Safety Analysis of Freeway segments with unobserved heterogeneity and Second order spatial effects

Eneliko Mujuni Mulokozi
University of Nevada, Las Vegas, mulokozi@unlv.nevada.edu

SAFETY ANALYSIS OF FREEWAY SEGMENTS WITH UNOBSERVED HETEROGENEITY AND SECOND ORDER SPATIAL EFFECTS

By

Eneliko Mulokozi

Bachelor of Science (Civil & Structural Engineering)
University of Dar es salaam
2003

Master of Science in Engineering
University of Nevada, Las Vegas
2013

A dissertation submitted in partial fulfillment of
the requirements for the

Doctor of Philosophy - Civil and Environmental Engineering

Department of Civil and Environmental Engineering
Howard R. Hughes College of Engineering
Graduate College

University of Nevada, Las Vegas
December 2015

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

November 10, 2015

This dissertation prepared by

Eneliko Mulokozi

entitled

Safety Analysis of Freeway Segments with Unobserved Heterogeneity and Second Order Spatial Effects

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Engineering – Civil Engineering
Department of Civil and Environmental Engineering and Construction

Hualiang Teng, Ph.D.
Examination Committee Co-Chair

Mohamed Kaseko, Ph.D.
Examination Committee Co-Chair

Alexander Paz, Ph.D.
Examination Committee Member

Mose Karakouzian, Ph.D.
Examination Committee Member

Djeto Assane, Ph.D.
Graduate College Faculty Representative

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

ABSTRACT

Safety analysis of Freeway segments with unobserved heterogeneity and second order spatial effects.

By

Eneliko Mulokozi

Dr. Hualiang (Harry) Teng, Examination Committee Chair
Associate Professor of Civil and Environmental Engineering and Construction
University of Nevada, Las Vegas

Safety analysis of freeway networks entails the quantification of crash frequency influencing factors which include roadway and traffic characteristics, environmental factors as well as human factors. This quantification can be used to detect locations with large impacts on the occurrence of crashes which in turn assist engineers and planners to improve safety levels of the network. Roadway characteristics are comprised of the physical elements of the road geometry such as section length, median and right shoulders, speed-exchange lanes, the number of main facility as well as geometry of the entrance from and exit to the main freeway facility. Traffic characteristics are comprised of traffic flow and vehicular volumes while environmental factors include weather conditions, pavement surface conditions, work zone areas conditions, and lighting conditions along the travel facility. Human factors are comprised of aging, aggressiveness while driving, mental stability, fatigue, alcoholism, acute psychological stress, suicidal behavior, drowsiness, and temporary distraction.

Variability in the crash frequency is captured by the interaction of the aforementioned factors either in a multiplicative or additive nature through the use of statistical model formulation. When all factors believed to influence the occurrence of crashes are included in a mathematical formulation and all the assumptions underlying the statistical model are met, variability in the crash frequency referred to as observed heterogeneity can be fully explained. However, not all information believed to generate crashes is available. Some of the factors are latent in nature and some are either not available at the time of analysis or require time and high cost to be established. When such conditions exist, a formulated model does not fully explain observed heterogeneity in the crash frequency. Lack of information to fully explain variability in crash frequency as a result of excluding some factors leads to unobserved heterogeneity problems which results in biased and inconsistent safety estimators.

Specifically, when observed crash counts are considered as clusters, analytical approach should consider the possibility of dependence within clustered crash counts. Correlation within clusters may be due to variation being induced by common unobserved cluster-specific factors. Ignoring cluster-effects increases the likelihood in drawing conclusion based on unrealistic inferences because safety estimator standard errors are likely to be underestimated and the usual conditional mean is no longer correctly specified. Cross sectional dependence may also arise when the crash counts have a spatial dimension due to contiguous freeway segments. Such conditions lead to what is known as spatial autocorrelation. This is the presence of spatial pattern in crash frequency over space due to geographic proximity whereby high values of crash frequency tend to cluster together in adjacent freeway segments or high crash frequencies are contiguous with low values of crash frequencies. When the distribution of crash frequency over space exhibit the aforementioned pattern, safety analysis techniques based on the distributional assumption of independence of crash frequency is violated.

This study has two objectives: First, analyze safety of freeway geometric features while accounting for the effect of unobserved influencing factors and cluster-specific effects; Second, analyze safety of freeway geometric elements in the presence of spatial autocorrelation due to geographical proximity effects. To achieve the first objective, four models are compared: Two are standard Poisson and Negative binomial regression models which do not account for cluster effects. The other two are mixed effects Poisson and Negative binomial regression models which in addition to fixed effects parts they account for the effects of randomness arising from heterogeneity and clustering.

The empirical results indicate that 13.9% of the variation in crash frequency is unaccounted for, which is an indication of the existence of unobserved factors influencing the occurrence of crashes. It is also revealed that weaving segments (EN-EX) had the highest between segment variance compared to non-weaving segments. More vehicles and short segments increased crash frequency while wider right shoulder decreased the crash frequency.

It is also observed that weaving segments decreased crash frequency compared to non-weaving segments. These results indicate that by allowing parameters to vary within the weaving and non-weaving segments it is possible to capture and quantify unobserved factors. Ignoring these factors results in biased coefficients because the estimate of the standard errors required determining inferential statistics will be wrong.

To achieve the second objective, Conditional Autoregressive models in Bayesian setting framework (CAR) is used. CAR models recognize the presence of spatial dependence which helps to obtain unbiased estimates of parameters quantifying safety levels since the effects of spatial autocorrelation is accounted for in the modeling process.

Based on CAR models, approximately 51% of crash frequencies across contiguous freeway segments are spatially autocorrelated. The incident rate ratios revealed that wider shoulder and weaving segments decreased crash frequency by factors of 0.84 and 0.75 respectively. The marginal impact graphs showed that an increase in longitudinal space for segments with two lanes decreased crash frequency. However, an increase of facility width above three lanes results in more crashes which indicates an increase in traffic flows and driving behavior leading to crashes. These results call an important step of analyzing contagious freeway segments simultaneously to account for the existence of spatial autocorrelation.

ACKNOWLEDGEMENT

The successful accomplishment of my dissertation work was possible through the support of Professor Hualiang (Harry) Teng, who continuously provided support during my graduate studies both at the masters and PhD levels. I also recognize and thank all committee members of my dissertation: Professor Moses Karakouzian, Professor Mohamed Kaseko, Professor Alexander Paz, and Professor Djeto Assane. Their assistance contributed to timely completion of my work. I also thank Professor Mohamed Kaseko for his financial support when I first came to UNLV.

DEDICATION

To my mother Judith, my wife Beatrice, and our beloved children Irene and Moses

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF DEFINITIONS

AADT Annual Average daily Traffic

AIC Akaike Information Criterion - a measure of the relative quality of statistical models for a given set of data

BIC Bayesian Information Criterion – a criterion for model selection among a finite set of models

CAR Conditional Autoregressive Model – a model designed to address spatial dependence

CI Credible Interval defined as an interval in the domain of a posterior probability distribution used for interval estimation.

EN-EX Freeway segment where “EN” means entrance and “EX” means exit

EX-EN Freeway segment where “EX” means exit and “EN” means entrance

EX-EX Freeway segment with all exits

EN-EN Freeway segment with all entrances

$f(Y/U\,,\omega)$ Probability distribution of crash frequency at level 1 of the multilevel model

f_U Probability distribution of random effects given the parameters of interest

$g(Y,U/\,\omega)$ Joint distribution of crash frequency and random effects vectors

GLM Generalized linear model - a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

PO Poisson regression model

ME PO Mixed effects Poisson regression model

NB Negative Binomial regression model

ME NB	Mixed effects Negative Binomial regression model
u_{oj}	Cluster effects residuals that summarize all unobserved factors generating crashes

CHAPTER 1 INTRODUCTION

1.0 Motivational background

Safety analysis of a road network involve the application of statistical models which can be used to explain variability in the safety performance measure in terms of the variability in the observed factors believed to influence crash occurrence (Simon, Matthew, & Fred L, 2011), However, not all relevant information is observed to absolutely describe the source of variance in crash events because of either latent nature of this information or not easily obtained when needed. Missing some of the relevant factors in addition to their variability leads to what is known as unobserved heterogeneity which represent the combined effect of all unobserved factors (Skrondal & Rabe-Hesketh, 2004).

When observation units are clustered in space, it is expected that there are shared unobserved heterogeneity that leads to intra-cluster dependence which violates distribution assumption of a statistical model. In addition to clustering effects and unobserved heterogeneity, crash events tend to co-vary in space leading to what is known as spatial autocorrelation, a type of spatial effects. Spatial autocorrelation is the spatial phenomenon exhibited in crash counts such that high crash counts are found near high values and low values are found near low values (Bolstad, 2005).

