement error correction approaches were presented in the 'Results' chapter and were useful in understanding the strengths and limitations of both the approaches. As mentioned earlier, these methods were used to evaluate the impacts of making measurement error corrections on the safety effects estimates of selected explanatory variables in regression models of expected crash frequency.

Safety Modeling Framework for Assessing Model Impacts of Traffic Pattern Estimates and Measurement Error Corrections

This section explains the basic safety modeling framework that was used to assess the potential modeling impacts of findings from the first two efforts on traffic pattern estimates and measurement error corrections. The method for modeling expected crash frequency, given the fact that crashes are count outcomes and the variance of crash counts is almost always greater than the mean (or overdispersion), is presented in this section. The modeling process was primarily focused on including information related to roadway geometrics (i.e., horizontal and vertical curve attributes), operational characteristics (i.e., traffic volume information), roadside features, and weather conditions, which are assumed to potentially have an impact on road safety. The datasets containing the explanatory and response variables required for road safety analysis are described in the next chapter, ‘Data Collection’.

Modeling Crash Data

The process of modeling crash data started from most frequently used regression models of expected crash frequency. This dissertation describes the process of variable inclusion or elimination to obtain the ‘best’ model specifications for the dependent variables (i.e., crash outcomes). Since crashes are count outcomes, the Poisson regression model was a common starting point in the crash frequency modeling process. These regression models were used by road safety researchers for crash frequency analysis for several decades. The Poisson regression models assume that the mean and variance of the crash frequency distribution are equal. In a Poisson regression model, the probability of a roadway segment having crashes per some time period is given by the following equation:

$$P(y_i) = \frac{EXP(-\mu_i){\mu_i}^{y_i}}{y_i!} \qquad \text{Equation (8)}$$

where:

$P(y_i)$ = probability of a roadway segment *i* having y_i crashes per some time period;

μ_i = Poisson parameter for a roadway segment i, which is equal to roadway segment's expected number of crashes per some time period (i.e., E[y_i]; and

$$\mu_i = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}) \qquad \text{Equation (9)}$$

where:

μ_i = dependent variable, expected number of crashes for a roadway segment i;

x_{ij} = independent, explanatory variables specifying traffic, geometric, and other characteristics of a roadway segment i;

β_j = regression parameters to be estimated that quantify the relationships between the explanatory variables and expected number of crashes (i.e., μ_i); and

β_0 = intercept (or constant term).

Although road safety researchers have earlier used Poisson model for crash frequency analysis, they often found that the characteristics exhibited by crash data make the application of Poisson model problematic. Specifically, Poisson regression models do not handle overdispersion (variance greater than the mean) and can produce biased results, if used otherwise (*36*). The negative binomial (NB) regression model is an extension of the Poisson model to account for the presence of possible overdispersion in the data. However, the NB models cannot handle underdispersed data and pose problems in the estimation of the dispersion parameter when the data are characterized by low sample mean values and small sample sizes (*29,36*).

The negative binomial regression model is derived by rewriting the Poisson parameter as follows:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \varepsilon_i) \qquad \text{Equation (10)}$$

where:

μ_i = dependent variable, expected number of crashes for a roadway segment i;

x_{ij} = independent, explanatory variables specifying traffic, geometric, and other characteristics of a roadway segment i;

β_j = regression parameters to be estimated that quantify the relationships between the explanatory variables and expected number of crashes (i.e., μ_i);

β_0 = intercept; and

ε_i = random error term, where exp (ε_i) is a gamma-distributed error term with mean 1 and variance α. In the NB-2 model, the variance in the number of crashes is written as $\mu_i + \alpha{\mu_i}^2$, with α referred to as the dispersion parameter.

The negative binomial regression model was used to help answer the first and second research questions in the dissertation on the modeling impacts of findings from the first two efforts on traffic pattern estimates and measurement error corrections. The analysis for these two studies seeks to compare general findings, parameter estimates, and other model properties resulting from estimation of the regression models for expected number of crashes. For the study on daily traffic patterns, the negative binomial regression model was used to explore the magnitude and direction of parameter estimates for variables representing new information (i.e., day and night traffic volumes), to better understand the impacts of these estimates on expected number of crashes. The negative binomial regression model was also useful to compare the model performance (i.e., pseudo R-squared value, dispersion parameter) for models with and without the new information, to see if the model benefits from including the new information.

For measurement error corrections, the intent of using negative binomial model was to attempt to characterize the impacts and benefits of using functional type measurement error correction approaches in regression models of expected crash frequency. This was done by comparing parameter estimates that quantify the relationships between the expected number of crashes and several explanatory variables, and the standard errors of the estimates in the regression models. The whole focus of the methodologies and interpretation in this dissertation was not only on the specific parameter estimates themselves (i.e., the magnitude of the parameter estimates), but was also on exploring and demonstrating the effect of application of these alternative approaches on the results of observational road safety studies.

Bayesian Framework with Informative Priors

This part of the research, as mentioned earlier, was motivated by an issue frequently encountered in road safety research. A majority of the road safety studies let the statistical diagnostics drive the functional form, estimation results drive the interpretations, and ‘start over’ with each study to see which model functional form fits the existing data best. In other words, statistical road safety modelers usually rely only on the data at hand to estimate model coefficients; they do not explicitly incorporate prior knowledge and end up ‘starting over’ the model specification and estimation process with each subsequent study. Hence, a choice of appropriate model functional form that provides reliable associations between explanatory variables and crash outcomes is less clear in road safety research (*34*).

Accommodating accumulated knowledge from past research is virtually impossible with conventional frequentist approaches unless the past and present data are combined to make it one complete dataset (*134*). This section presents a Bayesian approach, in which the current data along with prior information are incorporated into the analysis to make stronger statistical inferences on the parameter estimates of the selected right-hand side variables. This is hoped to ultimately help in drawing inferences on converging functional forms in regression models of expected crash frequency. Unlike in the frequentist approaches, the incorporation of prior information, i.e., expert opinions or criteria, and information from previous rigorous and well-defined observational studies is possible in the Bayesian framework through assigning prior distributions to the parameters. This portion of the research explored how incorporating prior knowledge of the safety effects of explanatory variables (i.e., regression coefficient estimates) into Bayesian model formulations yielded different conclusions with respect to model prediction results and safety effects estimates of selected right-hand side variables.

A large portion of previous road safety studies that employed a full Bayes approach as an alternative to the maximum likelihood estimation used noninformative priors to analyze the crash data. This lets the data 'speak for itself' and does not involve any subjectivity in the model estimation. However, there are examples of successfully incorporating informative priors with respect to the inverse dispersion parameter, and countermeasure effectiveness reported in previous studies, involving subjectivity in the model estimation (*149,189-190*).

These authors found that specifying informative priors from previous studies can result in more robust estimates under certain circumstances. However, these estimated priors changed from one study (or researcher) to another, because there is some level of subjectivity involved in selecting and weighting the studies that produce an informative prior (*190*). Consequently, the statistical inferences drawn from these different informative priors will be different. Because of this, some authors have stated that the complexity of Bayesian methods, specifically with informative priors, is a serious barrier to their application (*36*).

This dissertation further explored the use of prior knowledge along with the current data in interpreting the safety effects estimates of selected right-hand side variables and model prediction results. Keeping the problem statement in mind, Bayesian approach with an informative prior on the model parameters was employed in the current study. However, the noninformative and semi-informative priors were also employed in this study to draw comparisons between the results obtained from different types of priors. The Bayesian logic combines the 'subjective' or prior knowledge, typically in the form of statistical prior distributions, with 'objective' current information or data to derive meaningful posterior distributions. Bayesian statistics is built on Bayes' rule which defines the change in the probability of an event A after another event B occurs, and the philosophy of regarding the model parameters as random variables.

Given a model parameter θ, the random variable X follows a distribution with density, $f(x|\theta)$ (*191*). In the Bayesian analysis approach, the posterior distribution of θ given x is proportional to the product of the prior distribution of parameter θ and the likelihood (or sample information). The relative weights of the likelihood and prior are determined by the variances of the prior distributions, with smaller variance resulting in a greater weight of the prior towards determining the posterior (*192*). Furthermore, the conditional distribution of θ given the sample observations (or data) x is given as the posterior distribution of θ given x, denoted by $\pi(\theta|x)$, shown in the following equation:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \qquad \text{Equation (11)}$$

where:

$\pi(\theta)$ = prior distribution of parameter θ;

$f(x|\theta)$ = likelihood function; observed data given parameter θ;

$\pi(\theta|x)$ = posterior distribution of parameter θ given observed data; and

the denominator represents the marginal likelihood.

The process of making posterior inferences takes advantage of the Markov Chain Monte Carlo (MCMC) methods to overcome the computational complexity and difficulties in the Bayesian analysis approach (*190*).

The Bayesian analysis approach in this research assumes a Poisson-Gamma model for the analysis of crash data. The Poisson-Gamma model can be described mathematically as shown in the following equation:

$$y_i = \text{Poisson}(\mu_i) \qquad \text{Equation (12)}$$

where:

y_i is the observed number of crashes for a roadway segment *i*; and

μ_i is the expected crash frequency for a roadway segment *i*.

However, the expected crash frequency can be expressed as a function of the contributing factors for crash frequency and a multiplicative random effects component, which is shown in the following equation (*149,190*):

$$\mu_i = \theta_i * \tau_i \qquad \text{Equation (13)}$$

where:

θ_i = a function of the contributing factors for crash frequency; and

τ_i = multiplicative random effect that is assumed to be gamma distributed with mean 1 and variance of $1/\emptyset$; where $\emptyset$ is the inverse dispersion parameter (*193*).

In addition, the inverse dispersion parameter is also assumed to follow a gamma distribution with shape and scale parameters, a and b, respectively (i.e.,

$\emptyset \sim \text{gamma}(a, b)$). The mean and variance for the inverse dispersion parameter are as given in the following equations:

Mean $(\emptyset) = ba$ Equation (14)

Variance $(\emptyset) = b^2 a$ Equation (15)

where:

a, b are the shape and scale parameters, respectively.

Given that the inverse dispersion parameter follows gamma distribution, the dispersion parameter will follow an inverse gamma distribution, i.e., $\frac{1}{\emptyset} \sim \text{Inverse gamma}(a, \frac{1}{b})$). The mean and variance for the dispersion parameter are as shown in the following equations:

Mean $\left(\frac{1}{\emptyset}\right) = \frac{(\frac{1}{b})}{(a-1)}$ Equation (16)

Variance $\left(\frac{1}{\emptyset}\right) = \frac{(\frac{1}{b})^2}{(a-1)^2(a-2)}$ Equation (17)

where: a, 1/b are the shape and scale parameters, respectively.

These Poisson-Gamma models differ from the traditional Poisson regression in the sense that the Poisson-Gamma models can overcome the overdispersion issue that often exists in the crash data.

As mentioned earlier, the Bayesian philosophy asserts that almost always something is known or expected about model parameters before estimation (*192*). A substantial amount of published research is available that quantifies the relationship between some of the explanatory variables and expected number of crashes. Among all the explanatory variables, some of the variables are used more often than others in safety models because road safety researchers have access to the data pertaining to these variables. The variables include traffic volume, segment length, shoulder width, lane width, and horizontal curvature attributes like radius or degree of curvature among many others (*164-166,194-195*). In this study, the prior distributions for the selected explanatory variables were constructed using accumulated knowledge, i.e., regression coefficient estimates of the selected explanatory variables from previous rigorous and well-defined observational studies. The data and prior distributions corresponding to these selected explanatory variables were used to apply Bayesian analysis using the MCMC methods, and WinBUGS software. The methodology of constructing prior distributions from the previous studies is described in detail in the next section.

Incorporating Prior Information

Incorporating prior information is one of the most critical issues in Bayesian analysis. After having the cumulative body of past research, a decision has to be made as to how to include this information (i.e., regression coefficient estimates of explanatory variables from previous studies) in the form of a prior in Bayesian analysis. One option that is used by (*196*) in the field of clinical trials and that has been applied in the current study was to assume that all the regression coefficient estimates obtained from different studies, ρ_i, for a particular variable are similar to a common value ρ_c. Under this assumption, the ρ_i may be regarded as independently drawn from a common random distribution. In this case, it was appropriate and convenient to assume that the observed regression coefficients for a particular explanatory variable were drawn from a normal distribution, which is shown in the following equation:

$$\rho_i \sim N(\varphi, \omega^2) \qquad \text{Equation (18)}$$

where:

ρ_i are the regression coefficients of a particular explanatory variable obtained from previous studies;

φ is the mean; and ω^2 is the square of the standard error of the sample containing the regression coefficients estimates from previous studies.

Three types of priors were used in this study for employing Bayesian analysis: Noninformative priors, semi-informative priors, and informative priors. The detailed description on the three types of priors is given in the below section.

a) Noninformative Priors

These priors are also called vague priors, and are the most commonly used priors in the field of road safety research. These priors carry virtually no information about the likely true value of a parameter. They preserve the objectivity of the analysis, by giving maximum weight to the likelihood and minimum weight to the prior in the determination of the posterior distribution. In other words, the data itself will lead to the estimation of the parameter estimates of right-hand side variables and the importance of the prior will be minimized.

For the right-hand side variables mentioned above (i.e., traffic volume, segment length, shoulder width, lane width, and degree of curvature), noninformative priors were generated and used in this research. According to the previous literature, for regression parameters representing the explanatory variables and the intercept (or constant term), a normal distribution with mean zero and a large variance of 1000 was used to signify the noninformative prior. Similarly, the inverse dispersion parameter with mean value of one and a large variance 1000 (i.e., shape parameter of 0.001 and scale parameter of 1000) were used in these types of priors for Bayesian methodological framework (*190*).

b) Semi-informative Priors

In this study, a semi-informative (or weakly informative) prior was defined as a prior distribution that carried more information than a noninformative prior, but deliberately carried a smaller degree of information than was actually available. These priors were not considered to be overly objective, like the noninformative priors, but were not considered as totally subjective either. These priors give a majority of the weight to the likelihood (or data) and minimum weight to the prior in the determination of the posterior distribution. The purpose of using semi-informative priors rather than noninformative priors was typically to achieve some stabilization in the MCMC sampling estimation procedure. In this study, for the explanatory variables and intercept term, a normal distribution with the mean calculated from the sample of previous studies, and a variance of 100 was used to signify the semi-informative prior.

For the inverse dispersion parameter, a gamma distribution with the mean value calculated from the sample of previous studies, and a variance of 100 was used in these types of priors. More information on the calculation of mean of the parameter estimates for each of the explanatory variables, intercept, and inverse dispersion parameter from previous studies is given in the 'selection and weightage assignment for previous studies' section.

c) Informative Priors

An informative prior was defined as a prior distribution that carries a distinguishable and larger degree of information than a semi-informative prior. The purpose of using these types of priors was to fully utilize and mix prior (or external) knowledge about the parameters with the data. These priors introduce subjectivity in the analysis, by incorporating prior information on parameter estimates of selected explanatory variables, intercept term, and inverse dispersion parameter. The informative priors usually have a greater impact on the posterior distributions when compared to the noninformative or semi-informative priors. The subjective nature and the greater impact of these priors on the posterior distributions and results are the reason for their underutilization in the field of road safety research. In this study, for the explanatory variables and intercept term, a normal distribution with the mean and variance calculated from the sample of previous studies was used to signify the informative prior. Similarly, for the inverse dispersion parameter, a gamma distribution with the mean and variance calculated from the sample of previous studies was used to define the informative prior. The selection and weightage assignment for the previous studies which were used to calculate the mean and standard error for the selected explanatory variables, intercept, and inverse dispersion parameter is illustrated in the next section.

Selection and Weightage Assignment for Previous Studies

The studies were first selected based on the condition that the study setting was rural, two-lane highways, and the safety models included the selected explanatory variables for the analysis of expected total crash frequency. The Bayesian methodology incorporating semi-informative and informative priors consisted of the following two steps:

1. *Study Selection and Weightage*: select previous research involving the selected explanatory variables and weight the studies based on the methodology and

statistical rigor; and

2. *Weighted Statistic Calculation*: calculate the weighted mean and standard error (i.e., square root of the variance) for the assumed prior distribution for the sample of parameter estimates for selected explanatory variables, intercept term, and inverse dispersion parameter.

For the first step, this study assumed that all the available regression coefficient estimates from the previous studies differ in their quality, confidence in the results, and relevance to the study setting. Hence, to account for these differences and give more weightage to the most relevant studies and lesser weight to the lesser relevant previous work, this dissertation employed a subjective rating approach adapted from the Crash Modification Factors (CMF) clearing house star rating system (*197*). According to this system, the star rating was based on a scale (1 to 5), where a 5 indicates a statistically rigorous and robust study with highest reliability and a 1 indicates the least reliable study.

The rating was based on five different categories, with each category having three different options on which the score is based. The five different categories were: 1) study design, 2) sample size, 3) standard error, 4) potential bias, and 5) data source, with the three options for each category being excellent, fair, and poor (*197*). For more detailed information on the relative rating and categories, please refer to the CMF clearing house website. The total subjective score was calculated from the following equation and star rating (i.e., 1 to 5) was given as per the scores obtained for each previous study (*197*).

$$\text{Score} = (2 * \text{SD}) + (2 * \text{SS}) + \text{SE} + \text{PB} + \text{DS} \qquad \text{Equation (19)}$$

where:

SD = Study Design;

SS = Sample Size;

SE = Standard Error;

PB = Potential Bias; and

DS = Data Source

After assigning the star rating (from 1 to 5) for all the previous studies selected, the study weights were assigned based on the CMF star rating score. In other words, if the star rating for a study was 5, a weightage of 1 was given for the study. The study with a star rating of 1 was given a weightage of 0.2, i.e., the weights were assigned with decrements of 0.2 from a star rating of 5 to 1. Hence, a study with a star rating of 5 (i.e., a weight of 1) indicates a rigorous and reliable study, and the study with a star rating of 1 (i.e., a weight of 0.2) indicates a weak and less reliable study. This way, the most relevant studies were given more weightage when compared to the less reliable studies, while calculating the mean and standard error for the sample parameter estimates obtained from those studies, which is the second step. The weighted mean and the weighted standard error for the parameter estimates were calculated for all the variables, using the following equation:

$$P_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \qquad \text{Equation (20)}$$

where:

P_w = weighted mean, or weighted standard error;

w_i = weight for the study with ith observation;

x_i = value for the ith observation; and

N = number of total observations (i.e., regression parameter estimates and standard errors, in this case).

In this process of calculating the weighted mean and weighted standard error, the regression coefficient estimates along with the standard error (or t-statistic) were obtained from the previous relevant studies. For some studies with no standard error information for the parameter estimate, t-statistic was used to calculate the standard error (i.e., standard error being the ratio of parameter estimate and t-statistic). The studies with the parameter estimates, standard errors for selected explanatory variables, intercept term, and dispersion parameter, along with the CMF scores, star rating values, and the weightage for the previous studies calculated based on the above mentioned approach are given in Table 5. The error bar plots containing the mean and the standard errors for the selected explanatory variables, intercept, and dispersion parameter are shown in Figure 7 through Figure 10.

MCMC Techniques and Model Fit

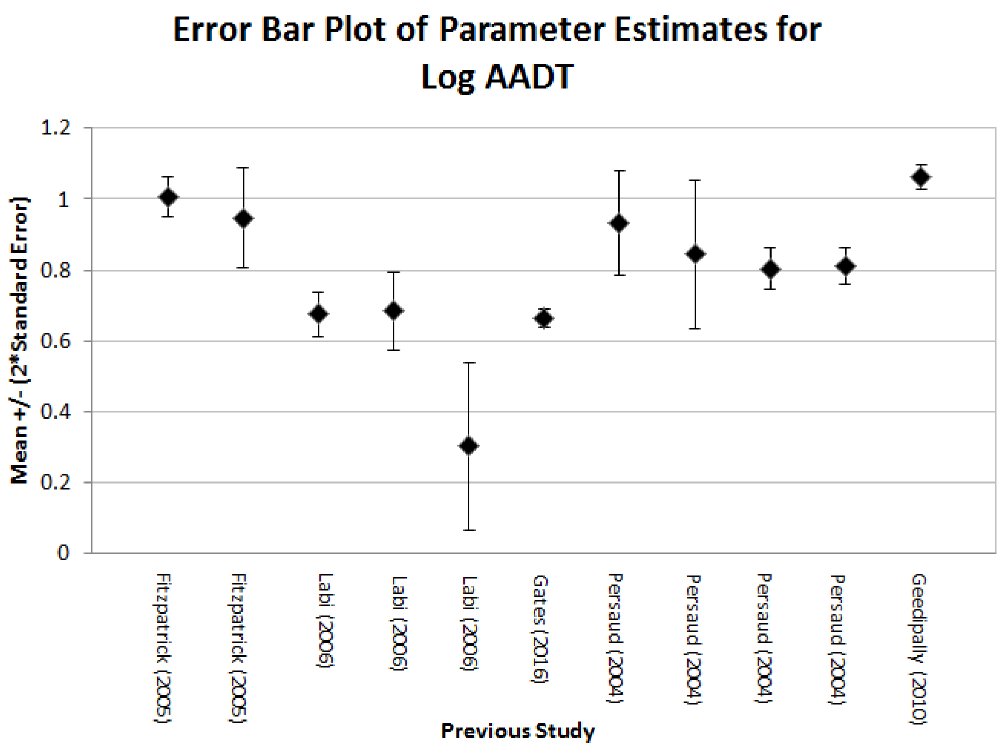
In this study, the Bayesian models incorporating three different types of priors were estimated using Markov Chain Monte Carlo (MCMC) methods. MCMC is a sampling-based approach to estimation that is well suited for Bayesian models (*192*). MCMC techniques provide a way of simulating from complex distributions by simulating from Markov chains which have the target distributions as their stationary distributions (*198*). The Gibbs sampler (a widely used MCMC technique), which enables simulation from multivariate distributions by simulating only from the conditional distributions, was used in this research. In each case of the type of prior, 11000 iterations of the algorithm were carried out, of which the first 4000 iterations were regarded as burn-ins and discarded from the simulations. These burn-ins were used to check if the chains had converged, before the posterior distributions were estimated.

Table 5 Summary of Previous Research and Weights Based on CMF Clearing House Score and Star Rating

Authors/Brief Study Description	Coeff. Estimate	Std. Error	CMF Score/ Star Rating/ Weightage	Weighted Mean	Weighted Standard Error
			Logarithm of AADT		
Fitzpatrick et. al. (2005)	1.004	0.028	7 / 3 stars / 0.6	**0.799**	**0.052**
	0.947	0.070			
Labi (2006)	0.675	0.032	7 / 3 stars / 0.6		
	0.683	0.056			
	0.301	0.118			
Gates et. al. (2016)	0.664	0.013	9 / 3 stars / 0.6		
Persaud et. al. (2004)	0.933	0.074	12 / 4 stars / 0.8		
	0.844	0.105			
	0.803	0.029			
	0.811	0.026			
Geedipally et. al. (2010)	1.063	0.017	7 / 3 stars / 0.6		
			Logarithm of Segment Length		
Fitzpatrick et. al. (2005)	0.851	0.024	7 / 3 stars / 0.6	**0.874**	**0.035**
	0.814	0.047			
Labi (2006)	0.893	0.041	7 / 3 stars / 0.6		
	0.999	0.081			
	0.753	0.066			
Gates et. al. (2016)	0.907	0.009	9 / 3 stars / 0.6		
Persaud et. al. (2004)	0.834	0.043	12 / 4 stars / 0.8		
	0.887	0.013			
	0.913	0.015			
			Shoulder Width		
Fitzpatrick et. al. (2005)	-0.060	0.007	7 / 3 stars / 0.6	**-0.057**	**0.011**
Harwood et. al. (2000)	-0.059	0.011	10 / 3 stars / 0.6		
Labi (2006)	-0.017	0.009	7 / 3 stars / 0.6		
	-0.039	0.013			
	-0.073	0.018			
Persaud et. al. (2004)	-0.036	0.015	12 / 4 stars / 0.8		
	-0.071	0.005			
Geedipally et. al. (2010)	-0.103	0.004	7 / 3 stars / 0.6		
			Degree of Curvature		
Harwood et. al. (2000)	0.045	0.008	10 / 3 stars / 0.6	**0.044**	**0.006**
Gooch et. al. (2016)	0.054	0.002	10 / 3 stars / 0.6		
Momeni (2016)	0.034	0.006	7 / 3 stars / 0.6		
	0.047	0.008			
			Lane Width		
Fitzpatrick et. al. (2005)	-0.137	0.033	7 / 3 stars / 0.6	**-0.157**	**0.065**
Harwood et. al. (2000)	-0.084	0.042	10 / 3 stars / 0.6		
Labi (2006)	-0.068	0.018	7 / 3 stars / 0.6		
	-0.093	0.032			
	-0.108	0.056			
Geedipally et. al. (2010)	-0.155	0.029	7 / 3 stars / 0.6		
Tarko et. al. (1999)	-0.453	0.251	5 / 2 stars / 0.4		
			Dispersion Parameter		
Fitzpatrick et. al. (2005)	0.421	0.036	7 / 3 stars / 0.6	**0.310**	**0.029**
	0.273	0.058			
Harwood et. al. (2000)	0.305	0.033	10 / 3 stars / 0.6		
Labi (2006)	0.257	0.016	7 / 3 stars / 0.6		

Table 5 continued

Authors/Brief Study Description	**Coeff. Estimate**	**Std. Error**	**CMF Score/ Star Rating/ Weightage**	**Weighted Mean**	**Weighted Standard Error**
		Dispersion Parameter			
Labi (2006)	0.218	0.019			
	0.202	0.026			
Geedipally et. al. (2010)	0.496	0.020	7 / 3 stars / 0.6		
Inverse Dispersion Parameter	*Cross-calculation*			***3.249***	***0.309***
		Intercept (Constant Term)			
Fitzpatrick et. al. (2005)	-5.098	0.397	7 / 3 stars / 0.6	**-5.940**	**2.355**
	-6.780	0.571			
Labi (2006)	-4.105	0.239	7 / 3 stars / 0.6		
	-3.910	0.557			
Gates et. al. (2016)	-5.292	0.106	9 / 3 stars / 0.6		
Persaud et. al. (2004)	-7.432	0.685	12 / 4 stars / 0.8		
	-6.541	0.845			
	-6.973	0.236			
	-5.817	0.218			
Cafisco et. al. (2007)	-5.861	34.09	5 / 2 stars / 0.4		
Geedipally et. al. (2010)	-6.462	0.306	7 / 3 stars / 0.6		



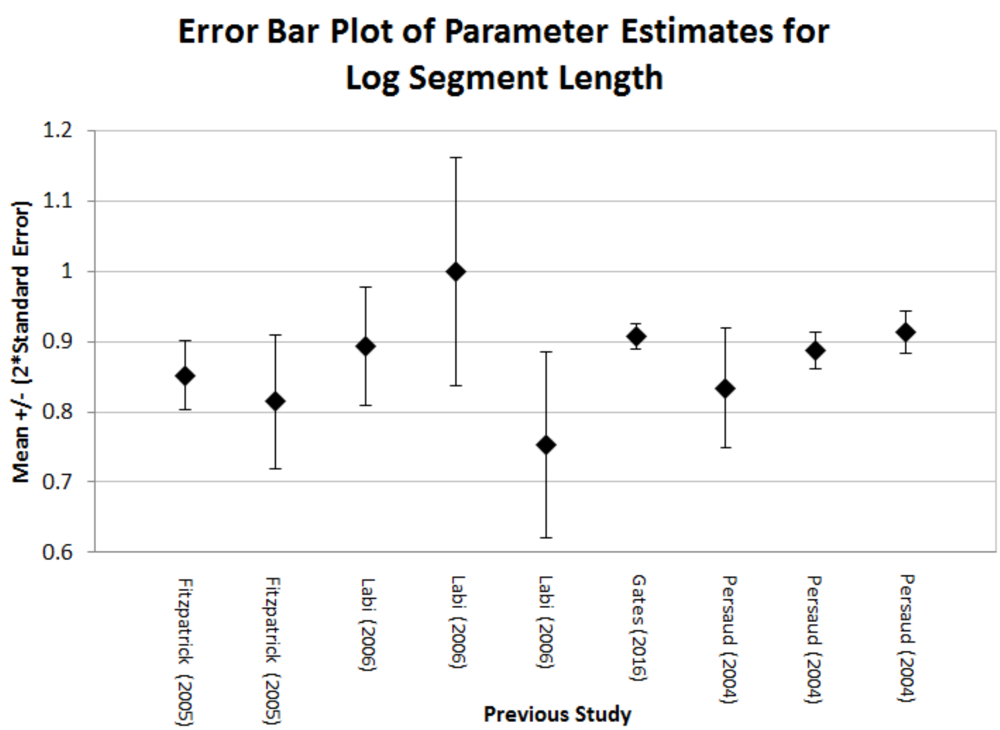
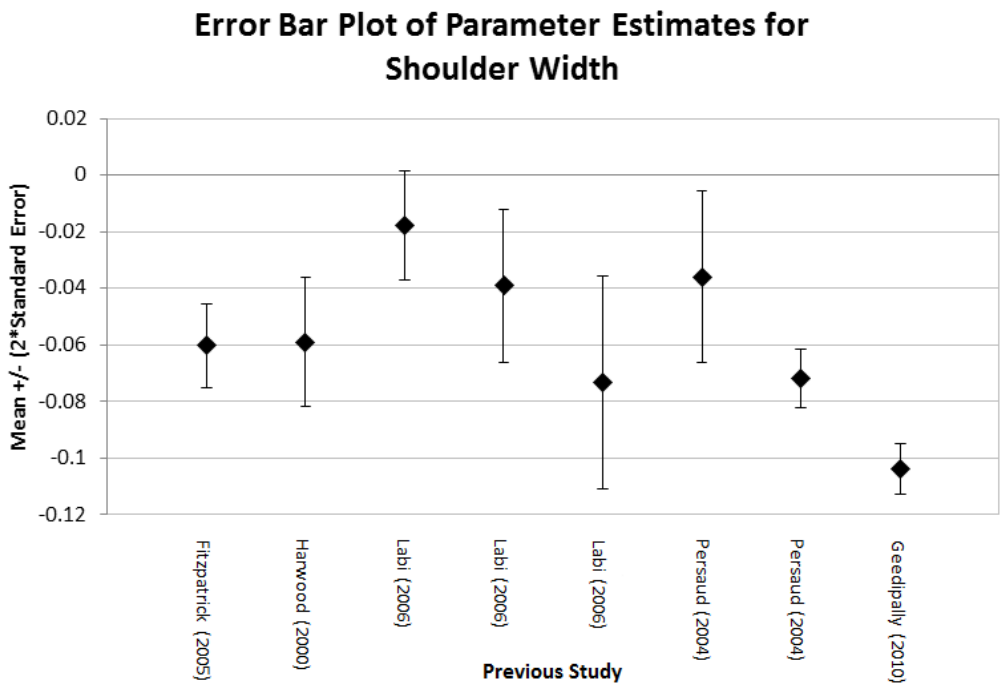


Figure 7 Error Bar Plots of Parameter Estimates for Log AADT and Log Segment Length from Previous Studies



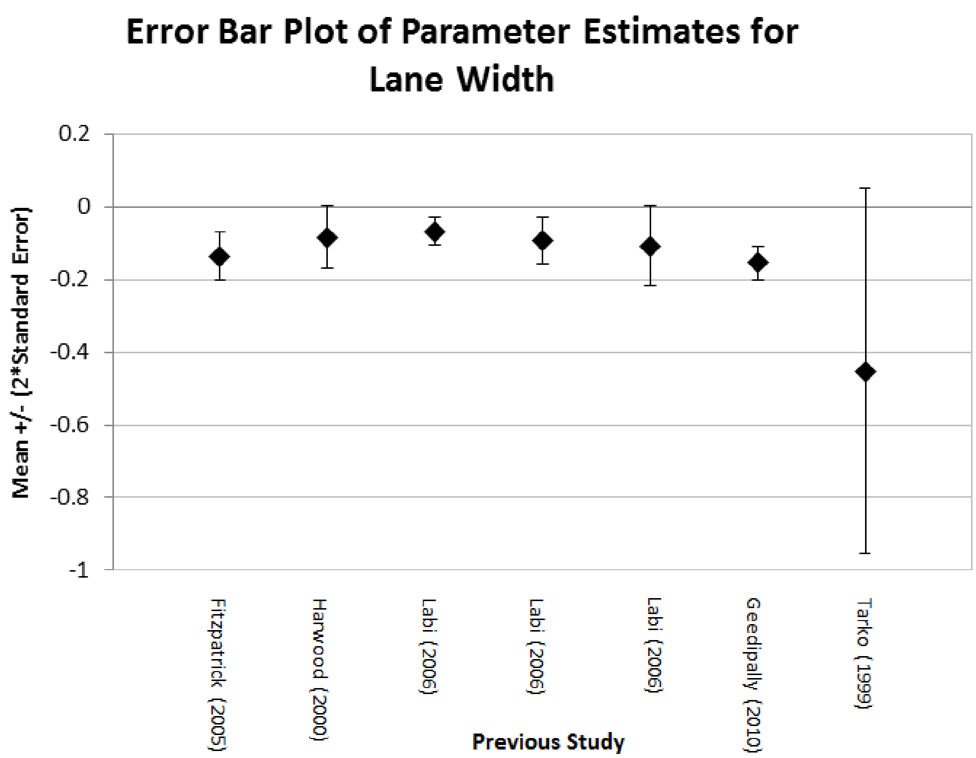
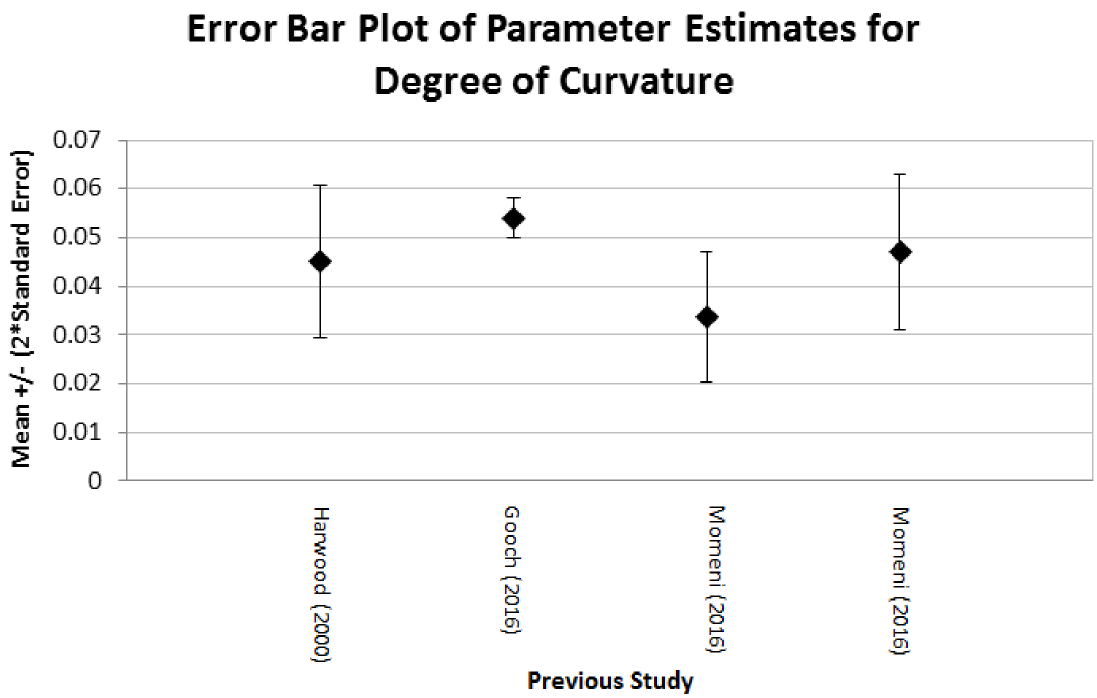


Figure 8 Error Bar Plots of Parameter Estimates for Shoulder Width and Lane Width from Previous Studies



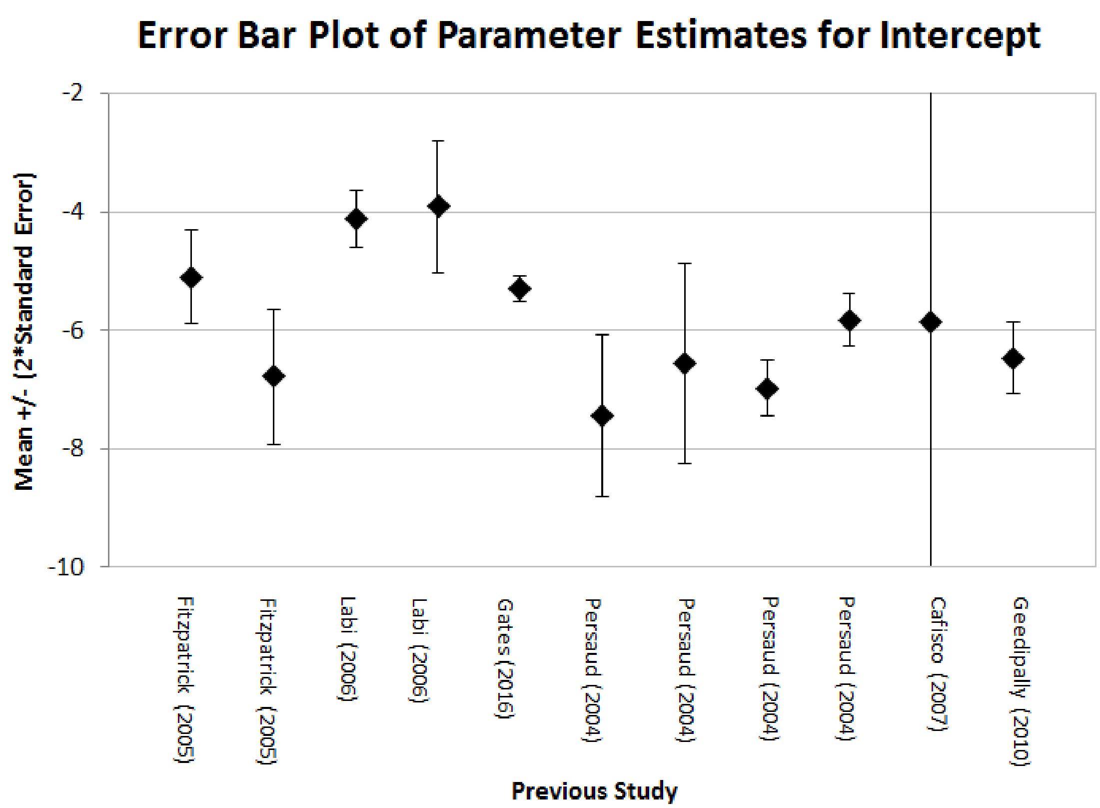


Figure 9 Error Bar Plots of Parameter Estimates for Degree of Curvature and Intercept Term from Previous Studies

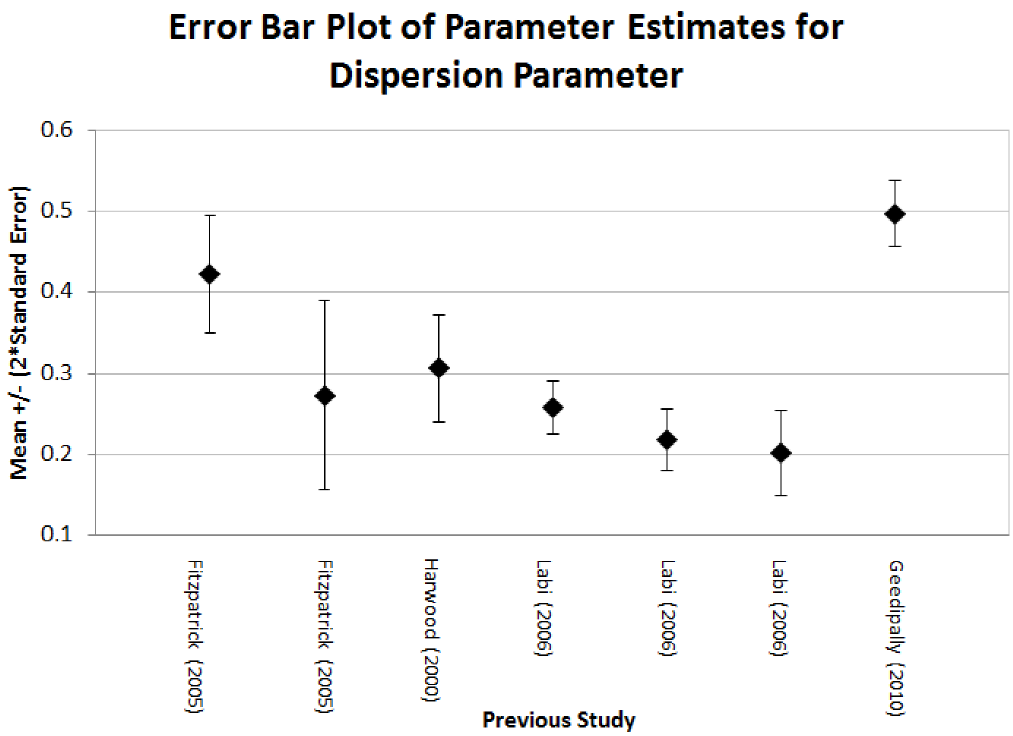


Figure 10 Error Bar Plot of Parameter Estimates for Dispersion Parameter from Previous Studies

The final model selection, from all the Bayesian models incorporating the different types of priors was aided by tools that help quantify the goodness-of-fit, which eventually helped in making useful comparisons across models. In this research, the deviance information criterion (DIC), which is the most popular criterion for Bayesian model selection and model comparison, was used to compare and select the "best" fitting Bayesian model. In other words, the DIC is a measure of model fit computed from the likelihood function with a penalty for complexity. Complexity is measured as the 'effective number of parameters in the model'. DIC is measured by the sum of deviance and twice the number of effective model parameters in the Bayesian model, which is shown in the following equation (*192*):

$$\mathrm{DIC} = D(\bar{\theta}) + 2pD \qquad \text{Equation (21)}$$

where:

$D(\bar{\theta})$ = the deviance of $\bar{\theta}$ as an expectation of θ; and

pD = the number of effective parameters in the model.