Understanding factors involved in crash generating mechanisms have been an active area of research in traffic safety. Alternative methods are proposed to addrress complex issuess relating to unobserved heterogeneity and spatial correlation in crash frequency. However, due to the complex nature of the factors leading to crash occurrence and complex data structure required to capture most of the sources leading to variability and spatially correlated crash frequency, the current research has not fully describe most of the concerns arising due to unobserved factors and spatially correlated crash frequency.

Addrressing these issues require the use of appropriate micro data structures that capture most of the variability in crash frequency. The current data structures used do not account for unobserved heterogeneity at all levels. For instance, (Venkataraman, Ulfarsson, Shankar, Oh, & Park, 2011) modeled the relationship between interstate crash occurrence and geometrics by using freeway segments at the interchange and noninterchange levels bu accounted only variation within the explanatory variables by estimating random parameters.

Furthermore, Venkataraman, *et al* (2011) defined interchange segments by the furthest merge and diverge ramp limits for each direction. This definition indicates that a segment contained more than one interchange. Furthermore, noneinterchange segments were defined as a continuous travel segment between two interchanges. However, at the segment level it is possible to encounter factors which are likely to be involved in crash generating mechanism thereby causing one segment to vary from another in terms of crash occurrence experiences. Such differential variability across segments arises from the fact that some of the may factors affecting the frequency and severity of crashes are not observable and can be attributed to a specific segment. It is also tru that the necessary data required may be nearly impossible to collect. If these unobserved factors, referred to as unobserved heterogeneity, are correlated with observed factors, biased parameter estimators will be estimated and incorrect inferences could be drawn.

Other studies such as (Anastasopoulos & Mannering, 2009) and (Veeraragavan & Dinu, 2011) have also addressed these complex issues but they did not capture variability in crash frequency at all levels. For instance, (Anastasopoulos, *et al*, 2009) used segments with homogeneous characteristics but accounted for random parameters only. However, even if a segment is homogeneous, there are still differences in terms of unobserved factors across these segmenst which need to be accounted for but were not captured. (Veeraragavan, *et al*, 2011) divided the highway into homogenous sections based on traffic volume, carriageway width, and shoulder width but only random parameters(slopes) were estimated. This leaves out variability at the segment level due to unobserved factors likely to be involved in crash generating mechanisms.

The purpose of this research is to conduct analysis of safety performance measures on freeway segments while accounting for the shared unobserved heterogeneity and spatial correlation. In this study, freeway data from the Las Vegas Area in State of Nevada as illustrated in Figure 1 below are used. The following subsections explain in detail the aforementioned problems in terms of the roadway geometric characteristics and methods available to address these problems. Furthermore, importance of the proposed research to transportation engineers is explained.

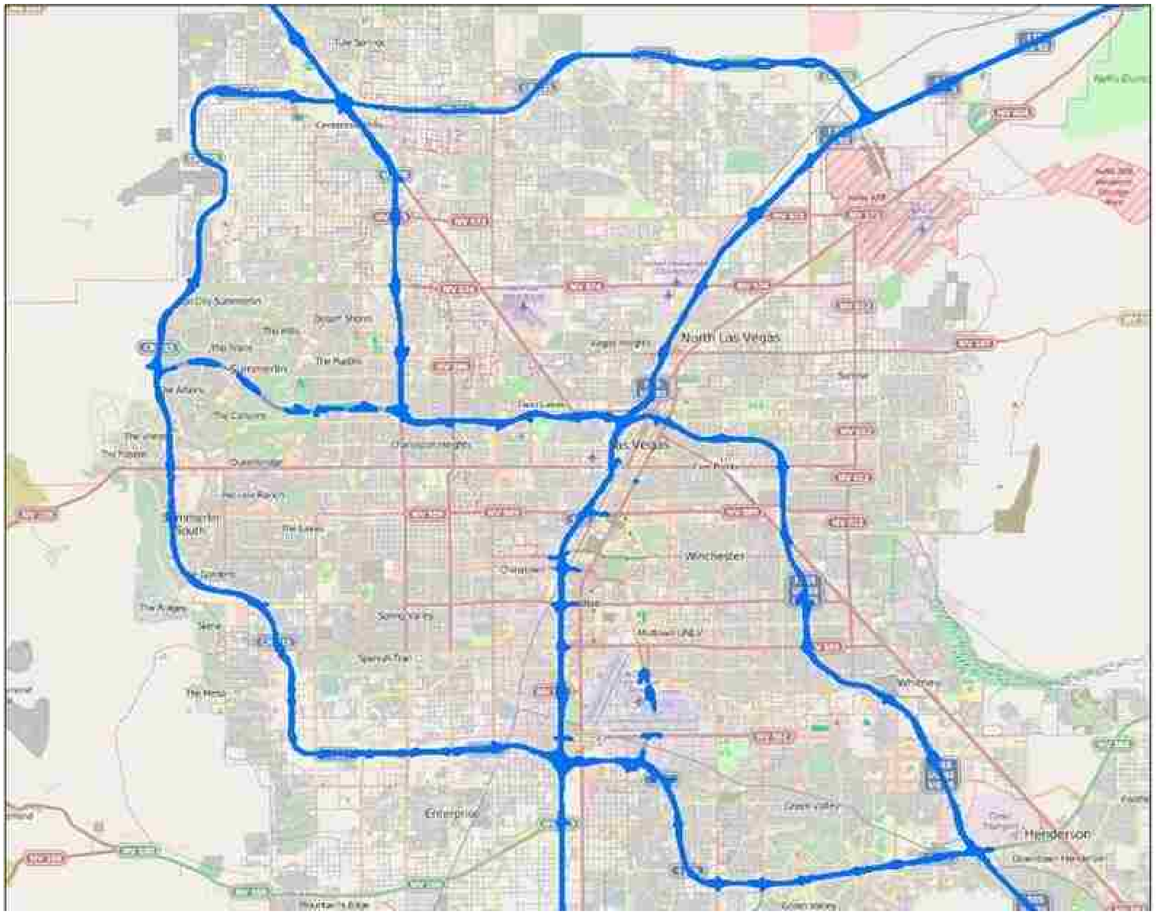

Figure 1: Freeway networks (blue) under study

1.1 Problem statement

The level of safety of freeways depends on the interaction of a number of factors including elements of the road geometry, traffic characteristics, environmental and human factors (Ogden, 1996). Elements of the road geometry constitute the number of lanes on the main facility, length of a segments, shoulders, medians, entrances to and exits from the main facility as well as auxiliary lanes. Traffic characteristics include traffic flows, traffic mix and vehicular speed while environmental factors are made up of weather and pavement surface conditions. Human factors include driving behavior leading to inappropriate and excessive speeding, age, fatigue, inattention, and impairment such as inadequate visibility. Safety level is manifested in terms of crash frequency or severity when the aforementioned factors interact in such a way that one or more of the factors do not execute an appropriate function.

Crash frequency, a measure of safety level, is related to the combination of influencing factors through a mathematical formulation which identifies individual factors contribution to the overall safety level (Sarhan, Hassan, & Adb El Halim, 2008; Shankar, Mannering, & Barfield, 1995; Chen, Liu, Lu, & Behzad, 2009; Liu, Chen, Lu, & Cao, 2010; Chen, Zhou, Zhao, & Hsu, 2011; Golob, Recker, & Alvarez, 2004; Joe, Greg, & Davey, 1999);. In the context of freeway networks, the mathematical model considers a freeway as made up of individual units of study on space from which crash frequency and influencing factors are observed and associated using an appropriate model. In order to quantify an individual factor contribution through a mathematical model, it has to be observed and measured either quantitatively or qualitatively on an appropriate cardinal scale. It further should exhibit variability in order to account for the observed variability within the crash frequency. When these conditions are met, a mathematical model employed can appropriately describe the safety level in terms of variability in crash frequency resulting from observed heterogeneity in influencing factors.

However, in practice, not all relevant factors are observed either in a pre-crash or post-crash environment. For instance, Chen, H. *et al* (2010) postulated the following safety performance function for evaluating left-side off-ramp at freeway diverge area:

$$Y_1 = (X_1)^{0.5060}(X_2)^{0.7412}\exp(-1.7518 + 0.1817X_3 + 0.8401X_4 - 0.7575X_5$$

where $X_2\ to\ X_5$ respectively represent mainline AADT, ramp AADT, ramp location (left-side or right side), length of deceleration lane, and ramp length. As it can be noted from the equation, other factors particularly environmental and human factors were not included. Shankar, V *et al*

(1996) showed that environmental factors such as extent of snow have an impact on the crash occurrence. Being fatigued significantly increases the risk of a crash. It makes a driver less aware of what is happening on the road and impairs his/her ability to respond quickly and safely if a dangerous situation arises. When a crash influencing factor cannot be observed and measured appropriately, it leads to unobserved heterogeneity since it omits important information in explaining the observed variability in crash frequency. A mathematical formulation which does not include all relevant factors leads to biased and inconsistent weights which quantify an individual factor contribution towards the overall safety of the network. Furthermore, decision makers are likely to arrive at incorrect decisions of improving safety levels through the use of countermeasures since the results are based on incorrect underlying factor interactions.