This parameter (i.e., DIC) was used to assess Bayesian model goodness-of-fit. A lower value of DIC indicates a better model fit. A difference of 5 between the DIC values does not indicate significant difference; however, a difference of a scale of 100 between the DIC values indicates that the model with lesser DIC value is the best model. The Bayesian analysis results obtained from this study (presented in the ‘Results’ chapter) can be further used to obtain more information regarding fully utilizing previous study results, instead of starting over with each study.

CHAPTER 4

DATA COLLECTION

The methods described in the previous chapter involved a prescription of needed data elements to carry out the statistical analysis. This chapter of the dissertation provides a detailed explanation on the site selection, data sources, and the data collection process of the variables required to employ the methods discussed in the previous chapter. The first part of this chapter presents details on the site selection and the states selected for analysis, and the second part of this chapter presents the list of variables that were collected for the safety analysis.

Site Selection

According to the Federal Highway Administration (FHWA), there are currently 4.09 million miles of roads in the United States, of which 2.98 million miles (72 percent) are located in the rural areas (*199*). Two-lane highways account for nearly 80 percent of rural roads and are critical for providing both mobility and accessibility to rural residents, agriculture, and industry (*53*). Average trends from 2003 to 2013 indicate that traffic fatalities on the nation's rural highway network accounted for about 55 percent of all traffic related deaths on the entire road network in the country (*200*). Specifically, in the year 2013, the crashes on the nation's rural, noninterstate routes resulted in 15,601 fatalities, which accounted for nearly half (48 percent) of the nation's 32,719 traffic deaths (*201*).

In terms of crash rates (crashes per vehicle-miles traveled), the worst performing are the rural two-lane roads with a record of 3.08 fatalities per 100 million miles of travel (*199,201*). This is obviously due to their relatively low traffic volume and consequently high severe crash frequency to traffic volume ratio for such roads. Historically, the most critical geometric design element that influences driver behavior and poses the most potential for crashes has been the horizontal curve (*202-203*). Crash rates are typically 1.5 to 4 times higher on horizontal curves than on the tangent segments (*203*). These disturbing numbers have spurred transportation safety professionals to seek more efficient and effective strategies to reduce the number and severity of these crashes. The extent of research that is possible with the available data sources made rural two-lane highways a case study for this research.

Data Sources

In order to carry out the methods outlined in the previous chapters, and to make the best use of the available data sources, data from the states of Utah and Washington were used in this analysis. Most of the data for Utah were obtained from Utah Department of Transportation's (UDOT) online data portal (including LiDAR data), UDOT Safety Team, and United States (U.S.) Census Bureau. The LiDAR data collection was executed by UDOT, and the data were openly available to the public through UDOT's online data portal for the years 2012 and 2014 (by the time this data collection was done).

The LiDAR-based dataset of roadway segments (i.e., horizontal curves in this case) differed the most from other traditional state databases in the context of location information. Usually a linear referencing system (i.e., route and milepost information) is used to locate a feature (i.e., segment or intersection) in state databases. However, with the LiDAR data, a spatial referencing system (i.e., geo-coding or latitude and longitude information) was available which could be used for employing spatial analysis. For the state of Washington, the data were obtained from the Highway Safety Information System (HSIS) database and Roadside Features Inventory Program (RFIP). The weather data for both the states were obtained from National Oceanic and Atmospheric Administration (NOAA). Table 6 includes data sources in detail, which were used to collect the data elements required for this research. For the Utah dataset, the geo-spatial information (i.e., latitude and longitude) was available for the horizontal curve segments. The hourly traffic volume data were also available publicly, for all the permanent traffic recorder stations through UDOT's website. Considering the spatial referencing system, data availability, and the applicability of spatial analysis methodology to the dataset, the dataset for Utah was used to answer the first research question on predicting the day and night traffic volumes for statistical road safety modeling and analysis.

Most of the safety studies have traditionally used electronically coded datasets (i.e., HSIS, state maintained databases) for road safety analysis. The dataset for Washington was developed using the HSIS data, and road safety researchers have used these kinds of databases on many occasions before. Hence, it was assumed that it was relatively easy to implement and answer the second and third research questions using the dataset for Washington, and relate to the findings obtained from this research. In summary, data availability, applicability of the methods, and variables available in each of the datasets for Utah and Washington dictated the selection of datasets to the respective research questions.

Table 6 Data Sources and Descriptions

Data Type	Description/Comment	Data Source
	State of Utah	
Crash data	Location, crash type, sequence of events	UDOT
Roadway data	Cross-section characteristics, number of lanes, functional class, horizontal features, traffic volume	LiDAR, Algorithms, and UDOT data portal
Hourly traffic volumes	Automatic Traffic Recorder (ATR) stations	UDOT
Socio-economic data	Population, and number of households	U.S Census Bureau, AGRC
Roadside data	Barrier	UDOT data portal, ArcGIS
Weather data	Temperature, snow, precipitation	NOAA
	State of Washington	
Crash data	Location, crash type, sequence of events	HSIS Database
Roadway data	Lane, shoulder, horizontal and vertical curve characteristics, traffic volume, posted speed limit	HSIS Database
Roadside data	Trees, guardrail, fixed objects	RFIP, ArcGIS
Weather data	Temperature, snow, precipitation	NOAA

Required Variables for Analysis

All crash data, roadway and geometric features, cross-section attributes, operational characteristics, roadside features, weather conditions, and socio-economic characteristics were collected using a number of different sources. The details on key parts of the data collection process and the variables used for analysis are outlined in the following sections.

Crash Data

For Utah, crash data were obtained for crashes occurring in rural, two-lane horizontal curves for 5 years, from 2009-2013. The crash data were obtained from UDOT separately from the UDOT online data portal. The location information of crashes was available through the GPS coordinates. For Washington, crash data were obtained for crashes occurring in rural, two-lane horizontal curves for 5 years, from 2008-2012. The crash data were obtained from the HSIS database accident and vehicle files. The accident and vehicle files identify the crash by case number, and the location information by state route number, and milepost. The crashes were merged onto the horizontal curve segments using the location information (i.e., latitude and longitude information for Utah; state route number and milepost information for Washington). This research does not include crashes involving pedestrians and bicyclists.

The variables contained in the crash inventory files include crash type, crash severity, time of day and day of week when the crash has occurred, weather condition, road surface condition, driver age, gender, vehicle make and type, and primary and secondary contributing factors for the crash, among many others. The traffic safety-related independent variables in this research were identified to differentiate between safety outcomes in terms of frequency. These variables were extracted from the above mentioned datasets for a period of 5 years. The safety outcomes this research focused on included:

- Total crashes (all types, and all severity levels); and
- Fatal and Injury (FI) crashes

Table 7 provides the descriptive statistics for all (total) crashes, fatal, injury, and property damage only (PDO) crashes on rural two-lane horizontal curves in the states of Utah and Washington.

Roadway and Traffic Data

Crash data were collected every year for 5 years in both the states. Similarly, the traffic volume data (i.e., AADT) were collected for 5 years, and was averaged for 1 year in the datasets. However, for the roadway and geometric features, and cross-section attributes, the data were collected for 1 year and were used throughout the duration of the study, if nothing changes. For Utah, roadway and geometric features, and cross-section attributes were obtained from the LiDAR data, through the UDOT online data portal.

Table 7 Descriptive Statistics for Road Crashes

Variable	Description	Obs.	Mean	Std. Dev	Min	Max
Tot_Crash_UT	Total Crashes in Utah	1710	0.703	1.520	0	18
Fatal_Crash_UT	Fatal Crashes in Utah	1710	0.016	0.131	0	2
Injury_Crash_UT	Injury Crashes in Utah	1710	0.180	0.576	0	8
PDO_Crash_UT	PDO Crashes in Utah	1710	0.506	1.160	0	13
Tot_Crash_WA	Total Crashes in Washington	6605	1.053	2.089	0	36
Fatal_Crash_UT	Fatal Crashes in Washington	6605	0.028	0.176	0	2
Injury_Crash_UT	Injury Crashes in Washington	6605	0.391	0.948	0	18
PDO_Crash_UT	PDO Crashes in Washington	6605	0.633	1.358	0	22

The data obtained from LiDAR for the horizontal curves in Utah were questionable. Hence extra steps were taken to visually check and verify the data associated with the horizontal curves in Google Earth. Some algorithms were used to verify the point of curve (PC) and point of tangent (PT) locations of the horizontal curve in Google Earth. The locations of the curves were identified by the state route number and milepost information. The curve data were linked with the roadway data (number of through lanes, and passing lanes), shoulder width, and traffic volume data obtained from the UDOT online data portal.

In addition to the AADT, the hourly traffic volume data were also used in the analysis. While aggregate measure of AADT was used for safety analysis, hourly volume data were utilized for traffic volume estimation during the ‘day’ and ‘night’ using spatial interpolation techniques. The hourly traffic volume data are obtained from the automatic traffic recorder (ATR) stations throughout the state of Utah for the years 2009-2013. UDOT has 100 ATR stations that provide an hourly count for 24 hours throughout the day and 365 days throughout the year. These data were available in the UDOT’s website in the form of an excel file. The number of lanes and functional classification characteristics were linked to the ATR stations in ArcGIS, using spatial join. This hourly traffic volume data, number of lanes, functional classification, and some of the socio-economic variables (discussed in the later part of this section) were used for the estimation of disaggregate volumes of traffic, i.e., day and night traffic volume estimation for the horizontal curves in Utah. The descriptive statistics for the roadway and geometric features, cross-section attributes, and operational characteristics for Utah are provided in Table 8.

For Washington, roadway and geometry features, and cross-section attributes were obtained from the roadway and curve file in the HSIS database. These attributes are located by state route number, and milepost information. GPS latitude and longitude information were not available for the curve data. The traffic, roadway, and cross-section attributes including posted speed, AADT, truck percent, functional classification, lane width, and shoulder width were linked to the horizontal curves using the route number and milepost information. The detailed vertical curve information, in addition to the information available in the curve file, was obtained during the study from Washington DOT. Some of the curves with higher segment lengths had multiple AADT's associated with them, and higher values of AADT were assigned to the curves in those cases. The descriptive statistics for the roadway and geometric features, cross-section attributes, and operational characteristics for Washington are provided in Table 9.

Roadside Data

For Utah, data for the only roadside feature available from the UDOT online data portal was the barrier data layer, since the UDOT database was actually built for asset management purposes. However, in the context of safety research, elements such as barrier length, offset, and type were considered to be important for analysis. The barrier length was calculated using start and end milepost information given in the barrier file.

Table 8 Descriptive Statistics for the Roadway, and Geometric Variables in Utah

Variable	Description	Obs	Mean	Std. Dev	Min	Max
AADT_UT	Average Annual Daily Traffic (veh/day)	1710	1169.86	1416.68	21.25	9382.50
Log_AADT_UT	Log Average Annual Daily Traffic	1710	6.56	1.01	3.05	9.14
Sh_Wth_UT	Shoulder Width (ft)	1710	2.97	1.83	0	15.50
Seg_Ln_UT	Segment Length (mi)	1710	0.18	0.11	0.05	1.20
Cur_Rds_UT	Curve Radius (ft)	1710	2089.20	1199.07	278	7350
Cur_Ang_UT	Horizontal Curve Angle (Degrees)	1710	34.30	24.89	3.42	176.57
Deg_Cur_UT	Degree of Curve	1710	3.75	2.25	0.78	20.61

Table 9 Descriptive Statistics for the Roadway, and Geometric Variables in Washington

Variable	Description	Obs	Mean	Std. Dev	Min	Max
AADT_WA	Average Annual Daily Traffic (veh/day)	6605	2758.51	2794.11	139	25844
Log_AADT_WA	Log Average Annual Daily Traffic	6605	7.52	0.90	4.93	10.15
Trk_Pct_WA	Truck Percent (%)	6605	17.58	8.93	0	66
Ln_Wth_WA	Lane Width (ft)	6605	11.50	0.84	9	20
Sh_Wth_WA	Shoulder Width (ft)	6605	4.19	2.24	0	15
Sp_Lm_WA	Posted Speed Limit (mph)	6605	53.53	7.29	25	65
Seg_Len_WA	Segment Length (mi)	6605	0.15	0.10	0.05	1.26
Cur_Rds_WA	Curve Radius (ft)	6605	2041.86	1596.30	250	9550
Cur_Ang_WA	Horizontal Curve Angle (Degrees)	6605	29.87	21.33	3.16	171.23
Deg_Cur_WA	Degree of Curve	6605	4.74	3.55	0.60	22.92
Grade_WA	Grade (%)	6605	0.22	2.71	-9.67	9.47
Grade_neg4	Indicator Variable 1=Grade < -4% 0=otherwise	6605	0.08	0.27	0	1
Grade_neg4_0	Indicator Variable 1=-4%<=Grade<0 0=otherwise	6605	0.26	0.43	0	1
Grade_0to4	Indicator Variable 1=0<=Grade<=4% 0=otherwise	6605	0.55	0.49	0	1
Grade_pos4	Indicator Variable 1=Grade>4% 0=otherwise	6605	0.11	0.31	0	1

For Washington, the roadside data were obtained from Roadside Features Inventory Program (RFIP), a system-wide program of limited scope undertaken by Washington DOT. This program was designed to provide information on the number, types, and locations of roadside features for the main purpose of safety analysis. Data were only available for a portion of road segments in Washington because the program was stopped before the roadside data from the entire state was collected. The data on roadside features were available in the form of an excel file and the location information of roadside features was available through the GPS coordinates (i.e., latitude and longitude).

In order to link the roadside data (available through GPS latitude and longitude) with roadway and geometric features, cross-section attributes, operational characteristics, and crash data (available through state route number, and milepost information), a linear referencing system (LRS) route feature class/layer was acquired from Washington DOT, which was a roadway network. Then the route field was identified and all the segments of a particular route were aggregated to result in one row per route number in a shapefile.

This layer was now combined with the roadway and geometric features, resulting in the projection of the rural, two-lane horizontal curves in the state of Washington in ArcGIS. Then the roadway and roadside features were linked using the GPS location information in ArcGIS. The roadside features considered for this analysis were the presence and locations of concrete barrier, guardrail, tree, and fixed object. The descriptive statistics for the roadside features of Utah and Washington are provided in Table 10.

Table 10 Descriptive Statistics for the Roadside Variables

Variable	Description	Obs	Mean	Std. Dev	Min	Max
			Utah			
Rtbr_Len_UT	Right Barrier Length (mile)	1710	0.0007	0.01	0	0.51
Ltbr_Len_UT	Left Barrier Length (mile)	1710	0.0005	0.008	0	0.18
Probr_Cur_UT	Proportion of Total Barrier in Curve (%)	1710	11	28	0	200
			Washington			
Cnbr_Len_WA	Concrete Barrier Length (mi)	72	0.12	0.11	0.01	0.60
Per_Cnbr_WA	Percent of Concrete Barrier in Curve (%)	72	47.44	32.56	3.07	100
Guar_Len_WA	Guardrail Length (mi)	1715	0.14	0.13	0.01	1.68
Per_Guar_WA	Percent of Guardrail in Curve (%)	1715	55.61	36.49	1.02	100
Tree_Cnt_WA	Tree Count	734	2.27	2.04	1	25
Tree_Dia_WA	Tree Average Diameter (ft)	734	5.66	2.69	0.98	12.42
Fix_Cnt_WA	Fixed Object Count	840	2.37	2.86	1	32

Weather Data

For both the states, weather data were obtained from the NOAA National Climatic Data Center. The climate data online search tool was used to obtain the past weather and climate data for both Utah and Washington for their respective study duration of 5 years. Data from the land-based stations included temperature, precipitation, wind, snowfall, relative humidity, and atmospheric pressure. The data were available at different time frame levels, i.e., daily, monthly, and on a yearly basis. However, aggregate yearly data were collected for this research, to be consistent with the time scale used for crash, roadway and geometric features, cross-section attributes, and operational characteristics data. The data elements regarding precipitation, temperature, and snowfall were used in this research. Table 11 provides detailed explanations and descriptive statistics for the weather variables used in the analysis.

Table 11 Descriptive Statistics for the Weather Variables

Variable	Description	Obs	Mean	Std. Dev	Min	Max
			Utah			
Mean_Temp_UT	Mean Temperature (in degrees F)	1710	48.15	5.19	36.51	61.95
Ndays_90_UT	No. of days with 90 plus temperature	1710	35.49	31.57	0	120.6
Ndays_32_UT	No. of days with 32 minus temperature	1710	166.40	39.53	62.80	238.20
Tot_Snw_UT	Total Snowfall (inches)	1710	62.41	58.48	2.58	322.34
NdayPr_1_UT	No. of days 1/10th inch precipitation	1710	38.41	15.06	14.20	87.20
NdayPr_5_UT	No. of days ½ inch precipitation	1710	7.24	4.47	2	32
NdayPr_10_UT	No. of days 1 inch precipitation	1710	1.28	1.18	0	9.80
Wntr_Clo_UT	Winter Closure Indicator 1=closure 0=otherwise	1710	0.10	0.30	0	1
			Washington			
Mean_Temp_WA	Mean Temperature (in degrees F)	2793	47.02	6.05	13.90	55.24
Ndays_90_WA	No. of days with 90 plus temperature	2793	1.94	0.97	1	5
Ndays_32_WA	No. of days with 32 minus temperature	2793	10.18	4.01	3	17
Tot_Snw_WA	Total Snowfall (inches)	2793	3.43	7.89	0.02	63.38
NdayPr_1_WA	No. of days 1/10th inch precipitation	2793	9.14	3.38	3	14
NdayPr_5_WA	No. of days ½ inch precipitation	2793	3.71	2.26	1	7
NdayPr_10_WA	No. of days 1 inch precipitation	2793	1.90	1.10	0	4

The weather data were obtained for all the weather stations in Utah and Washington for 5 years duration. The weather station was linked to the roadway segment based on the distance or proximity of the station to each curve. However, the terrain effects on weather patterns were not taken into account while linking the weather stations to the nearest curve segments. Data imputation was done for those weather stations that have more than 90 percent of the data available. In total, roughly 10 percent or less of the yearly temperature and precipitation data were imputed due to missing data.

Socio-Economic Data

Socio-economic data for the state of Utah were obtained from United States Census Bureau for the year 2010. The data include population counts, household counts, and ethnicity counts organized by different sized geographic areas, where the smallest unit of area is a 'census block' and the biggest is 'counties'. Majority of the studies on estimating the AADT included some kind of socio-economic variables, especially number of people and households in the area, and found them to be significant for the AADT estimation (*82,86*). The socio-economic data were pooled from the 2010 U.S. Census Bureau estimates on a census block level, allowing for a convenient extraction and linkage of the values of these socio-economic variables to the respective ATR station locations. The descriptive statistics for the population and household variables are provided in Table 12.

Table 12 Descriptive Statistics for the Population and Household Variables

Variable	Description	Obs	Mean	Std. Dev	Min	Max
Pop_UT	Population in the census block with ATR	100	46.74	92.37	0	399
Hou_UT	Number of Housing units in census block with ATR	100	22.50	44.56	0	204

CHAPTER 5

STATISTICAL ROAD SAFETY MODELING RESULTS

This chapter presents the results of the statistical road safety models of expected number of crashes on rural, two-lane horizontal curves, obtained using all the methods described in the 'Methodology' chapter using the variables collected and described in the 'Data Collection' chapter. These regression models of expected crash frequency help in determining how this "new" information, obtained from applying each of the methods described in the previous chapters, impacts the model estimation results and interpretations. Preliminary and final model specifications were obtained by using different combinations of explanatory variables in the models. The selected final model specification for each of the methods applied is also presented at the end of each section.

The presented models include the approach-specific goodness-of-fit indicators and the diagnostics that are required to evaluate the statistical road safety modeling results for regression models of expected total crash frequency and fatal plus injury crash frequency. The final models for each of the methods described in the previous chapter are recommended based on the comparison of the goodness-of-fit measures, variable selection, and appropriateness in the model specifications. This chapter also includes the interpretation of results extracted to demonstrate the effects of including new information specific to each of the methods, on regression coefficient estimates of selected right-hand side variables and model prediction results.

The choice of selecting possible explanatory variables to be included in the model specification was guided by the previous research findings and the substantive rationale for the statistical relationship between an explanatory variable and a response variable. Different model specifications were tested when sound reason for adding explanatory variables, or removing them from an equation, was available. At this time, the regression coefficient estimates for each of the explanatory variable included are also checked to see if there are no major changes in the coefficient estimates when new variables are added or removed from the model. The process of obtaining the best model specification was by successively adding or removing variables one at a time, based on its theoretical relevance, t-statistic, and the p-value. The p-value that is usually considered for statistical significance is 0.05. However, in this research, because of including new explanatory variables and the sample size constraints, a p-value of 0.15 is considered for including the explanatory variables in the model specification.

After all the variables are added or removed from the model, the R-squared value or the F-statistic is computed for the complete model to check if any of the variables previously added to the model can now be deleted or vice-versa. This procedure is continued until all of the variables not in the model specification have no significant effect on the response variable (i.e., the expected number of crashes). The final model specification will include all the significant variables that are very relevant to explaining the effects of the response variable. The preliminary model specification has all the possible significant explanatory variables in the model, whereas the final model specification contains the fundamental explanatory variables that are readily available. Hence, the final model specification adequately fits the data, and is interpretable.

Safety Effects of Traffic Pattern Estimates

This part of the research was to incorporate the day and night traffic volume estimates, obtained from kriging interpolation methods, along with other explanatory variables into statistical road safety models of expected crash frequency. Two dependent variables were modeled in this part of the research, total (all) crashes and fatal plus injury (FI) crashes. After predicting the day and night traffic volumes in the entire state of Utah (including all facility types and area types), these predictions were then exported only for a selected number of rural, two-lane horizontal curves that are of interest to this research. The hypothesis is that the sites with the same daily volume and geometric characteristics will differ in their safety performance if they differ in their distributions of day and night traffic volumes. Specific to the context of this research, the horizontal curves with higher proportions of traffic at night are expected to experience more crashes than similar curves with higher proportions of traffic during the day.

Preliminary NB Models for Total and FI Crashes

As mentioned in the previous chapter, a negative binomial (NB) regression model was used with the expected number of total and FI crashes as left-hand side variables, and selected traffic, geometric, and weather characteristics on the right-hand side including the AADT and predicted night-to-day traffic volume ratio from the kriging models. A negative binomial regression model for both the dependent variables with and without the predicted night-to-day traffic volume ratio were run iteratively using different groups of variables provided in the ‘Data’ chapter, as the regression model did not allow the inclusion of all the available variables into a single model. Table 13 and Table 14 show preliminary model specifications, for both the dependent variables (i.e., Total and FI

Table 13 Preliminary NB Regression Models for Total Crashes on Rural Two-Lane Horizontal Curves

Variable List	Model A				Model B			
	Coeff.	Std. Err.	Z Stat	P-Value	Coeff.	Std. Err.	Z Stat	P-Value
Log AADT	0.921	0.043	21.29	0.000	0.917	0.043	21.19	0.000
Predicted Night/Day Volume Ratio	---	---	---	---	3.935	2.594	1.52	0.129
Deg_Curv	0.066	0.016	4.26	0.000	0.066	0.015	4.27	0.000
Ndays32les	0.002	0.001	1.57	0.116	0.002	0.001	1.66	0.097
Ndayspre01	0.006	0.003	2.05	0.041	0.005	0.003	1.95	0.052
Constant	-7.873	0.380	-20.72	0.000	-8.285	0.468	-17.68	0.000
Segment Length	1 (Offset variable)				1 (Offset variable)			
Pseudo R^2	0.1372				0.1378			
Log-Likelihood	-1633.681				-1632.533			
Dispersion Parameter	0.665	0.086			0.662	0.086		

---: Variable not included in the specification

Table 14 Preliminary NB Regression Models for FI Crashes on Rural Two-Lane Horizontal Curves

Variable List	Model C				Model D			
	Coeff.	Std. Err.	Z Stat	P-Value	Coeff.	Std. Err.	Z Stat	P-Value
Log AADT	0.895	0.068	13.04	0.000	0.890	0.068	13.00	0.000
Predicted Night/Day Volume Ratio	---	---	---	---	7.508	4.146	1.81	0.070
Deg_Curv	0.120	0.023	5.14	0.000	0.122	0.023	5.25	0.000
Ndays32les	-0.0002	0.001	-0.14	0.890	0.00009	0..002	0.05	0.963
Ndayspre01	0.005	0.004	1.12	0.265	0.004	0.005	0.89	0.374
Constant	-8.780	0.608	-14.44	0.000	-9.598	0.762	-12.58	0.000
Segment Length	1 (Offset variable)				1 (Offset variable)			
Pseudo R^2	0.1283				0.1301			
Log-Likelihood	-773.251				-771.637			
Dispersion Parameter	0.792	0.208			0.771	0.204		

---: Variable not included in the specification

crashes) with and without the predicted night-to-day traffic volume ratio as a comparison between the results obtained from both models. The same explanatory variables were used in both the models to see the effects of these variables on both total crashes and FI crashes. These preliminary models were further used to obtain final model specifications for estimating each response variable, using goodness-of-fit indicators, variable inclusion, and appropriateness in the model specification.

The modeling process for the preliminary model specification started with inclusion of variables that were considered to be the potential candidates for inclusion into the final model specifications. Many explanatory variables related to the traffic characteristics, horizontal alignment, and weather-related variables were explored to be included in the preliminary model specification. The major difference between both preliminary and final model specifications is that the explanatory variables used in preliminary model specifications includes all the possible combinations of variables that were collected as part of this research. However, the final model specifications consist of the set of variables that were statistically significant at 85 percent or higher confidence interval, widely available in all the databases, and can be obtained for any type of road safety research.

Final NB Regression Model for Total Crashes

Table 15 presents the final model specification for the regression model of the expected number of total crashes, with and without the predicted night-to-day traffic volume ratio variable included in the model specification. The final model specification for the regression model of expected number of total crashes included natural logarithm of AADT, predicted night-to-day traffic volume ratio, and degree of curve as explanatory variables, and natural logarithm of segment length as the offset variable in the model. The positive coefficients that are slightly less than one for the natural logarithm of AADT in both Models E and F in Table 15 suggest that as the traffic volume increases on the roadway, the expected number of crashes also increase, but at a nonlinear rate. The parameter estimate of less than unity is consistent with the previous work on rural, two-lane highways (*166,204*).

Table 15 Final NB Regression Models for Total Crashes on Rural Two-Lane Horizontal Curves

Variable List	Model E				Model F			
	Coeff.	Std. Err.	Z Stat	P-Value	Coeff.	Std. Err.	Z Stat	P-Value
Log AADT	0.941	0.041	22.87	0.000	0.935	0.041	22.57	0.000
Predicted Night/Day Volume Ratio	---	---	---	---	4.060	2.62	1.55	0.122
Deg_Curv	0.073	0.015	4.78	0.000	0.072	0.015	4.76	0.000
Constant	-7.478	0.305	-24.49	0.000	-7.880	0.402	-19.58	0.000
Segment Length	1 (Offset variable)				1 (Offset variable)			
Pseudo R^2	0.1344				0.1351			
Log-Likelihood	-1638.943				-1637.748			
Dispersion Parameter	0.683	0.088			0.680	0.088		

---: Variable not included in the specification

The coefficient for the ratio of predicted night-to-day traffic volume ratio is positive, with an estimated value of 4.060 in Model F. This variable is statistically significant at 87 percent confidence level in this model. The positive coefficient verifies the earlier hypothesis that horizontal curves with higher proportions of traffic at night are expected to experience more crashes than similar horizontal curves with higher proportions of traffic during the day. The positive coefficient for the degree of curve suggests that as the radius increases, the expected number of crashes will decrease. In other words, the estimated model shows that if degree of curve increases by 100 percent, the crashes increase by more than 7 percent, with all the other variables being constant.

Final NB Regression Model for FI Crashes

Table 16 presents the final model specification for regression model of expected number of FI crashes, with and without the predicted night-to-day traffic volume ratio variable included in the model specification. Similar to the regression model of total crashes, all the coefficient estimates were as expected in both Models G and H as well.

Table 16 Final NB Regression Models for FI Crashes on Rural Two-Lane Horizontal Curves

Variable List	Model G				Model H			
	Coeff.	Std. Err.	Z Stat	P-Value	Coeff.	Std. Err.	Z Stat	P-Value
Log AADT	0.920	0.064	14.33	0.000	0.909	0.064	14.17	0.000
Predicted Night/Day Volume Ratio	---	---	---	---	7.890	4.12	1.91	0.056
Deg_Curv	0.125	0.022	5.46	0.000	0.126	0.022	5.55	0.000
Constant	-8.809	0.488	-18.04	0.000	-9.599	0.642	-14.93	0.000
Segment Length	1 (Offset variable)				1 (Offset variable)			
Pseudo R^2	0.1275				0.1290			
Log-Likelihood	-773.917				-772.119			
Dispersion Parameter	0.790	0.208			0.770	0.204		

---: Variable not included in the specification

The final model specification includes the same explanatory variables as the total crash model. As expected, the increases in natural logarithm of AADT and degree of curve were associated with an increase in the expected number of FI crashes at the 99 percent confidence level. The coefficient for the ratio of predicted night-to-day traffic volume ratio is positive, with an estimated value of 7.890 in Model H. This variable is statistically significant at 94 percent confidence level in this model. The positive coefficient verifies the earlier hypothesis on the expected number of night crashes being more than the day crashes, on similar horizontal curves with same traffic volumes.

From the final models, it can be seen that including the predicted night-to-day traffic volume variable seems to provide minor benefit in terms of model fit, when comparing Model F to Model E and Model H to Model G for regression models of expected number of total crashes and FI crashes, respectively. The pseudo R^2 is somewhat improved (larger) and the overdispersion parameter is improved (smaller). As mentioned earlier, it can be seen that regression coefficient estimates for predicted night-to-day traffic volume ratio in Models F and H are positive, which suggests that the expected number of crashes increase with an increase in the traffic volumes at night.

As explained above, this variable is statistically significant at 87 percent confidence level and 94 percent confidence level in the regression models of expected number of total and FI crashes, respectively. The parameter is informative, both in the sign and magnitude, and therefore would be a loss of key information if this is excluded from the model specifications. The parameter estimate provides useful information, and is assumed that statistical significance will be improved as sample size becomes larger.

The last part of this section presents the effects of the variables that represent the daily traffic pattern estimates on the CMFs of total and fatal plus injury crash models, analyzed in this research. Figure 11 and Figure 12 present the relationship between CMF for total and FI crashes, respectively, and horizontal curve radius, at a constant traffic volume of 8000 vehicles per day, with a varying percentage of traffic volume at night. The results presented in these figures were extracted to further interpret the regression coefficient estimates obtained from the final NB regression models of expected crash frequency.

These figures show that, for total crashes, horizontal curves with 10 percent of traffic volume at night have 18 percent fewer crashes, and the curves with 20 and 25 percent of traffic volume at night have 23 and 50 percent more crashes, when compared to curves with 15 percent of traffic volume at night. Similarly, for the FI crashes, horizontal curves with 10 percent of traffic volume at night have 30 percent fewer crashes, and the curves with 20 percent of traffic volume at night have 50 percent more crashes, when compared to curves with 15 percent of traffic volume at night. These findings show that the severe crashes are much higher at night than during the day, which is consistent with the previous literature (*39-40*).

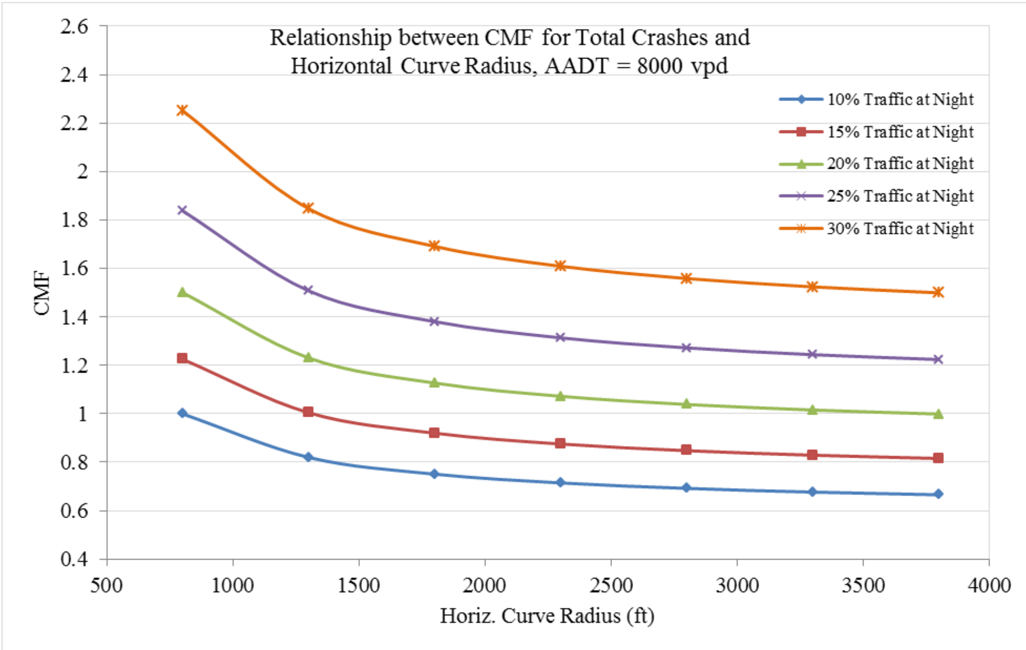


Figure 11 Relationship Between CMF for Total Crashes and Horizontal Curve Radius, at an AADT of 8000 veh/day

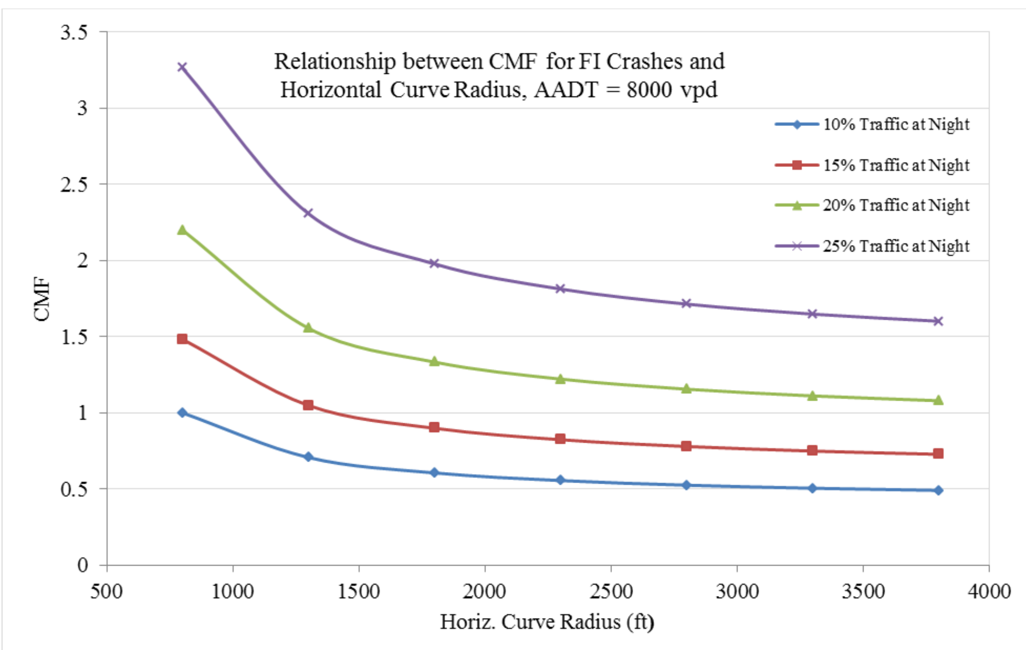


Figure 12 Relationship Between CMF for FI Crashes and Horizontal Curve Radius, at an AADT of 8000 veh/day

Safety Effects of Measurement Error Corrections

This part of the research was focused on evaluating the impacts of measurement error in AADT estimates on regression coefficient estimates of right-hand side variables in regression models of expected crash frequency. The magnitude of measurement error in AADT estimates was measured through measurement error variance values, and different values of measurement error variance were used to estimate the effects of measurement error on study results. These estimations, sometimes referred to as *sensitivity analysis*, were used in this part of the study to assess the robustness of the study findings in general, and quantify the plausible impact these specific errors have on parameter estimates of right-hand side variables in particular.

This section presents the study results after accounting for measurement error in AADT estimates, and how those results affected the overall regression model inferences. As explained in the 'Methodology' section, the measurement error (ME) in AADT estimates is accounted for in the regression models of expected total crash frequency in this research. Hence, the dependent variable that was modeled as part of this study is total crashes, as a function of several explanatory variables. The hypothesis is that the regression coefficient estimates in the naïve model (i.e., model without accounting for measurement error) bias towards the null value. Specific to the context of this research, the regression coefficient estimates of Log AADT bias towards zero, if measurement error is not accounted for in the regression models of expected crash frequency.

NB Models without Accounting for Measurement Error in Traffic Volumes

Table 17 presents the NB regression model specification for expected number of total crashes, as a function of explanatory variables that were statistically significant at 95 percent or higher confidence level. This model specification did not account for the measurement error in the Log AADT estimates. Hence, this regression model of expected total crash frequency is considered to be the naïve model, where the parameter estimates for Log AADT without accounting for measurement error are biased towards zero. The NB model specification for expected number of total crashes included natural logarithm of AADT, natural logarithm of segment length, speed limit, shoulder width, and degree of curve as explanatory (or right-hand side) variables in the model.

The positive coefficient for the natural logarithm of AADT in Table 17 suggests that as the traffic volume increases on the roadway, the expected number of crashes also increase, at a nonlinear rate. Similar to the natural logarithm of AADT, the regression coefficient estimate for natural logarithm of segment length was also positive, which captured the increase in expected crash frequency with an increase in segment length, due to increased exposure. The positive coefficients for speed limit and degree of curve were also associated with an increase in expected number of crashes as speed limit and degree of curve increases.