Another form of dependence occurs when units of analysis are clustered. A mathematical model which incorporates clustered freeway sections induces dependence because any two randomly selected sections in the same cluster tend to be alike compared to units selected from different freeway sections. If clustering is ignored in a mathematical model, inferential statistics are wrong because the standard errors of the weights are generally underestimated. This may wrongly lead a research to infer that an influencing factor has a real effect on the outcome when in fact the effect could be ascribed by chance.

In addition to the effect of unobserved heterogeneity in analyzing safety, homogeneity assumption of influencing factors is restrictive. Parameters that describe a factor’s effect apply to all categories of interest such as segment lengths or the type of a segment (Figures 2 & 3). Figure 2 shows two sites with different segment length while in Figure 2, the sites are differentiated by the ramp configuration: EX-EN as type 1 and EN-EX as type 2 where EX stands for exit and EN for entrance. In these cases an impact of an influencing factor is assumed to be constant for all the sites under study. Past research has identified that long segments in freeways experience few crash frequency in any observed period of study and crash events occurrence in weaving segments is different compared to non-weaving segments. However, for two segments with differential lengths, the homogeneity assumption characterizes the impact of these lengths to be the same based on the assumption that other factors influencing a crash are held constant. These results obscure the ultimate goals of safety analysis which aims to identify locations with higher level of impact on crash frequency in order to design appropriate countermeasures.

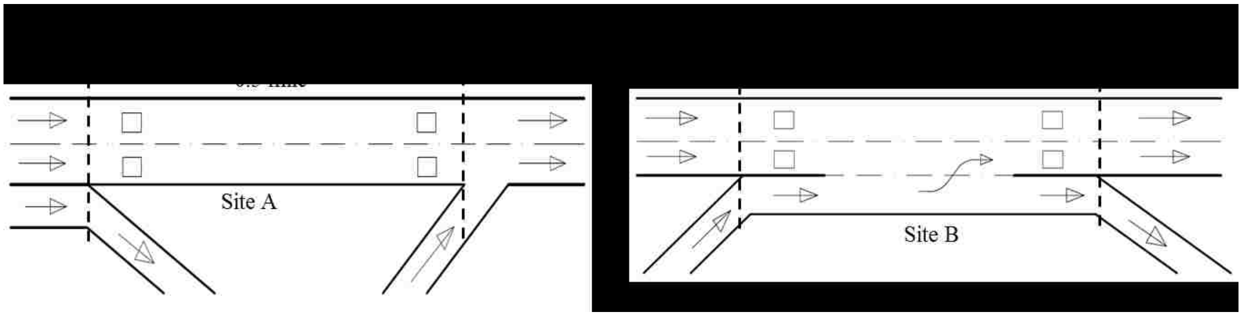

Figure 2: Two sites with different lengths

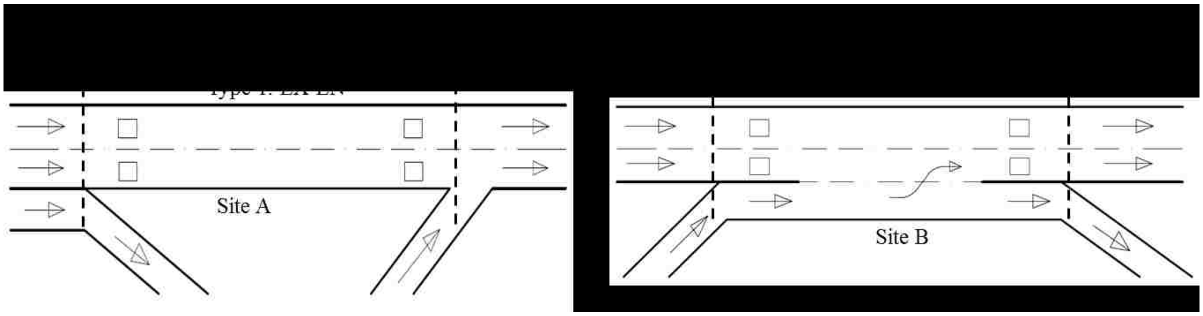

Figure 3: Two sites with different ramp configurations (EX-EN Vs EN-EX)

Spatial effects are likely to be experienced in the analysis process due to the special nature of crash frequency distribution. Spatial effects manifests through spatial autocorrelation in crash frequency across contiguous freeway segments. Spatial autocorrelation is the presence of spatial pattern in crash frequency over space due to geographical proximity (Ramosa, 2013; Li, Zhu , & Sui, 2007; Aguero-Valverde, 2014; Black & Thomas, 1998; Wang & Kockelman, 2013). When the distribution of crash frequency over space form patches of clusters such that high values of crash frequency tend to cluster together in adjacent freeway segments or high crash frequencies are contiguous with low values of crash frequencies then safety analysis techniques based on the distribution assumption of independence of crash frequency is violated. In this case the ultimate results quantifying the safety levels are biased.

Addressing the aforementioned issues require the use complex data structures and statistical models to be able to address sources of variability in crash frequency at a micro-level sections along the road network. This study proposes analysis safety on freeways to address these issues at a micro-level of data structures by dividing a freeway segment into small sections in space and capturing traffic characteristics associated with crash frequency from those sections. Due to variability across segments, the study further incorporates contextual variables at the segments level such as geometric characteristics and randomizing the model intercept to be able to capture factors which are unobserved but are likely to contribute to crashes. Furthermore, spatially correlated crash frequencies are addressed by considering events arising from contiguous freeway segments based on the notion that each freeway segment is likely to affect or influence crash events generated form the abutting segments.

1.2 Research hypotheses

This research assumes the existence of dependence within crash counts across sites on freeway network. Correlation within clusters may be due to variation being induced by common unobserved cluster-specific factors. Ignoring cluster-effects increases the likelihood in drawing conclusion based on unrealistic inferences because estimator standard errors are likely to be underestimated and the usual conditional mean is no longer correctly specified. It further assumes that the distribution of crashes extends beyond the influence areas of the divergence and convergence segments as well as weaving segments. By including these areas, data within the weaving and non-weaving segments can be clustered to quantify the variability of unobserved factors through the variance of random parameters using multilevel count models.

Cross-sectional dependence may also arise when the crash counts have a spatial dimension due to contiguity of units of observations. Based on spatial phenomenon, it is assumed further that there exists spatial autocorrelation in crash frequency over space due to geographic proximity. When the distribution of crash frequency over space form patches of clusters such that high values of crash frequency tend to cluster together in adjacent freeway segments or high crash frequencies are contiguous with low values of crash frequencies then safety analysis techniques based on the distributional assumption of independence of crash frequency is violated.

1.3 Objectives

Objective #1: To analyze safety of freeways using geometric and traffic characteristics while accounting for the effect of unobserved characteristics

To address the problem of unobserved characteristics, the observed crash counts are considered as clusters embedded within the segments and therefore analytical approach considers the possibility of dependence within clustered crash counts as a result of unobserved cluster-specific factors leading to correlations (Karlaftis, G.M., and Tarko, P.A.1998; Abdel-Aty, M., and Huang, H. 2010; Lord, D., and Mannering, F. 2010). Based on this data structure the impacts of geometric and traffic characteristics on observed crash frequency on freeway segments will be analyzed while controlling for the effects of unobserved influencing factors and cluster-effects by comparing four count models: two are standard Poisson and negative binomial regression models which do not account for cluster-effects while the other two are Mixed-effects Poisson and negative binomial regression models which in addition to fixed-effects parts they account for the effects of randomness arising from heterogeneity and clustering.

Defining the segments as stretches between their ramps locations allows the analysis to include crashes beyond the influence areas because the distribution of crashes in space is not limited to only the influence areas of the divergence and convergence segments as well as weaving segments. To this point data are clustered within the weaving and non-weaving segments to quantify the variability of unobserved factors through the variance of random parameters using multilevel count models. The mixed-effects models are estimated by Gauss-Hermite quadrature method because the distributions assumed for unobserved heterogeneity and cluster-effects are different and these have to be integrated out of the conditional mean.

Objective #2: To analyze the existence of spatial autocorrelation on contiguous freeway segments with ramps as natural delineators while controlling for traffic and geometric characteristics observed.