In other words, the estimated regression model shows that if speed limit is doubled, the crashes increase by more than 1 percent, with all the other variables being constant. Similarly, an increase in the degree of curve by 100 percent increases the crashes by more than 9 percent. The shoulder width variable appeared to be statistically significant with the regression coefficient estimate being negative. This indicates the beneficial influence of shoulder width; an increase in shoulder width is associated with a decrease in the expected crash frequency. This means that as shoulder width is doubled, the crashes decrease by more than 2 percent, with all the other variables being constant.

Table 17 NB Regression Models for Total Crashes on Rural Two-Lane Horizontal Curves with Error-Prone AADT Estimates

Variable List	Coeff.	Std. Err.	Z Stat	P-Value	95% Conf. Interval	
Log AADT	1.016	0.023	43.02	0.000	0.969	1.062
Log Segment Length	0.847	0.032	26.32	0.000	0.784	0.910
Speed Limit	0.013	0.002	5.00	0.000	0.008	0.018
Shoulder Width	-0.021	0.008	-2.49	0.013	-0.038	-0.004
Deg_Curv	0.095	0.005	16.95	0.000	0.084	0.106
Constant	-7.354	0.263	-27.94	0.000	-7.870	-6.838
Pseudo R-squared	0.1422					
Log-Likelihood	-7889.895					
Dispersion Parameter	0.659	0.032			0.598	0.727

NB Models Accounting for Measurement Error in Traffic Volumes

The results shown in this section reflect the effects of accounting for classical measurement error in Log AADT estimates in regression models of expected crash frequency. All of the tables in this section show the results of regression calibration and simulation extrapolation analyses accounting for the classical measurement error in Log AADT estimates, where the measurement error variance values ranged from 0.05 to 0.20, with increments of 0.05. Table 18 shows the regression coefficient estimates of the selected explanatory variables, along with the bootstrap standard errors and 95 percent confidence intervals while the measurement error variance in Log AADT estimates was 0.05. Figure 13 shows the naïve parameter estimate and SIMEX parameter estimate for Log AADT, obtained by quadratic extrapolation in SIMEX method. The parameter estimates for Log AADT obtained from these two approaches were greater than those obtained from naïve analysis. In other words, the parameter estimates for Log AADT obtained from RCAL and SIMEX methods were biased by 8.75 percent and 9.05 percent, respectively, from the naïve regression model parameter estimates (i.e., model without accounting for measurement error).

Table 18 RCAL and SIMEX Analysis Results for Total Crashes on Rural Two-Lane Horizontal Curves with ME Variance of 0.05 in Log AADT Estimates

Variable	Regression Calibration (RCAL)*				Simulation Extrapolation (SIMEX)*			
	Coeff.	*Bootstr Std. Err.*	*95% Conf. Interval*		*Coeff.*	*Bootstr Std. Err.*	*95% Conf. Interval*	
Log AADT	1.105	0.029	1.047	1.162	1.108	0.030	1.050	1.166
Log Seg. Len.	0.853	0.032	0.790	0.916	0.858	0.033	0.794	0.922
Speed Limit	0.014	0.003	0.008	0.020	0.015	0.003	0.009	0.020
Shoulder Width	-0.035	0.011	-0.056	-0.013	-0.032	0.010	-0.052	-0.011
Deg_Curv	0.098	0.006	0.086	0.111	0.100	0.006	0.088	0.113
Constant	-8.054	0.297	-8.637	-7.470	-8.155	0.302	-8.748	-7.563
Wald F-Statistic	562.06				574.71			
P-Value	0.00				0.00			

*All regression coefficient estimates are significant at 95% or more confidence level

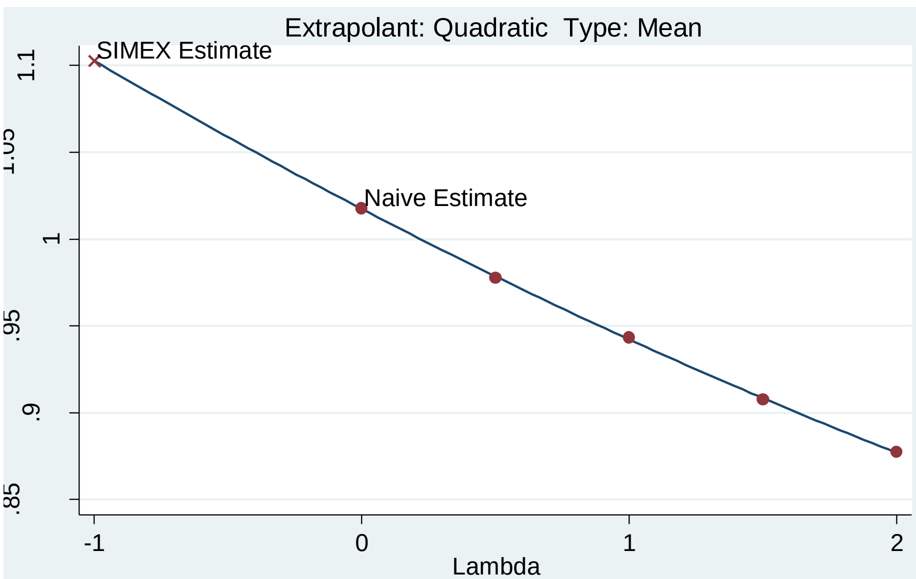


Figure 13 SIMEX Naïve and Quadratic Extrapolation Estimates Plot for Log AADT with ME Variance of 0.05

Similar to Log AADT, the parameter estimates for natural logarithm of segment length were also greater than the ones obtained from naïve method. The parameter estimates for posted speed limit and degree of curve obtained from the two approaches were also greater than the ones obtained from naïve method. In other words, the regression coefficient estimates for posted speed limit were biased by 8 percent and 15 percent, respectively, for RCAL and SIMEX methods. The regression coefficient estimates for shoulder width were lower and biased by 60 percent and 50 percent, respectively, for RCAL and SIMEX methods. All these regression coefficient estimates obtained from RCAL and SIMEX were significant at 99 percent or more confidence level, when measurement error variance values of 0.05 in Log AADT estimates were accounted for in the regression models of expected number of crashes.

Table 19 shows the regression coefficient estimates of the selected explanatory variables, along with the bootstrap standard errors and 95 percent confidence intervals while the measurement error variance in Log AADT estimates was 0.10. Similar to Figure 13, Figure 14 also shows the naïve parameter estimate and SIMEX parameter estimate for Log AADT, obtained by quadratic extrapolation in SIMEX method, but for ME variance of 0.10.

As expected, the parameter estimates for Log AADT obtained from these two approaches were greater than those obtained from naïve analysis. In other words, the parameter estimates for Log AADT obtained from RCAL and SIMEX methods were biased by 18 percent and 17 percent, respectively, from the naïve regression model parameter estimates (i.e., model without accounting for measurement error).

Similar to the previous results for ME variance of 0.05 in Log AADT estimates, the parameter estimates for natural logarithm of segment length from RCAL and SIMEX approaches were also greater than the ones obtained from naïve method. The parameter estimates for posted speed limit and degree of curve obtained from the two approaches were also greater than the ones obtained from naïve regression model. In other words, the regression coefficient estimates for posted speed limit were biased by 23 percent and 38 percent, respectively, for RCAL and SIMEX methods.

Table 19 RCAL and SIMEX Analysis Results for Total Crashes on Rural Two-Lane Horizontal Curves with ME Variance of 0.10 in Log AADT Estimates

Variable	Regression Calibration (RCAL)*				Simulation Extrapolation (SIMEX)*			
	Coeff.	*Bootstr Std. Err.*	*95% Conf. Interval*		*Coeff.*	*Bootstr Std. Err.*	*95% Conf. Interval*	
Log AADT	1.208	0.030	1.149	1.267	1.192	0.032	1.130	1.255
Log Seg. Len.	0.856	0.034	0.788	0.924	0.856	0.033	0.790	0.922
Speed Limit	0.016	0.003	0.010	0.022	0.018	0.003	0.012	0.024
Shoulder Width	-0.054	0.010	-0.074	-0.033	-0.044	0.011	-0.065	-0.023
Deg_Curv	0.101	0.006	0.089	0.113	0.103	0.006	0.090	0.115
Constant	-8.858	0.302	-9.451	-8.266	-8.948	0.327	-9.591	-8.306
Wald F-Statistic	582.21				558.56			
P-Value	0.00				0.00			

*All regression coefficient estimates are significant at 95% or more confidence level

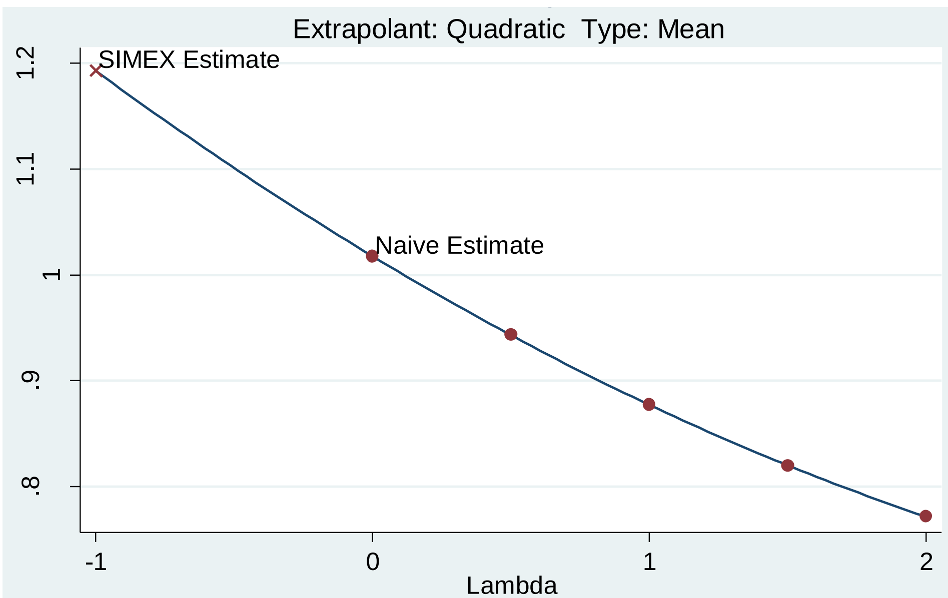


Figure 14 SIMEX Naïve and Quadratic Extrapolation Estimates Plot for Log AADT with ME Variance of 0.10

Similarly, the regression coefficient estimates for shoulder width were lower and biased by 150 percent and 100 percent, respectively, for RCAL and SIMEX methods. All these regression coefficient estimates obtained from RCAL and SIMEX were significant at 99 percent or more confidence level, when measurement error variance values of 0.10 in Log AADT estimates were accounted for in the regression models of expected crash frequency.

Table 20 shows the regression coefficient estimates of the selected explanatory variables, along with the bootstrap standard errors and 95 percent confidence intervals while the measurement error variance in Log AADT estimates was 0.15. Similar to the above figures, Figure 15 also shows the naïve parameter estimate and SIMEX parameter estimate for Log AADT, obtained by quadratic extrapolation in SIMEX method, but for measurement error variance of 0.15. As expected, the parameter estimates for Log AADT obtained from RCAL and SIMEX methods were greater and biased by 31 percent and 25 percent, respectively, from the naïve regression model parameter estimates (i.e., model without accounting for measurement error).

Similar to the previous results for ME variance of 0.10, the parameter estimates for natural logarithm of segment length from RCAL and SIMEX approaches were also greater than the ones obtained from naïve method. The parameter estimates for posted speed limit and degree of curve obtained from these two approaches were also greater than the ones obtained from naïve method, i.e., the regression coefficient estimates for posted speed limit were biased by 38 percent and 46 percent, respectively, for RCAL and SIMEX methods. The regression coefficient estimates for shoulder width were lower and biased by more than 200 percent and 160 percent, respectively, for RCAL and SIMEX analysis. All these regression coefficient estimates obtained from both approaches were significant at 99 percent or more confidence level, for measurement error variance values of 0.15 in Log AADT estimates in regression models of expected number of crashes. It was observed that as the magnitude of measurement error variance increases in Log AADT estimates, the parameter estimates tend to move away from zero, which agrees with the previous literature on measurement error models.

Table 20 RCAL and SIMEX Analysis Results for Total Crashes on Rural Two-Lane Horizontal Curves with ME Variance of 0.15 in Log AADT Estimates

Variable	Regression Calibration (RCAL)*				Simulation Extrapolation (SIMEX)*			
	Coeff.	*Bootstr Std. Err.*	*95% Conf. Interval*		*Coeff.*	*Bootstr Std. Err.*	*95% Conf. Interval*	
Log AADT	1.334	0.037	1.261	1.406	1.276	0.034	1.208	1.344
Log Seg. Len.	0.859	0.034	0.793	0.926	0.864	0.036	0.793	0.934
Speed Limit	0.018	0.003	0.012	0.024	0.019	0.003	0.013	0.026
Shoulder Width	-0.077	0.012	-0.100	-0.054	-0.056	0.011	-0.078	-0.033
Deg_Curv	0.105	0.006	0.093	0.117	0.107	0.006	0.094	0.119
Constant	-9.830	0.337	-10.491	-9.168	-9.649	0.342	-10.321	-8.977
Wald F-Statistic	520.29				504.72			
P-Value	0.00				0.00			

*All regression coefficient estimates are significant at 95% or more confidence level

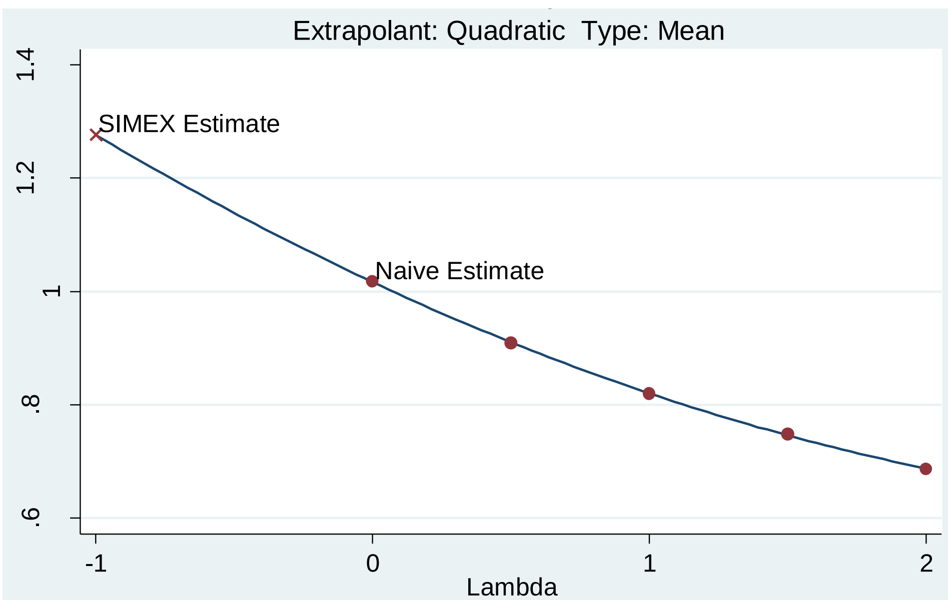


Figure 15 SIMEX Naïve and Quadratic Extrapolation Estimates Plot for Log AADT with ME Variance of 0.15

Table 21 shows the regression coefficient estimates of the selected explanatory variables, along with the bootstrap standard errors and 95 percent confidence intervals while the measurement error variance in Log AADT estimates was 0.20. Figure 16 shows the naïve parameter estimate and SIMEX parameter estimate for Log AADT, obtained by quadratic extrapolation in SIMEX method, but for measurement error variance of 0.20. As expected, the parameter estimates for Log AADT obtained from these two approaches

Table 21 RCAL and SIMEX Analysis Results for Total Crashes on Rural Two-Lane Horizontal Curves with ME Variance of 0.20 in Log AADT Estimates

Variable	Regression Calibration (RCAL)*				Simulation Extrapolation (SIMEX)*			
	Coeff.	*Bootstr Std. Err.*	*95% Conf. Interval*		*Coeff.*	*Bootstr Std. Err.*	*95% Conf. Interval*	
Log AADT	1.488	0.041	1.406	1.570	1.356	0.037	1.282	1.429
Log Seg. Len.	0.864	0.034	0.796	0.931	0.864	0.033	0.798	0.930
Speed Limit	0.021	0.003	0.015	0.027	0.021	0.003	0.015	0.028
Shoulder Width	-0.106	0.012	-0.131	-0.081	-0.068	0.012	-0.091	-0.045
Deg_Curv	0.109	0.006	0.097	0.121	0.109	0.006	0.096	0.122
Constant	-11.026	0.379	-11.770	-10.282	-10.322	0.374	-11.056	-9.587
Wald F-Statistic	512.83				463.56			
P-Value	0.00				0.00			

*All regression coefficient estimates are significant at 95% or more confidence level

were greater than those obtained from naïve analysis. In other words, the parameter estimates for Log AADT obtained from RCAL and SIMEX methods were biased by 46 percent and 33 percent, respectively, from the naïve regression model parameter estimates (i.e., model without accounting for measurement error).

Similar to the previous results for different values of measurement error variance; the parameter estimates for natural logarithm of segment length from RCAL and SIMEX approaches were also greater than the ones obtained from naïve method. The parameter estimates for posted speed limit and degree of curve obtained from these two approaches were also greater than the ones obtained from naïve method. In other words, the regression coefficient estimates for posted speed limit were biased by 60 percent for both RCAL and SIMEX methods.

The regression coefficient estimates for shoulder width were lower and biased by 400 percent and 200 percent, respectively, for RCAL and SIMEX approaches. All these regression coefficient estimates obtained from RCAL and SIMEX were significant at 99 percent or more confidence level, when measurement error variance values of 0.20 in Log AADT estimates were accounted for in the safety models.

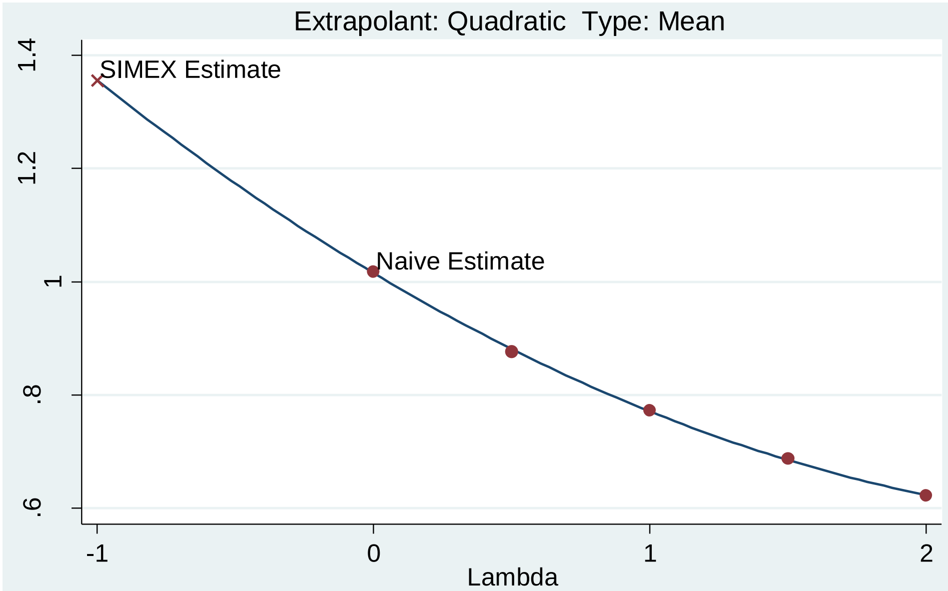


Figure 16 SIMEX Naïve and Quadratic Extrapolation Estimates Plot for Log AADT with ME Variance of 0.20

Overall, it can be said that the regression coefficient estimates obtained by both RCAL and SIMEX measurement error correction approaches were consistent in direction with the ones obtained from naïve NB regression models of expected crash frequency. It was also noted that the parameter estimates with a positive coefficient were larger and a negative coefficient were smaller, when measurement error was accounted for in the regression models when compared to the ones obtained from naïve regression model. This implies that the parameter estimates for the selected explanatory variables with measurement error were biased towards zero, when the error was not accounted for in the regression modeling of expected number of crashes. This is consistent with the previous work on measurement errors, which suggests that measurement error shrinks the parameter estimates towards zero (*41,117*). However, as the measurement error in Log AADT estimates was accounted for in the models, the parameter estimates started biasing away from zero.

The bootstrap standard errors corresponding to the RCAL and SIMEX approaches for the selected explanatory variables were larger when compared to the standard error estimates in naïve model estimation. This shows that the NB regression model without accounting for measurement errors in explanatory variables (i.e., Log AADT estimates in this context) underestimates the standard errors for the parameter estimates, falsely increasing the parameter significance in the model specification. This finding is consistent with the previous work, which states that the standard errors obtained from the naïve NB regression models are smaller when compared to the models that account for the measurement errors in explanatory variables (*133*). The parameter estimates were found to be biased for the error-free explanatory variables, along with the error-prone explanatory variables after the measurement error correction approaches were applied. Of all the correctly measured explanatory variables, the effect of measurement error in natural logarithm of AADT was greater in shoulder width, followed by posted speed limit, degree of curve, and natural logarithm of segment length.

These differences in the effects were found to be due to the correlation aspects between the explanatory variables. The correlation effect between Log AADT and shoulder width is higher, explaining the reason for larger effect than the other variables. However, the effects for all the explanatory variables increased with an increase in the ME variance in Log AADT estimates. Additionally, if the results are compared between RCAL and SIMEX approaches, SIMEX tend to preserve the precision of the regression coefficient estimates, and RCAL inclined to correct for a larger amount of effect bias in regression coefficient estimates, particularly for the error-prone variable, i.e., Log AADT estimates in this research.

Safety Effects of Including Prior Information

This part of the research employed a Bayesian methodological framework to incorporate accumulated knowledge (i.e., prior information) from the past research along with the data in the regression models of expected crash frequency. Three different types of priors were used in this part of the analysis: noninformative, semi-informative, and informative priors for the selected explanatory variables, intercept, and inverse dispersion parameter. These explanatory variables included: natural logarithm of AADT, natural logarithm of segment length, shoulder width, lane width, and degree of curvature. The reason for employing different types of priors was to compare and contrast between the results obtained from different types of priors. After employing the MCMC algorithm, the posterior means and standard deviations were calculated for the selected explanatory variables, along with the intercept term and the inverse dispersion parameter for all the different types of priors in this study.

Bayesian Analysis Results with Noninformative Priors

The results shown in this section reflect the effects of incorporating noninformative priors for the selected explanatory variables, intercept, and inverse dispersion parameter in the Bayesian analysis framework. The table presented in this section shows the posterior means and standard deviations calculated for all the variables, when a large variance was incorporated in the prior distributions. Table 22 shows the posterior means and other estimates of all the parameters, along with the 2.5 percent and 97.5 percent value of the parameter estimates in its posterior distribution. A quick comparison of the posterior means of the parameter estimates with the corresponding maximum likelihood estimates in Table 17 shows that the values are very close. This is to

Table 22 Posterior Means and Other Estimates for All Parameters Using Noninformative Priors

Variable List	Mean	Std. Dev.	MC Err.	2.5%	Median	97.5%
Log AADT	0.996	0.015	0.001	0.959	0.997	1.020
Log Segment Length	0.863	0.030	0.002	0.80	0.863	0.917
Shoulder Width	-0.011	0.007	6.3E-4	-0.026	-0.011	0.004
Degree of Curvature	0.089	0.005	4.2E-4	0.079	0.089	0.101
Lane Width	0.005	0.013	0.001	-0.028	0.004	0.028
Intercept Term	-6.545	0.168	0.018	-6.842	-6.567	-6.254
Inverse Dispersion Parameter	1.493	0.078	0.004	1.345	1.49	1.652
DIC	14669					

be expected because when noninformative priors are employed, less weight is given to the prior means and distributions, and more weight is given to the data in the estimation of the posterior distributions of the parameters.

The results from Table 22 show that Bayesian analysis with noninformative priors resulted in the posterior mean values as expected in terms of magnitude and direction for all the explanatory variables. The analysis resulted in a posterior mean value of 0.996 for natural logarithm of AADT and a posterior standard deviation of 0.015. Similarly, the posterior mean and standard deviation values for natural logarithm of segment length were found to be 0.863 and 0.030, respectively. The posterior estimates for shoulder width and degree of curve were very close to the maximum likelihood estimates obtained from the frequentist method, with the mean values being -0.011 and 0.089, respectively. Similarly, the standard deviation values for shoulder width and degree of curve were 0.007 and 0.005, respectively. The posterior mean value for lane width was positive and close to zero, with the standard deviation being 0.013. The distribution of lane width included zero, and the 2.5 percent to 97.5 percent values ranged from -0.028 to 0.028, respectively.

Bayesian Analysis Results with Semi-Informative Priors

The results shown in this section reflect the effects of incorporating semi-informative priors for the selected explanatory variables, intercept term, and inverse dispersion parameter in the Bayesian analysis framework. The table presented in this section shows the posterior means and standard deviations calculated for all the variables, when a relatively smaller variance (compared to that of the noninformative priors) of a magnitude of 100 was incorporated in the prior distributions. Table 23 shows the posterior means and other estimates of all the parameters, when semi-informative priors were employed.

The results from Table 23 show that Bayesian analysis with semi-informative priors resulted in a posterior mean of 0.985 for natural logarithm of AADT and a posterior standard deviation of 0.018. Similarly, the posterior mean and standard deviation for natural logarithm of segment length was found to be 0.880 and 0.032, respectively. The posterior estimates for shoulder width and degree of curve moved closer to the weighted average value of the priors, the mean values being -0.010 and 0.090, and standard deviation values being 0.007 and 0.006, respectively.

Table 23 Posterior Means and Other Estimates for All Parameters Using Semi-Informative Priors

Variable List	Mean	Std. Dev.	MC Err.	2.5%	Median	97.5%
Log AADT	0.985	0.018	0.002	0.951	0.981	1.021
Log Segment Length	0.880	0.032	0.003	0.810	0.884	0.933
Shoulder Width	-0.010	0.007	7.1E-4	-0.026	-0.009	0.005
Degree of Curvature	0.090	0.006	4.9E-4	0.079	0.091	0.102
Lane Width	0.002	0.019	0.002	-0.03	0.002	0.036
Intercept Term	-6.402	0.157	0.017	-6.675	-6.384	-6.106
Inverse Dispersion Parameter	1.492	0.075	0.003	1.358	1.489	1.653
DIC	14665					

The posterior mean value for lane width moved much closer to zero, with the standard deviation being 0.019. The mean value of the intercept term was -6.402, and standard deviation was 0.157. Overall, all of the posterior estimates obtained from using the semi-informative priors were closer in magnitude to the posterior estimates obtained from using noninformative priors. The only difference is that the posterior mean values using semi-informative priors moved closer to the weighted mean values of the priors (which were used to develop the prior distributions), which is in accordance with the definition of the semi-informative priors. The analysis with the semi-informative priors gives more weightage to the priors than the noninformative priors.

Bayesian Analysis Results with Informative Priors

The results shown in this section reflect the effects of incorporating informative priors for the explanatory variables, intercept term, and inverse dispersion parameter in the Bayesian analysis framework. The table presented in this section shows the posterior means and standard deviations calculated for all the variables, when a weighted variance, calculated from the prior knowledge, was incorporated in the prior distributions. Table 24 shows the posterior means and other estimates of all the parameters, along with the 2.5 percent and 97.5 percent values of the parameter estimates in their posterior distribution.

In these types of priors, more weight is given to the prior means and distributions, and lesser weight is given to the data in the estimation of the posterior distributions of the parameter estimates. These priors were helpful to see if the estimation results from the past studies could be incorporated into the current studies, instead of starting over with each study, using the Bayesian methodological framework. The overall effect of including informative priors on parameter estimates is described in this section.

Table 24 Posterior Means and Other Estimates for All Parameters Using Informative Priors

Variable List	Mean	Std. Dev.	MC Err.	2.5%	Median	97.5%
Log AADT	0.966	0.016	0.001	0.938	0.964	0.997
Log Segment Length	0.848	0.022	0.002	0.801	0.848	0.890
Shoulder Width	-0.028	0.006	5.04E-4	-0.041	-0.027	-0.016
Degree of Curvature	0.069	0.004	2.71E-4	0.061	0.069	0.076
Lane Width	-0.009	0.014	0.001	-0.03	-0.008	0.017
Intercept Term	-5.983	0.165	0.018	-6.357	-5.972	-5.733
Inverse Dispersion Parameter	1.721	0.085	0.003	1.562	1.720	1.896
DIC	14649					

The results from Table 24 show that Bayesian analysis using informative priors resulted in a posterior mean of 0.966 for natural logarithm of AADT and a posterior standard deviation of 0.016. This value of the posterior mean is closer to the prior weighted mean value of 0.799 when compared among the results from all different types of priors incorporated in this analysis. Similarly, the posterior mean and standard deviation for natural logarithm of segment length was found to be 0.848 and 0.022, respectively. The posterior estimates for shoulder width and degree of curve moved much closer to the weighted average value of the priors, the mean values being -0.028 and 0.069, respectively. The posterior mean for lane width was negative with a value of -0.009, also closer to zero, and in the same direction as the lane width prior used for the analysis. This posterior mean value of lane width parameter estimate was different (in direction) from those obtained when noninformative and semi-informative priors were used in the analysis. This means that the data resulted in a positive posterior mean value close to zero for lane width parameter. However, the analysis using informative prior gave more weightage to the prior resulting in a negative posterior mean value for lane width estimate.

The mean value of the intercept term obtained using informative prior was -5.983, and standard deviation was 0.165. Overall, the results obtained from employing different types of priors by changing the prior mean and variance values improved the Bayesian result estimation marginally. In other words, the posterior estimates found by using the informative priors are slightly different from those obtained by employing objective priors. This smaller difference in the parameter estimates is because of the fact that the data and prior distributions agreed on the parameter estimates. However, the Bayesian estimator with the smaller prior variance values (i.e., informative priors) seemed to perform better than the others because of the reduced standard errors of the posterior estimates and smaller confidence intervals for all the parameters in the model. This reflects the reduced uncertainty due to incorporating the information from the previous rigorous and well-defined observational studies. Hence, if the informative priors employed in the Bayesian analysis are accurate, and informative, many reliable and repeatable results can be obtained by employing them using the Bayesian methodological framework.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

This chapter presents the summary of the research findings for each of the three research questions addressed in this research, and describes the major research contributions and limitations for statistical road safety modelers, including researchers and practitioners. This chapter also provides recommendations for the future development of the research. The main objective of this research was to explore key limitations in both data and regression modeling approaches used in observational road safety studies and identify possible solutions. This dissertation develops methods and approaches that address these limitations and develop more informed and complete model specifications using detailed datasets and empirically-derived theory. Data from the states of Utah and Washington for rural, two-lane horizontal curves were used to develop a detailed dataset consisting of crash data, roadway, traffic and roadside data, weather, and socio-economic data, which influence the prediction and estimation of expected number of crashes.

Conclusions and Contributions

Observational road safety studies serve as a major source of knowledge for researchers and other decision makers on the expected road safety effects of highway and traffic engineering decisions. These decisions are further used in the policy development to identify the most important road safety problems, and the contributing factors associated with those problems. This is generally done by regression models of expected crash frequency, which are used to estimate the expected number of crashes, and in some cases, used to estimate the effects of right-hand side variables on expected number of crashes. Hence, it is extremely important to address data and modeling challenges that exist and have a considerable impact on the accuracy and repeatability of observational study results. The idea behind methodological approaches developed in this research was to explore some options to address the above-mentioned data and modeling approach challenges that will allow researchers and practitioners to accurately evaluate the expected road safety effects.

The study on incorporating new information on traffic pattern estimates in the regression models of expected crash frequency introduced the application of geo-spatial interpolation methods in road safety research. This supported the use of spatial factors such as population density as well as other socio-economic factors that have strong influences on the traffic patterns and ultimately crash occurrence in the study area. The inclusion of night-to-day traffic volume ratio in regression models of expected crash frequency provided additional knowledge on how the expected number of crashes change as a function of traffic patterns (shown in the results chapter). The inclusion of this new information provided useful insights into the safety performance, even with only a small impact on model prediction in this particular research.

All observational before-after studies use crash and traffic volume data for time period before and after the improvement of treated sites. With the new information available on the traffic patterns during the entire day, the changes in traffic volume can be analyzed throughout the day and useful conclusions can be drawn in these studies. This will be particularly useful for the 'prediction task' in before-after studies, where the prediction of the safety effect of the treated entities in the after period without the treatment being applied is done. So, instead of using the yearly indicators for the "unmeasured" independent variables (i.e., traffic volumes during the day and night) that change with time, the actual changes in the traffic volumes throughout the day (i.e., day and night) can be used in the prediction task to predict the expected number of crashes. Hence, including this additional information (as an additional variable in cross-sectional studies and instead of the yearly indicators in before-after studies) will help in improving the predictions in the expected number of crashes in observational studies. Additionally, relating the expected number of crashes to day and night traffic volumes will also assist in estimating more reliable crash modification factors (CMFs) in both before-after and cross-sectional studies.

Measurement error in traffic volume estimates is often ignored in road safety research. This research presented a study on accounting for measurement error in the independent variables (i.e., traffic volume estimates) in regression models of expected crash frequency, by introducing functional-type measurement error approaches. Currently, the regression models of expected crash frequency use error-prone AADT estimates, which introduce substantial bias in the regression coefficient estimates of all the independent variables included in the model specification. This bias was eliminated when RCAL and SIMEX methods were used in safety models, incorporating the measurement error variance corrections associated with the AADT estimates. The measurement error correction methods presented in this research assume only one error-prone explanatory variable (i.e., AADT).

The work presented in this part of the research is applicable to both observational before-after and cross-sectional studies in the context of road safety research. The AADT estimates for the study duration are often reported with error. Inaccurate reporting of AADT could lead to bias in the regression coefficient estimates of all the explanatory variables included in the safety models. As explained in the literature chapter, error-prone AADT estimates lead to the parameter estimates' being close to zero, i.e., underreporting of the safety effect occurs. Accounting for the measurement error in AADT estimates provides more accurate estimates of the safety effects for all the independent variables included in the model specification. This is particularly useful for the cross-sectional studies where regression models are used to estimate the effects of right-hand side variables on expected number of crashes, and a slight bias in the regression coefficient estimates can alter the results in this case.

The study on incorporating prior knowledge in the form of informative priors in regression models of expected crash frequency introduced the application of the Bayesian methodological framework in road safety research. This study supported the use of parameter estimates of selected explanatory variables from previous rigorous and well-defined observational studies that have an effect on the expected crash frequency. The inclusion of informative priors in this study provided a logical starting point for utilizing the previous study results, and ultimately converging on a similar model form and specification. The importance of incorporating an informative prior was demonstrated by comparing the results to the cases when more objective priors (i.e., noninformative and semi-informative priors) were used. This study provided useful insights on how informative priors can be incorporated into safety models, which could ultimately improve the accuracy and repeatability of observational study results, without having to start over with each study and rely on the data to obtain the parameter estimates.

The work presented in this part of the research is applicable to both before-after and cross-sectional studies in the context of road safety research. Each study typically 'starts over' and chooses a model functional form that fits the existing data best, without utilizing previous study results. This study provides a starting point by incorporating previous knowledge, and arriving at repeatable study results that could transition to a more formal development of road safety theory. This would be particularly useful in the cases of limited research funding for effective utilization of resources, and establishing a foundation for future road safety theory development.

Limitations and Recommendations

One of the limitations for the study on incorporating new information on traffic pattern estimates in regression models of expected number of crashes was estimating a combined regression model for both day and night crashes, i.e., as total crashes. In the current research, the alternative cases (i.e., separate models for day crashes and night crashes) were tested and the parameter estimates were unstable as specifications changed and were not statistically significant at high level of confidence. This might be because of the nature of the data, and particularly small sample sizes on rural, two-lane horizontal curves. As part of the potential future research efforts, the methodology should be tested in a state with a more dense set of permanent traffic counters. This is hoped to help develop separate and better performing safety models for day and night crashes as a function of day and night traffic volumes, respectively. Regression models accounting for different crash severity levels should also be developed, for day and night crash models.

Other limitations for this study include considering Euclidean distances for kriging model development, and not having enough validation data to further evaluate the kriging predictions. As part of the future research efforts, network distances should be considered and more information on hourly traffic volumes other than at permanent traffic counters need to be obtained and incorporated in the research. This will help in utilizing additional data collected over shorter period of time (e.g., 2-3 days) at locations where the permanent traffic recorder stations are not present for further validation of the kriging predictions. This study included covariates that have an effect on the average annual daily traffic in the study area. Inclusion and removal of covariates changes the traffic volume estimates, as shown earlier by the different model specifications with different sets of covariates in the semivariogram models. As part of potential future research efforts, more covariates that have an effect on the traffic volume estimation and that change with the season and study area should be included in the semivariogram model estimation and kriging interpolation analysis.

Another limitation for this study is the small improvement in the model diagnostics, i.e., Pseudo R-squared and dispersion parameter, when the ratio of night-to-day traffic volume variable was introduced in the model specification. The other coefficient estimates did not change much and remained stable when this additional explanatory variable was introduced in the model. The CMF figures, which are based on the model specification, showing the relationship between the total crashes (or FI crashes) and horizontal curve radius at a specific AADT value for different ratios of night-to-day traffic volumes showed the actual difference in the estimated safety performance. These differences were small in this particular case study application, but the statistical significance of the parameter estimate for this new predictor variable would likely continue to increase as the sample size becomes larger. Hence, as part of the future research efforts, these day and night traffic volume estimates should be developed for all site and area types, which when included in safety models, are hoped to explicitly show greater differences in the safety performance during the day and night.

The major limitation for the study on accounting for measurement error in the independent variables in regression models of expected crash frequency was the calculation of the measurement error variance in AADT estimates. Typically in measurement error problems, external validation data with measurements of both true values of explanatory variable and the mismeasured values are used to estimate the measurement error variance. However, this study utilized the coefficient of variance equation for AADT from an unpublished study in the literature and sensitivity analysis by considering a range of values for the percentage of variance for measurement error. This was done because a 'gold standard' value for AADT was not available.

Additionally, collecting more data for the accurate AADT estimates for the entire study period (i.e., 5 years) in this research was not feasible. Hence, the measurement error variance was calculated using the equations that were available along with some assumptions. This was done to show the potential application of the measurement error correction approaches to regression models of expected crash frequency to correct for measurement error. As part of potential future research efforts, to advance the understanding of the impacts of measurement error on parameter estimates of explanatory variables and enable straightforward implementation of measurement error correction methods in the context of road safety, the calculation of the true measurement error variance (applicable area-wise or county-wise) in AADT estimates should be done. This will further help in obtaining reliable results and making meaningful interpretations from those results. If the calculation of true measurement error variance cannot be done, then methods that utilize partial information to arrive at near true value of measurement error variance need to be explored.

Another limitation for this study was not validating the measurement error models using any traffic volume data and performing sensitivity analysis using the ME variance equation from the literature. The number of permanent traffic recorder stations in Washington was not enough to carry out the validation of the ME correction methods. Future research efforts should focus on including all sites and area types for analysis, so all the available ATR stations, along with other short-term counts available, can be used for validating the models. As more traffic volume data become available, it is hoped that the future research efforts will be focused on deriving a near ‘gold standard’ value for AADT estimates, which will help in the calculation of ME variance, providing exact estimates of the safety effects of explanatory variables with respect to the response variable when RCAL and SIMEX methods are employed.

Other limitations for this study include strong assumptions of functional-type measurement error correction approaches applied in this research. Future research efforts to correct for measurement error should definitely make use of the structural equation modeling or other methods that are flexible and involve fewer assumptions. Consideration of structural-type measurement error approaches, which consider the error-prone explanatory variables to be random, can also address the true measurement error variance value limitation and be useful for making reliable interpretations.

The major limitation for the study on incorporating prior knowledge in Bayesian framework was the development of informative priors using the regression coefficient estimates from previous studies. This study utilized the CMF clearing house star rating system to weight the studies based on the methodology, and statistical rigor. Future research efforts should focus on systematically evaluating the various elements or categories that contribute to the weightage of rigorous and well-defined studies. These studies should explicitly evaluate selected multiple categories using multiple-criteria decision analysis and weight them using a structured technique like analytic hierarchy process. The resulting weights should be used in calculating the weighted mean and variance for prior distributions of explanatory variables, intercept term, and dispersion parameter.