Under this objective, two count models will be used: Non-spatial and Spatial Poisson models (Lee D. 2011 & 2014). This analysis is based on the results to be obtained in the previous section which aims to identify the existence of unobserved heterogeneity in freeway segments causing crash frequency variation unaccounted for by traffic and geometric characteristics. It should be noted that the existence of spatial autocorrelation is an indication of the presence of unobserved factors unaccounted for which are manifested through the residual spatial autocorrelation. To this point the hypothesis is that crash frequency observed in contiguous freeway segments exhibit spatial phenomenon leading to spatial autocorrelation and a value of the spatial autocorrelation parameter significantly away from 0 is an indication of the existence of spatial phenomenon across adjacent segments.

This study investigates the existence of spatial autocorrelation in crash frequency across contagious freeway segments. Spatial autocorrelation is the presence of spatial pattern in crash frequency over space due to geographic proximity. When the distribution of crash frequency over space form patches of clusters such that high values of crash frequency tend to cluster together in adjacent freeway segments or high crash frequencies are contiguous with low values of crash frequencies then safety analysis techniques based on the distributional assumption of independence of crash frequency is violated. However, Conditional autoregressive models are set up, in a Bayesian modeling framework, to include terms which help to identify and quantify residual spatial autocorrelation for neighboring observation units. Models which recognize the presence of spatial dependence helps to obtain unbiased estimates of parameters quantifying safety levels since the effects of spatial autocorrelation is accounted for in the modeling process. Figure 4 below shows types of inputs and expected outputs for both objectives.

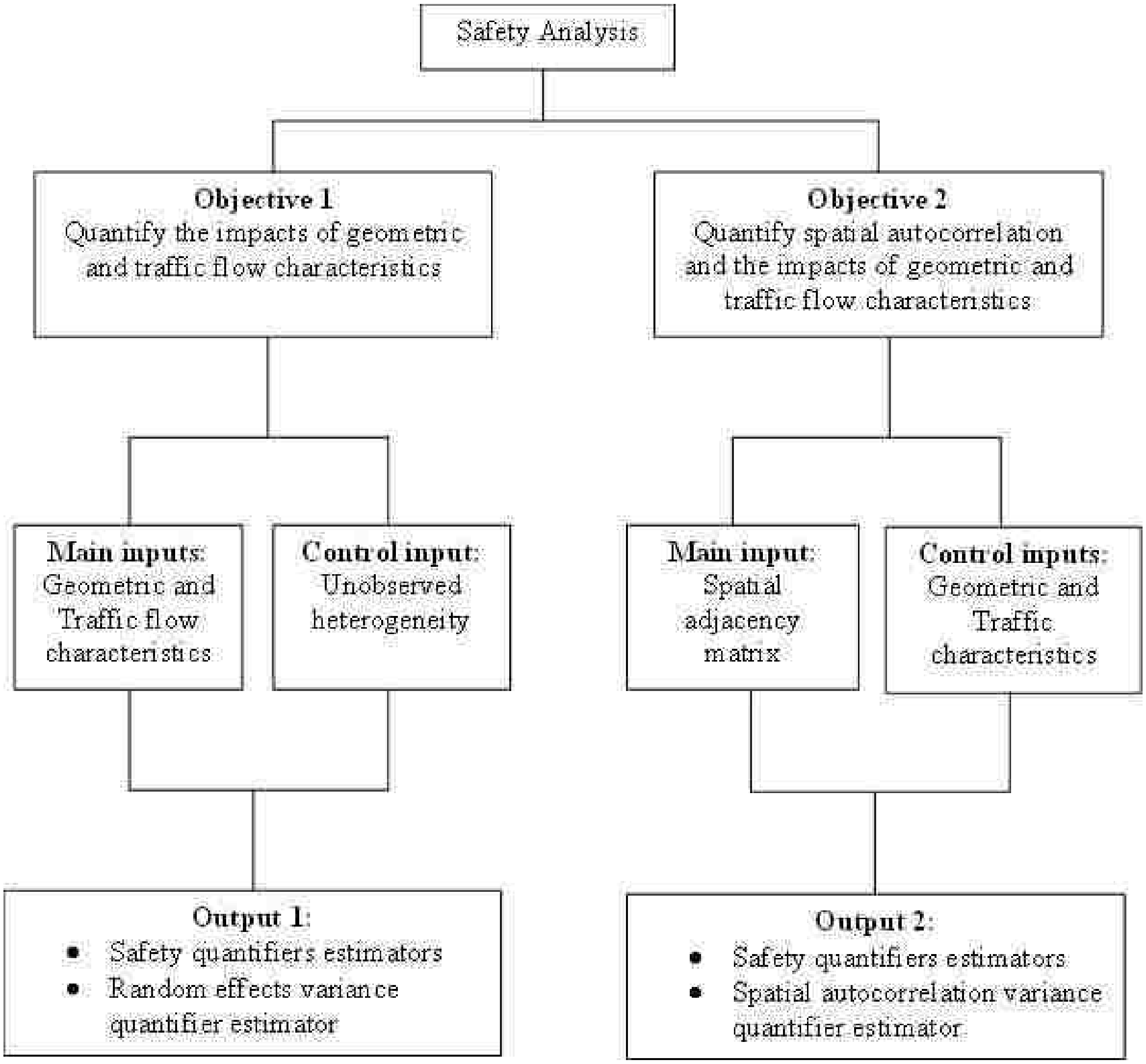
Safety Analysis
Objective 1
Quantify the impacts of geometric and traffic flow characteristics
Objective 2
Quantify spatial autocorrelation and the impacts of geometric and traffic flow characteristics
Main inputs:
Geometric and Traffic flow characteristics
Control input:
Unobserved heterogeneity
Main input:
Spatial adjacency matrix
Control inputs:
Geometric and Traffic characteristics
Output 1:
Safety quantifiers estimators
Random effects variance quantifier estimator
Output 2:
Safety quantifiers estimators
Spatial autocorrelation variance quantifier estimator

Figure 4: Inputs and outputs of study objectives

To appropriately accomplish the objects indicated on Figure 4 and obtained reliable estimators to address issues of unobserved heterogeneity and spatially correlated crash frequency, Figure 5 below displays research methodology to be focused on. Each objective as shown above will follow the same research methodology but with different statistical model setting as explained in detail in chapters 3 and 4 of the statistical methodology.

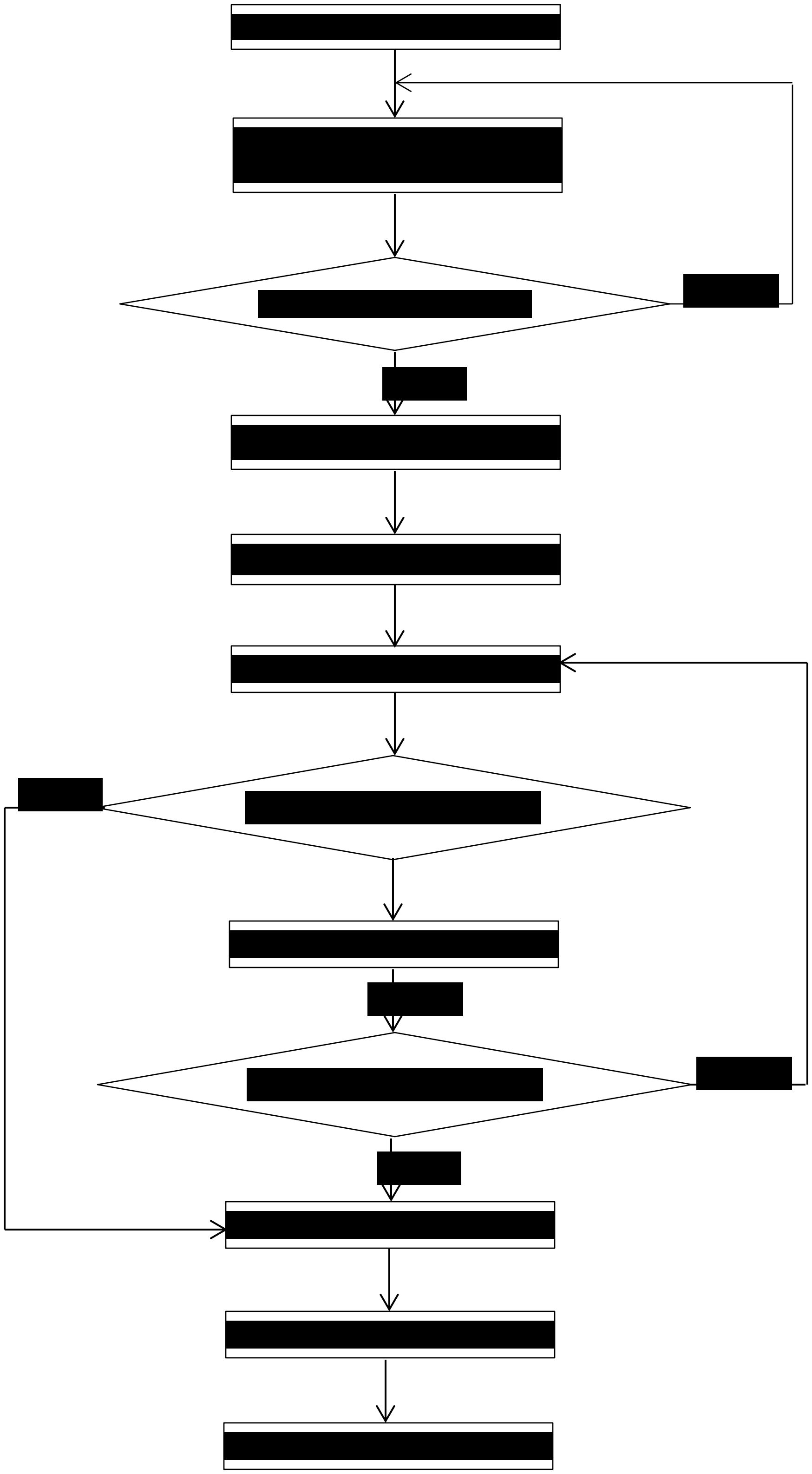

Figure 5: Flow chart of research approach

1.4 Study contributions

Limitations observed on the current literature suggest the need to address complex issues of unobserved heterogeneity and spatially correlated crashes by using complex data structures to help in understanding sources of variability and correlated crash frequency from abutting areas. This research differs from the previous research activities by using a different segmentation concept which addresses these issues at a micro-level data structure. Unobserved heterogeneity issues are addressed by dividing a freeway segments into small sections on which sensors are embedded and traffic characteristics can be observed easily and associated with crash frequency. Such setting helps to introduce more variability in crashes for successive sections within a freeway segment. An example of such a data structure is shown in Figure 6.

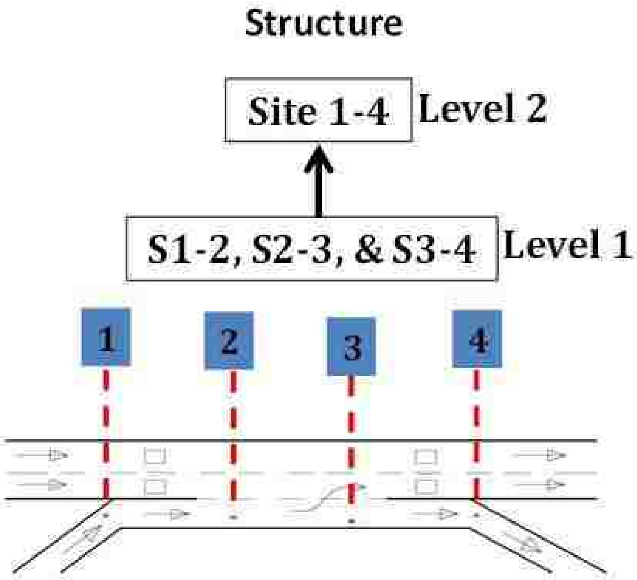

Figure 6: Proposed data structure to address unobserved heterogeneity

Figure 6 is a clear picture which shows that more variability is likely to be captured when a freeway segment is further subdivided into small sections. This implies that variations arising from each sub-section are likely to be captured. Furthermore, variability at the segment level is also possible by including context factors which vary at the segment level but are constant within a segment. These factors include geometric elements associated with a particular freeway segment. Since the model setting further requires the model intercept to vary across segments, more factors which are unknown or cannot be collected are also addressed. Current literature has address only among the sources of variability in crash frequency. This study addresses both variability by allowing slopes and intercept of the regression functions to vary across the freeway segments.

Issues of spatially correlated can be addressed by focusing on freeway segments sharing an arbitrary border as shown in Figure 7. Congestion levels on the network are likely to cause secondary crashes and in the event that the congestion levels are high spill over will occur due to shock waves on the segments behind another segment. Spill over may also occur when an incident a primary crash has occurred in the close proximity of a segments sharing border which in turn may lead to a secondary crash on another segment. Based on this concept and in addition to the primary crash occurrence, it is clear that each segment is likely to affect the other.

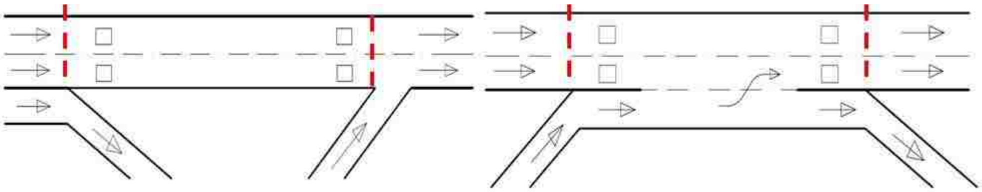

Figure 7: Contagious freeway segments to address spatially correlated crash frequency

In addition to the above contribution of the study in addressing the arising complex issues through the use of micro-level data structure, under objective 1, the model framework gives an engineer, the cluster of the actual segments with above average crash frequency. Spotting out these segments is a way of screening the network to find locations with need for improvement and efficient allocation of resources in the sense that improvement resource can be allocated to areas where there is a need.

The model can also be applied in the before-after studies in road safety of actual roadway sections where the geometric elements were changed. This study will consider the analysis period to be 2013 and therefore the findings provide an estimated model of what safety levels was if changes in geometric elements were made prior to the analysis period. This comprise an after study results. Data prior to the analysis period can be used to obtain a prediction model and through extrapolation, the model predicts what would have been the safety levels in the absence of changes in the geometric elements (Hauer, E. 1997). The difference of safety levels before and after studies indicates improvement.

The sign and direction of improvement is a function of specific changes made. For instance, an additional lane on the main facility between on-and off-ramps is likely to indicate better positive safety levels due to an added freeway capacity expected. Negative safety levels may results from narrowing a geometric element such as a shoulder for the purpose of adding high occupancy vehicle lanes. It is therefore important to conduct before-after studies to quantify actual levels of safety once changes are made in a network.

The model which addresses spatial effects can be applied as a discriminant model. This is based on the fact that spatial effects terms are included in the modeling process. Theories on the estimation process require integrating out these effects and summarize them in terms of variance, a method which leaves out the actual influences of the remaining factors. Based on the graphical plots to be produced a researcher can point out locations on the freeway network from which its factors exhibited more impacts on the crash frequency. Another important application is based on the natural interpretation of most of the regression coefficients. Negative coefficients in most of the cases mean the corresponding factors had a negative impact and therefore by increasing those factors help reduce crash frequencies on freeways.

1.5 Organization of the dissertation

This dissertation is comprised of eight chapters: chapter one describes motivational background, problem statement, research hypothesis as well as study objectives. Chapter two contains literature review while chapter three describes multilevel count model for clustered data where variant of traditional and hierarchical count models are detailed. Chapter four describes models designed for spatial data. Specifically Conditional Autoregressive Models (CAR models) are discussed in conjunction with generalized Poisson regression model. The CAR models account for spatial autocorrelation through a binary spatial weight matrix supplied by CAR models in a form of CAR priors. Chapter 5 describes data collection and input while chapters six and seven discuss model results. Chapter eight includes conclusions and recommendations.

CHAPTER 2 LITERATURE REVIEW

2.0 Non-spatial modeling

This section reviews safety modeling practices for non-spatial modeling of freeway segments. Non-spatial modeling entails only the mathematical association of crash frequency with geometric, traffic, environmental and human characteristics of a freeway segment without involving its spatial effect. In this case, the expected crash frequency of a segment is assumed to be fully described by only non-spatial characteristics which do not incorporate location characteristics such as inverse distance or spatial contiguity matrices. For clarity, a freeway segment – here referred to as an observational unit - is defined as the length of a roadway between two points on space over which traffic and physical characteristics remain the same (TRB, 2010). Three types of segments have been an area of research: weaving segments, freeways merge segments, and freeway diverge segments. Weaving segments are characterized by closely spaced merge and diverge areas as well as the presence of one movement crossing the path of another in the absence of signals. (Roess, Prassas, & McShane, 2011). Merge segments are formed when two traffic streams form a single stream while diverge segments are formed when a single traffic stream separates into two streams.

Geometrically, these segments are comprised of deceleration and acceleration lanes acting as speed change lanes, the number of through lanes to facilitate through movements, on- and off-ramps for entering and exiting vehicles respectively. In addition to these geometric elements, segments are also differentiated when traffic operation is such that the terminals operate independently (AASHTO, 2011).