Another limitation for this study includes the prior distribution assumptions of explanatory variables, intercept, and the inverse dispersion parameter. These are assumed to be normally distributed and gamma distributed, respectively. Since there is a slight improvement in the Bayesian estimation as the prior variance changes (or becomes smaller), different prior distributions should be looked into to see the sensitivity of the results to the type of the prior distributions. In other words, more research is needed to determine the behavior of the parameter estimates of explanatory variables and intercept term. Potential research efforts should focus on more rigorous processing of prior information and computing posterior distributions under a range of uncertainty levels, to account for the assumptions on prior distributions.

Other limitations for this study include the selection of the explanatory variables for this research. The link between lane width and safety has not been fully explored in the literature yet. Most researchers indicate that the safety benefit of widening a lane bottoms out somewhere between 11ft and 12ft. Future research efforts should focus on extending the Bayesian analysis by incorporating lane width as a categorical variable (e.g., 10ft, 11ft, and 12ft, and above) instead of a continuous variable. This should be done to weigh the empirical evidence substantially and incorporate what has been learned from the previous research to the current data to draw meaningful conclusions on safety effects of debated explanatory variables, like lane width. Future research efforts should also focus on extending the analysis by adding more stochastic explanatory variables and replicating the Bayesian analysis framework at other site types.

Overall, as new and more complete data become available; these three different methodologies to improve the road safety effect estimation and prediction using multivariate regression models can be applied without any prior assumptions (lack of data). The first method, kriging, can be employed in a larger site/area type, leading to significant difference in the model diagnostics with and without including the new information on daily traffic patterns. This may be possible due to increases in the sample size, and more complete and accurate traffic volume data at many other locations, in addition to the ATR stations. The second method on ME corrections can be employed by generating the ME variance equation from a near ‘gold standard’ value of AADT estimates if new additional data on traffic volumes can be obtained. These changes in employing the methodologies without assumptions will be helpful to the road safety managers in deriving useful conclusions. Eventually, these findings are hoped to lead to determining specific impacts on safety effects estimates and model prediction results, specific to each of the methods.

REFERENCES

1. WHO. Global Status Report on Road Safety: Supporting a Decade of Action. Geneva, Switzerland, 2013.

2. WHO. Global Status Report on Road Safety: Time for Action. Geneva, Switzerland, 2009.

3. Hakkert, A. S., and V. Gitelman. Thinking about the History of Road Safety Research: Past Achievements and Future Challenges. *Transportation Research Part F*, Vol. 25, 2014, pp. 137-149.

4. Blincoe, L., T. R. Miller, E. Zaloshnja, and B. A. Lawrence. The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised). Washington D.C., 2015.

5. FARS. NCSA Data Source Website. *National Highway Traffic Safety Administration (NHTSA)*. http://www-fars.nhtsa.dot.gov/Main/index.aspx. Accessed 2015.

6. NHTSA. U.S. Department of Transportation Announces Decline in Traffic Fatalities in 2013. *National Highway Traffic Safety Administration (NHTSA) 50-14*, 2014. http://www.nhtsa.gov/About+NHTSA/Press+Releases/2014/traffic-deaths-decline-in-2013. Accessed 2015.

7. NHTSA. Traffic Safety Facts: Research Note. Washington D.C., 2014.

8. Insurance Institute for Highway Safety. Saving Lives: Improved Vehicle Designs bring down Death Rates. Virginia, 2015.

9. NHTSA. An Analysis of the Significant Decline in Motor Vehicle Traffic Fatalities in 2008. Washington D.C., 2010.

10. Cathey, J. S. The Effects of Alcohol Regulation and Legislation on Traffic Fatalities in the United States: A Quantitative Analysis. Tennessee, 2009.

11. Waller, P. F. Challenges in Motor Vehicle Safety. *Annu. Rev. Public Health*, Vol. 23, 2002, pp. 93-113.

12. CDC. Motor-Vehicle Safety: a 20th Century Public Health Achievement. 1999.

13. Foot, H. C., A. J. Chapman, and F. M. Wade. *Road Safety: Research and Practice*. Praeger Publishers, New York, 1981.

14. Elvik, R., M. Kolbenstvedt, B. Elvebakk, A. Hervik, and L. Braein. Costs and Benefits to Sweden of Swedish Road Safety Research. *Accident Analysis and Prevention*, Vol. 41, 2009, pp. 387-392.

15. Farmer, E., and E. G. Chambers. A Study of Personal Qualities in Accident Proneness and Proficiency. *Industrial Health Research Board Report*, 1929, p. Report No. 55.

16. Farmer, E. The Method of Grouping by Differential Tests in Relation to Accident Proneness. 1925.

17. NHTSA. Research. *National Highway Traffic Safety Administration*, 2015. http://www.nhtsa.gov/Research. Accessed 2015.

18. Hauer, E. Cause, Effect and Regression in Road Safety: A Case Study. *Accident Analysis and Prevention*, Vol. 42, no. 4, 2010, pp. 1128-1135.

19. Li, C. L. Propensity Score Matching in Observational Studies with Multiple Time Points, 2011. https://etd.ohiolink.edu/rws_etd/document/get/osu1313420291/inline. Accessed 2014.

20. Rosenbaum, P. R. Observational Study. *Encyclopedia of Statistics in Behavioral Science*, 2005, pp. 1451-1462.

21. Rubin, D. B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 1974, pp. 688-701.

22. Cochran, W. G. The Planning of Observational Studies of Human Populations (with Discussion). *Journal of the Royal Statistical Society*, 1965, pp. 134-155.

23. Hosman, C. A. Methods to Control for Overt and Hidden Biases in Comparative Studies, 2011. https://deepblue.lib.umich.edu/handle/2027.42/86498. Accessed 2016.

24. Holland, P. W. Statistics and Causal Inference. *Journal of American Statistical Association*, 1986, pp. 945-960.

25. Rosenbaum, P. R., and D. B. Rubin. Reducing Bias in Observational Studies using

Subclassification on the Propensity Score. *Journal of American Statistical Association*, 1984, pp. 516-524.

26. Gross, F. Case-Control Analysis in Highway Safety: Accounting for Sites with Multiple Crashes. 2011. http://onlinepubs.trb.org/onlinepubs/conferences/2011/RSS/2/Gross,F.pdf. Accessed 2015.

27. Hauer, E. *Observational Before-After Studies in Road Safety*. Pergamon Press, Oxford, UK, 1997.

28. Harwood, D. W., F. M. Council, E. Hauer, W. E. Hughes, and A. Vogt. Prediction of the Expected Safety Performance of the Two-lane Rural Highways. McLean, VA, 2000.

29. Sasidharan, L. Causal Modeling Approach to Determine the Effectiveness of Traffic Safety Countermeasure. Pennsylvania, 2011.

30. Hauer, E., D. W. Harwood, F. M. Council, and M. Griffith. Estimating Safety by Empirical Bayes Method: A Tutorial. *Transportation Research Record:Journal of the Transportation Research Board, No.1784*, 2002, pp. 126-131.

31. Harwood, D. W., E. Bauer, I. B. Potts, D. J. Torbic, K. R. Richard, K. Rabbani, E. Hauer, and L. Elefteriadou. Safety Effectiveness of Intersection Left and Right-Turn lanes. McLean, VA, 2002.

32. Persaud, B., and C. Lyon. Empirical Bayes before–after safety studies: Lessons learned from two decades of experience and future directions. *Accident Analysis and Prevention*, Vol. 39, 2007, pp. 546-555.

33. Hauer, E., D. Harwood, F. M. Council, and M. S. Griffith. Estimating Safety by the Empirical Bayes Method: A Tutorial. *Transportation Research Record:Journal of the Transportation Research Board, 02-2181*, 2002, pp. 126-131.

34. Bonneson, J., and J. Ivan. Theory, Explanation, and Prediction in Road Safety: Promising Directions. Washington D.C., 2013.

35. Srinivasan, R., J. Baek, S. Smith, C. Sundstrom, D. Carter, C. Lyon, B. Persaud, F. Gross, K. Eccles, A. Hamidi, and N. Lefler. Evaluation of Safety Strategies at Signalized Intersection: NCHRP Report 705. Washignton D.C., 2011.

36. Lord, D., and F. L. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research, Part A* , 2010.

37. Donnell, E. T., R. J. Porter, and V. N. Shankar. A Framework for Estimating the Safety Effects of Roadway Lighting at Intersections. *Safety Science*, 2010, pp. 1436-1444.

38. NHTSA. Passenger Vehicle Occupant Fatalities by Day and Night - A Contrast. *Traffic Safety Facts: Research Note*, 2007. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810637. Accessed 2015.

39. Wang, C., and J. N. Ivan. Representing Traffic Exposure in Multi-Vehicle Crash Prediction for Two-Lane Highway Segments. in *Presented at 79th Annual Meeting of the Transportation Research Board*, Washington DC, USA, 2000.

40. Owens, D. A., and M. Sivak. Differentiation of Visibility and Alcohol as Contributors to Twilight Road Fatalities. *Human Factors*, 1996, pp. 680-689.

41. Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models, a Modern Perspective*. Chapman & Hall, CRC Press, Boca Raton, FL, 2006.

42. Chesher, A. The Effect of Measurement Error. *Biometrika*, 1991, pp. 451-462.

43. Mannering, F. L., and C. R. Bhat. Analytic Methods in Accident Research : Methodological Frontier and Future Directions. *Analytic Methods in Accident Research*, Vol. 1, 2014, pp. 1-22.

44. Wundersitz, L. N., and T. P. Hutchinson. Identifying and Improving Exposure Measures. Adelaide, Australia, 2008.

45. Ivan, J. N. New Approach for Including Traffic Volumes in Crash Rate Analysis and Forecasting. *Transportation Research Record: Journal of the Transportation Research Board*, 2004, pp. 134-141.

46. Hakkert , A. S., and L. Braimaister. The Uses of Exposure and Risk in Road Safety Studies. Netherlands, 2002.

47. Elvik, R. *The Handbook of Road Safety Measures, 2nd ed*. Bingley: Emerald Book Serials and Monographs, 2009.

48. Bossche, V. d., G. Wets, and T. Brijs. The Role of Exposure in the Analysis of Road Accidents: A Belgian Case-study. in *Transportation Research Board Annual Meeting CDROM*, Washington DC, USA, 2005, pp. 1-16.

49. Cameron, M., and J. Oxley. Investigation of Improved Exposure Data for the Assessment of Road Safety. Clayton, Victoria, 1995.

50. Wolfe, A. C. The concept of exposure to the risk of a road traffic accident and an overview of exposure data. *Accident Analysis and Prevention*, 1982, pp. 337-340.

51. Kam, B. H. A disaggregate approach to crash rate analysis. *Accident Analysis & Prevention*, 2003, pp. 693-709.

52. Haji, A. G. Towards a Road Safety Development Index (RSDI): Development of an International Index to Measure Road Safety Performance. Norrköping, Sweden, 2005.

53. Ivan, J. N., and C. Zhang. Investigation of a New Approach for Representing Traffic Volumes in Highway Crash Analysis and Forecasting. Storrs, Connecticut, 2008.

54. Qin, X., J. N. Ivan, N. Ravishanker, J. Liu, and D. Tepas. Bayesian estimation of hourly exposure functions by crash type and time of day. *Accident Analysis and Prevention*, 2006, pp. 1071-1080.

55. Greene-Roesel, R., M. Chagas Diogenes, and D. R. Ragland. Estimating pedestrian accident exposure: Protocol report. 2007.

56. Chipman, M. L., C. G. MacGregor, A. M. Smiley, and M. Lee-Gosselin. The role of exposure in comparisons of crash risk among different drivers and driving environments. *Accident Analysis and Prevention*, 1993, pp. 207-211.

57. Zhang, C., J. Ivan, W. ElDessouki, and E. Anagnostou. Relative Risk Analysis for Studying the Impact of Adverse Weather Conditions and Congestion on Traffic Accidents. in *Transportation Research Board 84th Annual Meeting CD ROM*, Washington DC, USA, 2005.

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

59. Agarwal, M., T. Maze, and R. Souleyrette. Impacts of Weather on Urban Freeway Traffic Flow Characteristics and Facility Capacity. in *Proceedings of Mid-Continent Transportation Research Symposium*, Ames, Iowa, 2005.

60. Zhou, M., and V. Sisiopiku. Relationship between volume-to-capacity ratios and accident rates. *Transportation Research Record: Journal of Transportation Research Board*, 1997, pp. 47-52.

61. Fridstrom, L., J. Ifver, S. Ingebrigtsen, R. Kulmala, and L. Thomsen. Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to The Variation in Road Accident Counts. *Accident Analysis and Prevention.*, 1995, pp. 1-20.

62. ERSO. Uses exposure data for road safety analysis. *European Road Safety Observatory*, 2007. http://ec.europa.eu/transport/wcm/road_safety/erso/data/Content/uses_exposure_data_for_road_safety_analysis.htm. Accessed 2015.

63. Ivan, J., C. Wang, and N. Bernardo. Explaining Two-lane Highway Crash Rates Using Lane Use and Hourly Exposure. *Accident Analysis and Prevention*, Vol. 32, 2000, pp. 787-795.

64. Mountain , L., M. Maher, and B. Fawaz. The Influence of Trend on Estimates of Accidents at Junction. *Accident Analysis and Prevention*, Vol. 30, 1998, pp. 641-649.

65. Maher, M., and I. Summersgill. A Comprehensive Methodology for the Fitting of Predictive Accident Models. *Accident Analysis and Prevention*, Vol. 28, 1996, pp. 281-296.

66. Hauer, E. On Exposure and Accident Rate. *Traffic Engineering and Control*, Vol. 36, 1995, pp. 134-138.

67. Persaud, B., and L. Dzbik. Accident Prediction Models for Freeways. *Transportation Research Record: Journal of Transportation Research Board*, Vol. 1401, 1993, pp. 55-60.

68. Garbarino, S., L. Nobili, M. Beelke, F. D. Carli, and F. Ferrillo. The Contributing Role of Sleepiness in Highway Vehicle Accidents. *Sleepiness and Vehicle Accidents*, Vol. 24, no. 2, 2001, pp. 203-206.

69. Langlois, P., M. Smolensky, B. His, and F. Weir. Temporal patterns of reported single-vehicle car and truck accidents in Texas, USA, during 1980–1983. *Chronobiol. Int.* , Vol. 2, 1983, pp. 131-146.

70. Zafar, R., and D. J. Wilson. Traffic Crashes and Traffic Flow Conditions On Auckland's SH1 Southern Motorway. in *IPENZ Transportation Group Conference*, Auckland, Australia, 2011, pp. 1-12.

71. Golob, T. F., W. Recker, and Y. Pavlis. Probabilistic models of freeway safety performance using traffic flow data as predictors. *Safety Science*, 2008, pp. 1306-1333.

72. Golob, T. F., and W. W. Recker. A method for relating type of crash to traffic flow characteristics on urban freeways. *Transportation Research Part A: Policy and Practice*, 2004, pp. 53-80.

73. Martin, J. Relationship between Crash Rate and Hourly Traffic Flow on Interurban

Motorways. *Accident Analysis and Prevention*, Vol. 34, 2002, pp. 619-629.

74. Chang, J., C. Oh, and M. Chang. Effects of Traffic Condition (v/c) on safety at freeway facility sections. in *Transportation Research E-Circular, Fourth International Symposium on Highway Capacity Proceedings*, 2000, pp. 200-208.

75. Persaud, B. N., and K. Mucsi. Microscopic accident potential models for two-lane rural roads. *Transportation Research Record: Journal of Transportation Research Board*, 1995, pp. 134-139.

76. Ceder, A., and M. Liveh. Relationships between road accidents and hourly traffic flow-i analyses and interpretation. *Accident Analysis and Prevention*, 1982, pp. 19-34.

77. Gwynn, D. W. Relationship of accident rates and accidents involvement with hourly volume. *Traffic Quart.*, 1967, pp. 407-418.

78. DelDOT. Automatic Traffic Recorder (ATR) Stations. *Delaware DOT Manuals*, 2003. https://www.deldot.gov/information/pubs_forms/manuals/traffic_counts/2003/pdf/atr_stations.pdf. Accessed 2014.

79. MnDOT. Transportation Research Synthesis: Power Sources for Automatic Traffic Recorders. 2012.

80. Fekpe, E., D. Gopalakrishna, and D. Middleton. Highway Performance Monitoring System Traffic Data For High-Volume Routes: Best Practices And Guidelines. Washington, D.C., 2004.

81. Gadda, S. C., K. M. Kockelman, and A. Magoon. Estimates Of Aadt: Quantifying The Uncertainty. in *World Conference on Transportation Research*, 2007.

82. Wang, X., and K. M. Kockelman. Forecasting Network Data: Spatial Interpolation of Traffic Counts from Texas Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2009, pp. 100-108.

83. Bagheri, E., M. Zhong, and J. Christie. US8805610 B2, 2014.

84. Lam, W. K., Y. F. Tang, K. S. Chan, and M. L. Tam. Short-Term Hourly Traffic Forecasts Using Hong Kong Annual Traffic Census. *Transportation*, 2006, pp. 291-310.

85. Tang, Y. F., W. K. Lam, and P. L. Ng. Comparison of Four Modeling Techniques for Short-Term AADT Forecasting in Hong Kong. *Journal of Transportation Engineering*, 2003, pp. 271-277.

86. Zhao, F., and S. Chung. Contributing Factors of Annual Average Daily Traffic in a Florida County: Exploration with Geographic Information System and Regression Models. *Transportation Research Record: Journal of the Transportation Research Board*, 2001, pp. 113-122.

87. Lam, W. H.K., and J. Xu. Estimation of AADT from Short Period Counts in HongKong – A Comparison between Neural Network Method and Regression Analysis. *Journal of Advanced Transportation*, 2000, pp. 249-268.

88. Shamo, B., E. Asa, and J. Membah. Linear Spatial Interpolation and Analysis of Annual Average Daily Traffic Data. *Journal of Computing in Civil Engineering*, 2015.

89. Avanti Engineering Group. (2014) Use of Spatial Interpolation to Estimate Interurban Traffic Flows from Traffic Counts. Accessed 2015.

90. Selby, B., and K. M. Kockelman. SPATIAL PREDICTION OF AADT IN UNMEASURED LOCATIONS BY UNIVERSAL KRIGING. in *90th Annual Meeting of the Transportation Research Board* , Washington DC, 2011.

91. Eom, J. K., M. S. Park, Y. Heo, and L. F. Huntsunger. Improved Prediction of Annual Average Daily Traffic for Non-Freeway Facilities by Applying Spatial Statistical Method. in *Proceedings of the 85th Annual Meeting of the Transportation Research Board*, Washington D.C., 2006.

92. Matheron, G. Principles of Geostatistics. *Economic Geology*, 1963, pp. 1246-1266.

93. Krige, D. G. A Statistical Approach to Some Mine Valuations and Allied Problems at the Witwatersrand. Johannesburg, South Africa, 1951.

94. Bayraktar, H., and F. S. Turalioglu. A Kriging-Based Approach for Locating a Sampling Site in the Assessment of Air Quality. *Stochastic Environmental Research and Risk Assessment*, 2005, pp. 301-305.

95. Emery, X. Simple and Ordinary Kriging Multigaussian Kriging for Estimating Recoverable Reserves. *Mathematical Geology*, 2005, pp. 295-319.

96. Zimmerman, D. A., G. D. Marsily, C. A. Gotway, and M. G. Marietta. A Comparison of Seven Geostatistically Based Inverse Approaches to Estimate Transmissivies for Modelling Advective Transport by Groundwater Flow. *Water Resource Research*, 1998, pp. 1273-1413.

97. Zhao, F., and N. Park. Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic. *Transportation Research Record: Journal*

of the Transportation Research Board, 2004, pp. 99-107.

98. Parmentier, B., B. McGill, A. M. Wilson , J. Regetz, W. Jetz, R. P. Guralnick, and M. Schildhauer. An Assessment of Methods and Remote-Sensing Derived Covariates for Regional Predictions of 1 km Daily Maximum Air Temperature. *Remote Sensing*, 2014, pp. 8639-8670.

99. Ijaz, H. Spatial Interpolation of Sulfate Concentration in Groundwater Including Covariates Using Bayesian Hierarchical Models. *Water Quality, Exposure and Health*, 2015, pp. 339-345.