Based on the fact that traffic volumes along the freeway segment vary, the manual provides guidelines which requires that consistency has to be maintained in the basic number of lanes designated over a given length of a route. This is independent of the changes in the levels of traffic volumes and lane balance needs. According to the lane balance principle, the number of lanes beyond the merging point of the entrance should not be less than the sum of traffic lanes on the merging roadways. At exits, the number of lanes on the main facility should be equal to the number of lanes on the exit minus one. An example of lane balance principle can be seen on Figure 8 below for a diverge segment. At location "A' there are five lanes including the auxiliary lanes and the sum of lanes at locations "B" and "C" minus one equals five. According to the lane balance principle, the segment is lane balanced.

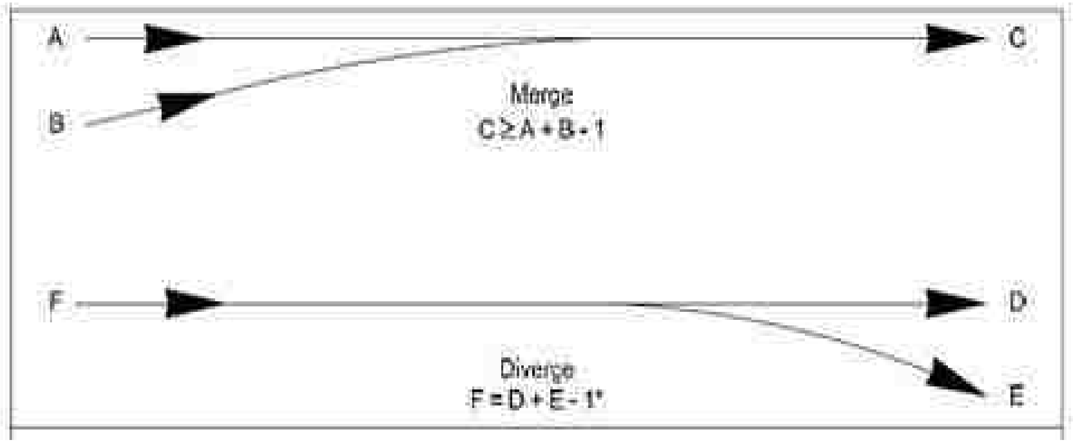

Figure 8: Lane - balance principle at exit
(Source: http://www.wsdot.wa.gov/publications/manuals/fulltext/M22-01/1360.pdf)

2.0.1 Observed heterogeneity counts modeling

Geometric design elements have been an active area of research due to their influence on crash occurrence (Garber & Hoel, 2009). The levels of influence of these elements are quantified by associating crash frequency, a measure of the number of crashes experienced on the site of interest per period, with numerical values measured on the influencing factors. For instance, the study done by (Liu, Chen, Lu, & Cao, 2010) employed a generalized linear model with log link to investigate how lane arrangements on freeway mainlines and ramps affect safety of freeways with closely spaced entrance and exit ramps based on consistency of basic number of lanes and lane balance principles. Their study area involve a stretch of 1000ft upstream of the merging area at the entrance and downstream area of the gore are including a distance between entrance and exit.

Furthermore, (Liu, *et al*, 2010) identified a total of seven different lane arrangements which included (1) segments with continuous auxiliary lanes between segment terminals where segment coded as type B had an auxiliary lanes ended up with a two lane exit and those identified as type C had auxiliary lanes ended with one lane exit, (2) Type A segments with one lane entrance followed by one lane exit from which some of the segments had a tapered type at type entrance and exit while others had parallel type entrance and tapered type exit, (3) Segments with one lane entrance and two exit where the outer lane is dropped at the exit gore and a taper is provided at the exit, (4) segments with two lane entrance followed by two-lane exit with lane drop for one type and the other with a lane drop for only the outer lane. Based on their main characteristic of lane balance coordination and basic number of lanes, the research team found that segments with continuous auxiliary lanes connecting closely spaced entrance and exits ramps, and where the exit ramp ends with two lanes (designated as Type B), had the highest average crash frequency, crash rate, and percentage of fatal plus severe injury crashes. They also found that more lanes, ramp AADT and posted speed limit increased crash frequency.

The study done by (Sarhan, Hassan, & Adb El Halim, 2008) used Poisson and negative binomial models to investigate safety performance of freeway sections in relation to speed-change lanes. With a focus on the influence of the characteristics of speed change lanes (acceleration and deceleration lanes) of merge and diverge areas as well as weaving segments on the freeway safety performance revealed that traffic volume and the number of travel lanes on the main facility increased crash frequency. The effect of the number of lanes implies that more lanes increases the number of lane changes which in turn is likely to increase crash frequency. Long lengths of acceleration and deceleration lanes decreased crash frequency because drivers can easily complete merging or diverging tasks. Their findings also indicated that segments with extended acceleration lanes increased crash frequency compared to segments with limited acceleration lanes. Extended acceleration lanes in this case are expected to have double function: (1) used as an acceleration lanes for vehicles merging, and (2) used as a deceleration lane for merging traffic.

(Chen, Zhou, & Liu, 2014), observed that short lengths of deceleration lanes on diverge areas has the highest counts because vehicles do not have enough space to reduce speeds smoothly and weave to the exit ramps. Optimal deceleration lane lengths between 500ft and 700ft were strongly recommended. Their segments consisted of exit sections with parallel and tapered design.

(Golob, Recker, & Alvarez, 2004), found that location where a crash has occurred as well as type of movement performed by the vehicles involved were the most significant factors influencing crash occurrence on weaving sections with sideswipe collisions having the highest likelihood of occurrence. Improvement in signage, lighting, pavement resurfacing, and enforcement on posted speed limits as well as implementation of changeable message signs warning of potential hazards have been recommended in their study.

(Chen, Liu, Lu, & Behzad, 2009) used generalized linear model with a log link and focused on diverge sections with an area spanning 1500ft upstream and 100ft downstream of painted nose and found that freeway and ramp AADT, posted speed limit on freeway, deceleration lane length, right shoulder width and the type of exit ramp significantly affected the safety performance of freeway diverge areas. In addition to freeway and ramp traffic demand as determinants of the number and arrangement of lanes on freeway exit ramps, this research added the benefit to quantify the impacts of different exit ramp types on the safety performance of freeway diverge areas.

(Joe, Greg, & Davey, 1999), focused on the ramps (entrance and exits) to estimate accident frequencies as a function of ramp and mainline traffic characteristics. Used the definition of speed-change lane as the length from the painted gore point to the end of the lane taper, their findings indicated that increasing speed-change lane reduced the number of crashes. The team also revealed that an increase in the annual average daily traffic of the mainline facility increased the number of crashes (Cirillo, 1970), worked on the same subject and found that lengthening acceleration lane is more beneficial compared to lengthening deceleration lane.

(Chen, Zhou, Zhao, & Hsu, 2011), concentrated on diverging areas by comparing off-ramps on the right-side of the main facility with those on the left-side of the facility based on ramp lane configurations. Count models were estimated with segmentation involving 1500ft upstream and 1000ft downstream of the diverge areas. Their results indicate that for one-lane exit ramps, length of deceleration lane and AADT on main facility increased the number of crashes. The left-side ramp increased crashes compared to the right-side ramp.

(Shankar, Mannering, & Barfield, 1995), used Poisson and Negative binomial models to associate geometric elements and weather factors with type of crash frequencies. A total of six models were estimated with type of crash counts comprised of sideswipe, rear-end, parked vehicle, fixed objects, overturn, and same direction accident frequencies. Their results indicated that both geometric and environmental factors have an impact on crash frequencies though the magnitude of impact is different across different types of crash frequency.

(Bonneson, Geedipaly, & Pratt, 2014), developed safety prediction methods for freeways and interchanges to address freeway segments and speed-change lanes safety by including crash modification factors (CMFs) that describe the observed relationship between crash frequency and freeway geometric and traffic characteristics. The research team found that crashes on curved freeway segments with shoulder rumble strips were more frequent than on curved segments without shoulder rumble strips.

The aforementioned research activities considered only the fixed part of the models from which the effects of geometric and traffic characteristics are considered constant across the sites. In this study, we extend further the analysis on the effects of geometric features by incorporating items which accounts for unobserved factors in count models to allow parameters to vary across segments in order to capture and quantify unobserved factors. This approach avoids biased coefficients in multilevel settings because the estimate of the standard errors will be correct.

2.0.2 Unobserved heterogeneity counts modeling

2.0.2.1 Random parameter counts modeling

(Venkataraman, *et al* 2011) employed random parameter negative binomial model to analyze crash counts on using data from Washington State’s Interstate system for investigating heterogeneity issues in the mean of slope regression coefficients. Both traffic and geometric characteristics of segments were used in the analysis including lighting type proportions by length, shoulder width proportions, lane cross-section proportions, number of vertical curves in a segment, the shortest horizontal curves in a segment length, the largest degree of curvature in a segment, the smallest vertical curve gradients, and the largest vertical curve gradients in a segment.