100. Mair, A., and A. Fares. Comparison of Rainfall Interpolation Methods in a Mountainous Region of a Tropical Island. *Hydrologic Engineering*, 2011, pp. 371-383.

101. Snepvangers, J., G. Heuvelink, and J. Huisman. Soilwater content interpolation using spatio-temporal kriging with external drift. *Geoderma*, 2003, pp. 253-271.

102. Hudson, G., and H. Wackernagel. Mapping temperature using kriging with external drift: theory and an example from scotland. *International Journal of Climatol*, 1994, pp. 77-91.

103. Valesco, e.a. Merging Radar and raingauges data to estimate rainfall fields: An improved geostatistical approach using non-parametric Spatial models. in *Sixth International Symposium on hydrological applications of weather radar*, Melbourne, Australia, 2004.

104. Kravchenko, A., and et. al. Estimation of Mean Annual Precipitation in Wyoming Using Geostatistical Analysis. in *American Geophysical Union 16th Annual Hydrology Days*, Fort Collins, Colorado, 1996, pp. 271-282.

105. Romann, A. Evaluating the Performance of Simulation Extrapolation and Bayesian Adjustments for Measurement Error. Vancouver, Canada, 2008.

106. Guolo, A. Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research*, 2008, pp. 555-580.

107. Guo, Y. Multiple Imputation for Measurement Error Correction Based on a Calibration Sample. Ann Harbor, Michigan, 2010.

108. Buonaccorsi, J. P. *Measurement Error: Models, Methods, and Applications*. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL, 2010.

109. Armstrong, B. K., E. White, and R. Saracci. *Principles of Exposure Measurement in*

Epidemiology. Oxford Medical Publications, New York, 1992.

110. Gleser, L., R. Carroll, and P. Gallo. The Limiting Distribution of Least Sqaures in an Errors-in-Variables Regression Model. *Annals of Statistics*, 1987, pp. 220-233.

111. Carroll, R., P. Gallo, and L. Gleser. Comparison of Least Squares and Errors-in-Variables Regression, with Special Reference to Randomized Analysis of Covariance. *Journal of the American Statistical Association*, 1985, pp. 929-932.

112. Cabral, C. R., V. H. Lachos, and C. B. Zeller. Multivariate measurement error models using finite mixtures of skew-Student t distributions. *Journal of Multivariate Analysis*, 2014, pp. 179-198.

113. Fung, K. Y., and D. Krewski. On Measurement Error Adjustment Methods in Poisson Regression. *Environmetrics*, 1999, pp. 213-224.

114. Bang, H., Y. L. Chiu, J. S. Kaufman, M. D. Patel, G. Heiss, and K. M. Rose. Bias Correction Methods for Misclassified Covariates in the COx Model: Comparison of Five Correction Methods by Simulation and Data Analysis. *Journal of Stat Theory and Practice*, 2013, pp. 381-400.

115. Messer, K., and L. Natarajan. Maximum Likelihood, Multiple Imputation, and Regression Calibration for Measurement Error Adjustment. *Statistics in Medicine*, 2008, pp. 6332-6350.

116. Freedman, L. S., D. Midthune, R. J. Carroll, and V. Kipnis. A Comparison of Regression Calibration, Moment Reconstruction and Imputation for Adjusting for Covariate Measurement Error. *Statistics in Medicine*, 2008, pp. 5195-5216.

117. Carroll, R. J. Measurement Error in Epidemiologic Studies. College Station, Texas, 2016.

118. Pierce, D. A., and A. M. Kellerer. Adjusting for Covariate Errors with Nonparametric Assessment of the True Covariate Distribution. *Biometrika*, 2004, pp. 863-876.

119. Buzas, J. S., T. D. Tosteson, and L. A. Stefanski. Measurement Error. 2003.

120. Hardin, J. W., H. Schmiediche, and R. J. Carroll. The Regression Calibration Method for Fitting Generalized Linear Models with Additive Measurement Error. 2016.

121. Prentice, R. L. Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model. *Biometrika*, 1982, pp. 331-342.

122. Carroll, R. J., and L. A. Stefanski. Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 1990, pp. 652-663.

123. Rosner, B., D. Spiegelman, and W. C. Willett. Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Measurement Error: The Case of Multiple Covariates Measured with Error. *American Journal of Epidemiology*, 1990, pp. 734-745.

124. Rosner, B., W. C. Willett, and D. Spiegelman. Correction of Logistics Regression Relative Risk Estimates and Confidence Intervals for Systematic Within Person Measurement Error (with Discussion). *Statistics in Medicine*, 1989, pp. 1051-1069.

125. Cook, J., and L. A. Stefanski. A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, 1995, pp. 1314-1328.

126. Stefanski, L. A., and J. Cook. Simulation extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 1995, pp. 1247-1256.

127. Greene, W. F., and J. Cai. Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics*, 2004, pp. 987-996.

128. Hwang, W. H., and S. Y. Huang. Estimation in capture-recapture Models when covariates are subject to Measurement Errors. *Biometics*, 2003, pp. 1113-1122.

129. Wang, C. Y., N. Wang, and S. Wang. Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, 2000, pp. 487-495.

130. Solow, A. R. On fitting a population model in the presence of observation error. *Ecology*, 1998, pp. 1463-1466.

131. Khattak, A. J., and F. M. Council. Effects of work zone presence on injury and non-injury crashes. *Accident Analysis and Prevention*, 2002, pp. 19-29.

132. Venugopal, S., and A. Tarko. Safety models for rural freeway work zones. *Transportation Research Record. Journal of the Transportation Research Board*, 2000, pp. 1-9.

133. El-Basyouny, K., and T. Sayed. Safety performance functions with measurement errors in traffic volume. *Safety Science*, 2010, pp. 1339-1344.

134. Jang, H., S. Lee, and S. W. Kim. Bayesian analysis for zero-inflated regression models with the power prior: Applications to road safety countermeasures. *Accident Analysis and Prevention*, Vol. 42, 2010, pp. 540-547.

135. Amador, L., and D. Mrawira. Bayesian regression in pavement deterioration modeling: revisiting the AASHO road test rut depth model. in *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, Washington DC, 2011.

136. Gelman, A., and J. Hill. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press, Cambridge, UK, 2007.

137. Mitra, S., and S. P. Washington. On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis and Prevention*, Vol. 39, 2006, pp. 459-468.

138. Guerrero-Barbosa, T. E., and G. E. Amarís-Castro. Application of Bayesian techniques for the identification of accident-prone road sections. *DYNA*, Vol. 81, no. 187, 2014, pp. 209-214.

139. Tuzel, O. Y., and K. Ozbay. A comparative Full Bayesian before-and-after analysis and application to urban road safety countermeasures in New Jersey. *Accident Analysis and Prevention*, Vol. 42, 2010, pp. 2099-2107.

140. Miranda-Moreno, L. F., L. Fu, F. Saccomano, and A. Labbe. Alternative Risk Models for Ranking Locations for Safety Improvement.. *Transportation Research Record:Journal of Transportation Research Board*, Vol. 1908, 2005, pp. 1-8.

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

142. Persaud, B., C. Lyon, and T. Nguyen. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1665, 1999, pp. 7-12.

143. Higle, J. L., and J. M. Witkowski. Bayesian Identification of Hazardous Sites. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1185, 1988, pp. 24-35.

144. Lee, S., J. Choi, and S. W. Kim. Bayesian approach with the power prior for road safety analysis. *Transportmetrica*, Vol. 6, no. 1, 2010, pp. 39-51.

145. Morris, W. K., P. A. Vesk, and M. A. McCarthy. Profiting from pilot studies:

Analysing mortality using Bayesian models with informative priors. *Basic and Applied Ecology*, Vol. 14, 2013, pp. 81-89.

146. Clyde, M. Bayesian model averaging: A tutorial: Comment. *Statistical Science*, Vol. 14, 1999, pp. 401-404.

147. Dennis, B. Discussion: Should ecologists become Bayesians?. *Ecological Applications*, Vol. 6, 1996, pp. 1095-1103.

148. McCarthy, M. A. *Bayesian methods for ecology*. Cambridge University Press., Cambridge, UK, 2007.

149. Lord, D., and L. Miranda-Moreno. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson- gamma models for modeling motor vehicle crashes: a Bayesian perspective. *Safety Science*, Vol. 46, 2008, pp. 751-770.

150. Washington, S., and J. Oh. Bayesian methodology incorporating expert judgment for ranking countermeasure effectiveness under uncertainty: example applied to at grade railroad crossings in Korea. *Accident Analysis and Prevention*, Vol. 38, 2006, pp. 234-247.

151. Yu, R., and M. Abdel-Aty. Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. *Accident Analysis and Prevention*, Vol. 56, 2013, pp. 51-58.

152. Choi, S., H. Kim, and S. W. Kim. Propriety of the power prior for the weibull regression model. *The East Journal of Theoretical Statistics*, Vol. 14, 2004, pp. 89-101.

153. Ibrahim, J. G., and M. H. Chen. Power distribution for regression models.. *Statistical Science*, Vol. 15, no. 1, 2000, pp. 46-60.

154. Chen, M. H., A. K. Manatunga, and C. J. Williams. Heritability estimates from human twin data by incorporating historical prior information. *Biometrics*, Vol. 54, no. 4, 1998, pp. 1348-1362.

155. Knecht, C. S. Crash Prediction Modeling for Curved Segments of Rural Two-Lane Two-Way Highways in Utah. Provo, Utah, 2014.

156. Ona, J. d., L. Garach, F. Calvo, and G. Munoz. Relationship Between Predicted Speed Reduction on Horizontal Curves and Safety on Two-Lane Rural Roads in Spain. *Journal of Transportation Engineering*, Vol. 140, no. 3, 2014.

157. Lord, D., M. A. Brewer, K. Fitzpatrick, S. R. Geedipally, and Y. Peng. Analysis of Roadway Departure Crashes on Two-Lane Rural Roads in Texas. College Station, Texas, 2011.

158. Aram, A. Effective Safety Factors on Horizontal Curves of Two-Lane Highways. *Journal of Applied Sciences*, Vol. 10, no. 22, 2010, pp. 2814-2822.

159. Voigt, A. Evaluation of Alternative Horizontal Curve Design Approaches on Rural Two-Lane Highways. College Station, Texas, 1996.

160. Zegeer, C. V., R. J. Stewart, F. M. Council, D. W. Reinfurt, and E. Hamilton. Safety Effects of Geometric Improvements on Horizontal Curves. *Transportation Research Record: Journal of the Transportation Research Board*, 1992, pp. 11-19.

161. Zegeer, C. V., D. W. Reinfurt, W. W. Hunter, J. Hummer, R. Stewart, and L. Herf. Accident Effects of Sideslope and Other Roadside Features on Two-Lane Roads. *Transportation Research Record: Journal of the Transportation Research Board*, 1988, pp. 33-47.

162. Bella, F. Coordination of Horizontal and Sag Vertical Curves on Two-Lane Rural Roads: Driving Simulator Study. *International Association of Traffic and Safety Sciences*, no. 39, 2015, pp. 51-57.

163. Bauer, K. M., and D. W. Harwood. Safety Effects of Horizontal Curve and Grade Combinations on Rural Two-Lane Highways. *Transportation Research Record: Journal of Transportation Research Board*, no. 2398, 2013, pp. 37-49.

164. VTT Communities. Safety Evaluation of Different Kinds of Cross-Section on Rural Two-Lane Roads. *Safestar Deliverable 4.1*, 2015. http://plaza.ufl.edu/katrinas/route/xSection.pdf. Accessed 2016.

165. Gross, F., P. P. Jovanis, and K. Eccles. Safety Effectiveness of Lane and Shoulder Width Combinations on Rural, Two-Lane, Undivided Roads. *Transoportation Research Record, Journal of Transportation Research Board*, no. 2103, 2009, pp. 42-49.

166. Fitzpatrick, K., W. H. Schneider IV, and E. S. Park. Comparisons of Crashes on Rural Two-Lane and Four-Lane Highways in Texas. College Station, Texas, 2005.

167. Miaou, S. P., and H. Lum. Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis and Prevention*, Vol. 25, no. 6, 1993, pp. 689-709.

168. Okamoto, H., and M. Koshi. A method to cope with the random errors of observed

accident rates in regression analysis. *Accident Analysis and Prevention,* Vol. 21, no. 4, 1989, pp. 317-332.

169. Zegeer, C. V., R. C. Deen, and J. G. Mayes. The Effect of Lane and Shoulder Widths on Accident Reductions on Rural Two-Lane Roads. Commonwealth of Kentucky, 1980.

170. Fitzpatrick, K., E. L. Park, and W. H. Schneider. Potential accident modification factors for driveway density on rural highways. *Transportation Research Record: Journal of the Transportation Research Board*, 2008, pp. 49-61.

171. Boodlal, L., E. T. Donnell, R. J. Porter, D. Garimella, T. Le, K. Croshaw, S. Himes, P. Kulis, and J. Wood. Factors Influencing Operating Speeds and Safety on Rural and Suburban Roads. McLean, Virginia, 2015.

172. Zhong, C., V. P. Sisiopiku, K. Ksaibati, and T. Zhong. Crash Prediction on Rural Roads. in *3rd International Conference on Road Safety and Simulation*, Indianapolis, USA, 2011.

173. Souleyrette, R. R., M. Caputcu, T. J. McDonald, R. B. Sperry, Z. N. Hans, and D. Cook. Safety Analysis of Low-Volume Rural Roads in Iowa. Ames, Iowa, 2010.

174. Mayora, J. M.P., and R. L. Rubio. Relevant Variables for Crash Rate Prediction in Spain's Two Lane Rural Roads. in *Transportation Research Board, 82nd Annual Meeting*, Washington D.C., 2003.

175. Garber, N. J., and E. A. Kassebaum. Evaluation of Crash Rates and Causal Factors for High-Risk Locations on Rural and Urban Two-Lane Highways in Virginia. Richmond, Virginia, 2008.

176. Rengarasu, T. M., T. Hagiwara, and M. Hirasawa. Effects of Road Geometry and Season on Head-on and Single-Vehicle Collisions on Rural Two-Lane Roads in Hokkaido, Japan. in *Eastern Asia Society for Transportation Studies, Vol.6,* 2007.

177. Valverde, J. A., P. P. Jovanis, and P. G. Knight. Advanced Road Safety and Weather Warning System. Pennsylvania, 2006.

178. TimeandDate. Sunrise, Sunset, and Moon Times. http://www.timeanddate.com/astronomy/usa/salt-lake-city. Accessed 2016.

179. SunriseSunset. Sunrise Sunset Calendar. http://www.sunrisesunset.com/calendar.asp. Accessed 2015.

180. Oliver, M. A., and R. Webster. A Tutorial Guide to Geostatistics: Computing and

Modeling Variograms and Kriging. *Catena*, 2013, pp. 57-69.

181. Cressie, N. *Statistics for Spatial Data Revised Edition*. Wiley: New York, 2015.

182. Majani, B. S. Analysis of External drift kriging algorithm with application to precipitation estimation in complex orography. Enschede, The Netherlands, 2007.

183. Fuller, W. *Measurement Error Models*. John Wiley & Sons, New York, 1987.

184. Armstrong, B. Measurement Error in Generalized Linear Models. *Communications in Statistics*, Vol. 14, 1985, pp. 529-544.

185. Carroll, R. J. (2011) Measurement Error in Covariates: Session 2, May27, 2011.

186. Hardin, J. W., H. Schmiediche, and R. J. Carroll. The Simulation Extrapolation Method for Fitting Generalized Linear Models with Additive Measurement Error. 2016.

187. Lederer, W., and H. Kuchenhoff. (2016) A Short Introduction to the SIMEX and MCSIMEX. Accessed 2016.

188. Apanasovich, T. V., and R. J. Carroll. SIMEX and standard error estimation in semiparametric measurement error models. *Electronic Journal of Statistics*, 2009, pp. 318-348.

189. Miranda-Moreno, L. F., S. Heydari, D. Lord, and L. Fu. Bayesian road safety analysis:Incorporation of past evidence and effect of hyper-prior choice. *Journal of Safety Research*, 2013, pp. 31-40.

190. Heydari, S., L. F. Miranda-Moreno, D. Lord, and L. Fu. Bayesian Methodology to Estimate and Update SPF Parameters under Limited Data Conditions: A Sensitivity Analysis. in *International Road Safety and Simulation Conference*, Rome, Italy, 2013.

191. Duan, Y. A Modified Bayesian Power Prior Approach with Applications in Water Quality Evaluation. Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, Dissertation 2005.

192. Washington, S. P., M. G. Karlaftis, and F. L. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2011.

193. Anastasopoulos, P., and F. Mannering. A Note on Modeling Vehicle Accident

Frequencies with Random Parameters Count Models. *Accident Analysis and Prevention*, 2008, pp. 153-159.

194. McCarthy, P. Effect of Speed Limits on Speed Distributions and Highway Safety: A Survey of the Literature. Washington D.C., 1998.

195. Shinar, D. Speed and Crashes: A Controversial Topic and an Elusive Relationship. In Managing Speed. Washington D.C., 1998.

196. Turner, R. M., S. G. Thompson, and D. J. Spiegelhalter. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. in *Clinical Trials: Workshop Article*, 2005, pp. 108-118.

197. CMF Clearinghouse. Star Quality Rating. *Crash Modification Factors Clearinghouse*. http://www.cmfclearinghouse.org/sqr.cfm. Accessed 2016.

198. Smith, E. Bayesian Modeling of Extreme Rainfall Data. 2005.

199. ARTBA. Transportation FAQs. *American Road & Transportation Builders Association*. http://www.artba.org/about/transportation-faqs/. Accessed 2015.

200. NHTSA. Traffic Safety Facts: 2012 Data. Washington D.C., 2014.

201. TRIP. Rural Connections: Challenges and Opportunities in America's Heartland. Washington D.C., 2015.

202. Krammes, R. A., R. Q. Brackett, M. A. Shafer, J. L. Otteson, I. B. Anderson , K. L. Fink, K. M. Collins, O. J. Pendleton, and C. J. Messer. Horizontal Alignment Design Consistency for Rural Two-Lane Highways. Washington D.C., 1995.

203. Zegeer, C., J. Twomey, M. Heckman, and J. Hayward. Safety Effectiveness of Highway Design Features Volume II, Alignment. Washington D.C., 1992.

204. Labi, S., and CATS. Effects of Geometric Characteristics of Rural Two-Lane Roads on Safety. Indiana: Pudue University, FHWA/IN/JTRP-2005/2, 2006.

205. FHWA. Highway Safety Improvement Program Manual. Washington D.C., 2010.

206. Hilbe, J. *Negative Binomial Regression, 2nd Edition*. Cambridge University Press, New York, 2011.

207. Kennedy, P. *A Guide to Econometrics*. MIT Press, 2003.

208. Gates, T., and et. al. Safety and Operational Impacts of Differential Speed Limits on Two-Lane Rural Highways in Montana. Helena, Montana, 2016.

209. Persaud, B. N., R. A. Retting, and C. A. Lyon. Crash Reduction following Installation of Centerline Rumble Strips on Rural Two-Lane Roads. *Accident Analysis and Prevention*, 2004, pp. 1073-1079.

210. Cafisco, S., G. L. Cava, and A. Montella. Safety Index for Evaluation of Two-Lane Rural Highways. *Transportation Research Record: Journal of Transportation Research Board*, 2007, pp. 136-145.

211. Geedipally, S. R., S. Patil, and D. Lord. Examination of Methods to Estimate Crash Counts by Collision Type. *Transportation Research Record: Journal of Transportation Research Board*, 2010, pp. 12-20.

212. Tarko, A. P., S. Eranky, K. C. Sinha, and R. Scinteie. An Attempt to Develop Crash Reduction Factors Using Regression Technique. in *Transportation Research Board 78th Annual Meeting*, Washington D.C., 1999.

213. Gooch, J. P., V. V. Gayah, and E. T. Donnell. Quantifying the Safety Effects of Horizontal Curves on Two-Way, Two-Lane Roads. *Accident Analysis and Prevention*, 2016, pp. 71-81.

214. Momeni, H. Identifying Effective Geometric and Traffic Factors to Predict Crashes at Horizontal Curve Sections. Manhattan, Kansas, Accessed 2016.