Their results indicated that curvature effects were found to be random which implies that the effects varied from one segment to the other. It was also found that the largest degrees of curvature, as well as the smallest and largest vertical curve gradient variables, were found to exhibit randomness. Furthermore, traffic characteristics expressed as the logarithm of average daily travel and the median and point of lighting proportions were found to exhibit randomness in influencing crash frequency.

(Anastasopoulos & Mannering, 2009) investigated factors that influence the frequency of crashes by using a random effect model in which an intercept is allowed to vary for accounting heterogeneity issues based on Poisson and negative binomial count data models. The research team used crash data collected on 322 roadway segments with homogeneous characteristics from rural interstate highways in Indiana. Pavement, geometric, and traffic characteristics were included in modeling process. Pavement characteristics included friction, international roughness index (IRI), average pavement condition rating (PCR), and rutting (good and excelling states). Geometric characteristics included segment length, median and right shoulders, average horizontal curve degree, number of vertical curves per mile, and ramp density while traffic characteristics constituted average daily traffic of passenger cars and average daily percent of combination trucks in a traffic stream. Their results indicated that random-parameters negative binomial outperformed Poisson models both in a fixed-effect and random-effects settings.

The decision rules to either apply fixed effects or random effect is applied is based on the significance of the standard deviation of the parameter density. If it is found to be different from zero, then the parameter is fixed across the population of roadway segments. Based on this rule, it was found that international roughness index (IRI), rutting indicator variable, roadway segment length, presence of median barrier, interior shoulders, horizontal curves degree of curvature, and average annual daily traffic (AADT) exhibited randomness characteristics with standard deviation of the parameters distribution being different from zero.

(Dinu & Veeraragavan, 2011) employed random parameter model approach for predicting crashes on two-lane undivided rural highways in India. Highway segments were divided into homogeneous sections based on traffic volumes, carriageway width, and shoulder width and from these sections geometric and environmental characteristics were collected and the modeling approach accounted for the difference in traffic volume levels and composition by separating day-time and night time crashes. Their results showed that predictors related to traffic composition had standard deviation significantly different from zero when modeled as random parameters. This implies that the associated characteristics exhibit heterogeneity behavior in influencing crash occurrence.

With respect to the day-time crash frequency model, it was shown that hourly traffic volume, length of highway segment, proportion of cars and motorized two-wheelers in traffic, driveway density, width of shoulder and horizontal and vertical curvature were found to significantly influence crash occurrence frequencies while the night-time model indicated that hourly traffic volume, length of highway segment, proportion of buses, cars, and trucks, driveway density, and vertical curvature were found to be significant. However, proportion of cars and tracks decreased crash frequency with remaining factors leading to an increase in crash frequency.

Furthermore, their results indicated that traffic, vehicle, and geometric characteristics exhibited randomness in influencing crash frequency. These results reflect the existence of variability associated with different characteristics. In particular it was observed that different vehicle types may likely lead to randomness in traffic volume and composition while driver behavior is likely to influence variability observed in road geometric characteristics and access factors such as driveway density.

(Garnowski & Manner, 2011) also extended the traditional count model by allowing random parameters across observations to investigate factors related to crashes on German highway clover leaves and diamond interchanges while accounting for the possible existence of parameter heterogeneity. Their study included crash data, traffic and geometric characteristics observed on the aforementioned highway connectors. Ramp length, length of the deceleration lane, total width of the lanes on ramps, width of the shoulder lanes, curve angles and their lengths, as well as number of lanes both on ramp and main facility constituted geometric characteristics. Traffic characteristics included separate flow in terms of average daily traffic of vehicular mix and total flow for all vehicle mix combined. Their results showed that negative binomial model with random coefficients outperformed the Poisson base model. In addition to model selection between competing models, it was revealed that average daily traffic of passenger cars, absolute total deflection angle, truck percentage, and steeper curves were statistically significant. With respect to random coefficients it was observed that steeper curves, length of the deceleration lane, and position of the steepest curve exhibited randomness.

In addition to the aforementioned techniques of investigating heterogeneity issues by randomizing regression coefficients, grouping or clustering information had been used in safety research activities to reduce heterogeneity through the selection of homogeneous data to be analyzed. (Wong & Chung, 2008) used clustering and classification approaches to analyze heterogeneous crash data from the perspective of crash occurrence; the aim being to compare group-specific characteristics and to examine the observed heterogeneity among crash groups. Grouping rules based on set theory and logistic regression model techniques applied on driver, trip, behavior and environmental characteristics as well as crashes.

Their findings revealed the existence of heterogeneity across distinct features associated with groups. In addition to this finding, it was also observed that driver characteristics such as age, gender and occupations were statistically significant while trip purpose and time were statistically significance with respect to trip characteristics. Statistical significance with respect to behavior and environmental characteristics were found in speed limit, road shape, sight distance and obstruction, weather, lighting, traffic control mechanisms as well as cell phone use and drinking condition of drivers.

Median crossover crashes also have been an area of active research. (Shankar, Albin, Milton, & Mannering, 1998) evaluated median crossover likelihoods based on clustered crash counts in Washington state using random effects negative binomial as well as cross-sectional negative binomial model with the formal outperformed the latter model with respect to unobserved spatial and temporal effects. Random effects negative binomial assumed that location-specific effects with respect to overdispersion are randomly distributed across groups. Their study constituted sections without median barriers on divided state highways and characteristics such as the number of crashes, traffic and geometric characteristics were observed.

Geometric characteristics included length, shoulder and median widths, total number of facility lanes, grades as well as number of curves. Traffic characteristics included average daily traffic and speed limit while vehicle characteristics constituted single-and double-unit truck percentages as well as total truck percentages. Their results indicated that location and time – specific variables with respect to random effects negative binomial model were significant. Accounting individual factors, it was observed that average daily traffic had a negative impact on crash occurrence which implies a counterintuitive effect, a notion which reflects the existence of negative association with other influencing factors. Furthermore, median width had a positive impact which reflects the existence of narrow median width and flatter median fore slopes for wide medians. Other factors such as the interaction of shoulder width and friction variables on horizontal curves positively affected median cross over frequencies.

2.0.2.2 Multilevel counts modeling

Multilevel modeling approaches, also known as hierarchical models have been given much attention in traffic safety research for its ability to account for unobserved heterogeneity (Huang & Abdel-Aty, 2010) proposed the use of multilevel framework in a different setting of data structures related to traffic safety as a function of clustering processes. The research team revealed five different levels from which traffic safety data structures can be observed and the levels ranged from geographic region level to occupant level with the intermediate levels constitute traffic site, traffic crash, and driver-vehicle unit levels.

The aforementioned levels of data structure were demonstrated in their study based on real-word datasets. To insist the existence of levels, this study cited two examples from their study related to both crash frequency and crash severity. Related to crash frequency, the research team indicated possible existence of levels involving nested groups where county is at the highest level representing geographic region level. Within the geographic level (county level) it was observed that corridor components of the roadway facility constitute the next level nested into the county level and at the lowest level intersection characteristics are observed. With respect to crash severity, example of possible levels observed included grouping of characteristics at the road segment and traffic crash levels with the formal level being at the highest level of the hierarchy.

In the severity level model, average daily traffic and driver age characteristics were observed to explain differential severity outcomes while accounting for the possible existence of segment-specific random effects as a function of cross-road segment heterogeneity. Segment level covariates were also included to determine variable threshold values of the severity outcomes. Their results indicated that segment-specific random effects exist as a function of omitted confounding factors associated with road segments. Furthermore, it was observed that the proposed multilevel crash severity prediction model was better compared to ordinary ordered model.

With respect to the crash frequency data structure, it was pointed out that crashes that occurred within a same county are likely to exhibit dependence when compared to the observed crash frequency at the corridor level while intersections within the same corridor may exhibit correlated characteristics due to omitted corridor-specific characteristics. In their study, random effects were investigated at the corridor level by specifying corridor-specific heterogeneity. The results revealed the existence of within-corridor correlation.

(Yu, Abdel-Aty, & Ahmed, 2013) employed Bayesian hierarchical models to analyze the effect of mountainous freeway hazardous factors on crash frequencies from Colorado state with a focus on season differential impacts on safety levels. Based on the fixed effects model, over-dispersed Poisson model with no correlation and the over-dispersed correlated Poisson model, two types of models were investigated: (1) seasonal model and (2) single-vehicle Vs multi-vehicle crash models. To account for unobserved heterogeneity, two random effects, segment season specific random effect and segment only specific random effect, were included in modeling process.

Furthermore, to account for the effect of observed heterogeneity, traffic and geometric characteristics as well as weather factors were included in modeling. Seasonal model included daily vehicle mile traveled, grades, and average speed for the crash segment as well as visibility and temperature during crashes. For single-vehicle and multi-vehicle crash model, precipitation, speed and occupancy factors were included. With respect to seasonal model, their results indicated that significant season effect exists. Specifically, it was observed that steeper slopes experience a higher crash frequency and the upgrade segments are safer compared to downgrade segments. Also it was observed that vehicle mile of travel and precipitation volumes had a positive impact on crash frequency while visibility conditions decreased crash frequency.

Single-vehicle Vs multi-vehicle crash model also revealed the existence of factors influencing crash occurrence. In this model, it was observed that crashes are more likely to happen at the segments with a sudden high precipitation, lower speed at the crash segment and higher occupancy at the upstream segment 5-10 minutes before the crash time increased the likelihood of crashes. It was further observed that single-vehicle is more influenced by other compared to upstream and downstream traffic status. With respect to multi-vehicle, it was evident that traffic variables influence crash occurrence and their factors are more associated with congestion levels.

2.0.3 Summary

The aforementioned review on non-spatial modeling painted unique differences across models which do not account for the existence of unobserved heterogeneity compared to those which accounts for unobserved heterogeneity. Observed heterogeneity can be modeled by counts models based on the assumption that crash distributions across observational units are independent and heterogeneity across crash frequency can be capture by a gamma-distributed random effects which accounts for the existence of over-dispersion across crash counts. The use also assumes that variability within the observed crash counts can also be accounted by only the observed covariates used. However, these models are limited by the fact that when omitted variable cases arise, then the models are likely to encounter heterogeneity issues leading to unobserved heterogeneity. Under these models, it is also assumed that the impact of influencing factors is the same across the site under investigation, a factor which may not be true in real world.

To account for the aforementioned limitations, researchers tried to build in randomness in the modeling safety on the highways whereby a regression coefficient has random effects in addition to its fixed effects. This implies that differential impact expected within the same factor can be realized when the standard deviation of the distribution of a regression coefficient is significantly away from zero. By allowing some or all of the parameters to vary, heterogeneity is accounted for and the biasness and inconsistency of the estimates are likely to be avoided and therefore inferential statistics are reliable.

Further extension from randomizing slopes is to allow the intercept to vary across the observed units of analysis. This has proved to be of great importance when data structures are such that groups exists and have nesting structures leading to a multilevel or hierarchical structure. In this regard multilevel counts models have been applied to capture the existence of unobserved heterogeneity at higher levels which are likely to influence the outcomes at the lowest levels of groups. However, groupings which lead to different levels depend on different research purposes and introduce levels of dependences within the outcomes observed in the same group. This means unobserved heterogeneity can be observed at different higher levels involved in the formation of data structure.

2.1 Spatial modeling

Crash events occur spatially along the highway network and including spatial effects in crash prediction models help to explain variability observed in crash frequency and avoid making inference on biased estimates. (Black & Thomas, 1998) employed a network autocorrelation analysis to examine accident distributed along the segments of a highway system and found a significant level of positive spatial autocorrelation. (El-Basyouny & Sayed, 2009) used Gaussian conditional autoregressive and multiple membership models on 281 urban segments in Vancouver, Canada and found that spatial autocorrelation across urban segments explained approximately 87.6% variability in crash rates for CAR model while it was approximately 98.5% for multiple membership models. In addition to these findings, it was also revealed that AADT, business land use, number of lanes between signals and density of unsignalized intersections had significant positive impact on the number of crashes.

(Guo, Wang, & Abdel-Aty, 2010) employed Bayesian count and Gaussian models to incorporate corridor-level and intersection proximity spatial autocorrelations in predicting crash rate and crash frequency and it was revealed that the size of an intersection, traffic conditions for both through and turning movements and the coordination of signal phase have significant impacts on intersection safety. This implies that closeness of coordinated intersections is likely to stimulate differential driving behavior compared to isolated intersections. (Ozbay & Yanmaz-Tuzel, 2010), developed models to investigate the impacts of increase in lane width, installation of median barriers, and vertical and horizontal improvements in the road alignment on crash rates. The research team revealed that the use of random effects and hierarchical model structures explained better spatial and temporal effects in the crash rate. It was further shown that individual crash reduction factors indicated a decrease in crash rates after the specific treatment is applied. In case of limited data availability, it was shown that Bayesian models with spatial structures are likely to reduce biasness in model parameters.

(Arthur, 2015), identified the existence of spatial autocorrelation based on Moran's I statistics applied on neighboring network intersections. To be able to apply the concepts of spatial autocorrelation on intersections as opposed to network roadway segments, the analysis considered the roadways as links and the intersections and the adjusted frequencies of collisions as areas. The Moran's I statistics values indicated the existence of clusters of collision frequencies while graphing these values identified a temporal fluctuation that follows a diurnal pattern which indicates clustering patterns. It was also revealed that daytime pattern suggests a high frequency of collisions on major arteries during the day, especially over rush hour where it would be reasonable to assume a more clustered pattern.

(Wang & Kockelman, 2013), used Poisson-based multivariate conditional autoregressive (CAR) models estimated by Bayesian Markov Chain Monte Carlo methods to examine the relationship between pedestrian crash counts across tracts areas and various attributes characterizing the network, land use and demography. The results indicated the existence of positive spatial autocorrelation across neighborhoods as a result of the existence of latent heterogeneity or missing variables that trend in space which are likely to generate spatial clustering of crash counts. In addition to spatial autocorrelation identification, their results also showed that there is a greater association of residences and commercial land uses with pedestrian crash risk across different severity levels due to high potential conflicts between pedestrian and vehicle movements.

(Miaou & Song, 2005), used a multivariate spatial Generalized Linear Mixed Model (GLMM) to model crashes by injury severity type simultaneously and to rank sites by crash cost rate as decision parameter in ranking. Ranking results were based on relative standards which imply that rank and select among a predetermined group of sites based on their relative risk levels. The results showed including spatial effects components in modeling processes improved the overall goodness-of-fit performance of the model and affected the ranking results for site improvements. Ramos, 2014 used a simultaneous equation modeling for crash rate of freeway segments to account for unobserved variables. Spatial effects have also been investigated in connection to travel demand models (Kwigizile, 2007). The results further revealed that including CAR model in modeling process accounts for the degree of overdispersion.

2.1.0 Summary

The aforementioned review shows that crash frequency exhibit autocorrelation in space which in this case is called spatial autocorrelation (also known as second degree spatial effects). A number of research activities have been conducted to investigate the existence of spatially correlated crash frequency on arterials on particular for closely spaced signalized intersections and along areas located between intersections known as midblock.

However, other areas along the road network are likely to exhibit spatial autocorrelation. One of these areas has been observed for freeway segments which share an arbitrary border. This is because there shared unobserved heterogeneity which are likely to propagate from one area to the next and may influence crash occurrence in the same trend. This implies that segments sharing border exhibit spatial autocorrelation property in such a way high values of crash frequency are found near other values of crash frequency leading to positive spatial autocorrelation or high values found near low values leading to negative spatial autocorrelation.

This study investigates the existence of spatial autocorrelation in crash frequency for abutting freeway segments and builds a spatial model which incorporates spatial effects terms. This is important in safety prediction models which assume the existence of spatial dependence in the response variables and which intends to obtain unbiased and consistence estimators.

CHAPTER 3 METHODOLOGY: MULTILEVEL COUNT MODEL

3.0 Introduction

The term multilevel is used to refer to analysis models for hierarchically structured data with variables defined at all levels of the hierarchy (Schnabel, Little, & Baumert, 2000); Raudenbush & Bryk, 2002). It offers a method of decomposing various sources of variability in the response variable. Due to the hierarchical nature and the existence of levels resulting from clustering (grouping) of individuals at higher levels, multilevel models are appropriate when researcher interests include the requirement to decompose various sources of variability in the response.

As aforementioned, clustered data arise when units of observations form groups (Dobson, 2002; Faraway, 2006; Scott, Simonoff, & Marx, 2013; Snijders & Bosker, 1999; West, Welch, & Gatecki, 2007; University of Bristol, 2015). Each group is known as a cluster (West *et al.* 2007) and observations within each cluster are known as units of analysis. Models designed to analyze such data structure are called multilevel models because the units of analysis forms the lowest level of the hierarchical structure termed as level 1 and successive groups forms the higher levels. For clarity and referring to the context of this research, consider Figure 9 below which represents a freeway segment as a stretch between ramps and within such a segment, sensors are located specifically for counting the number of vehicles within two different subsections: subsection 1-2 (for sensor A) and subsection 2-3 (for sensor B).

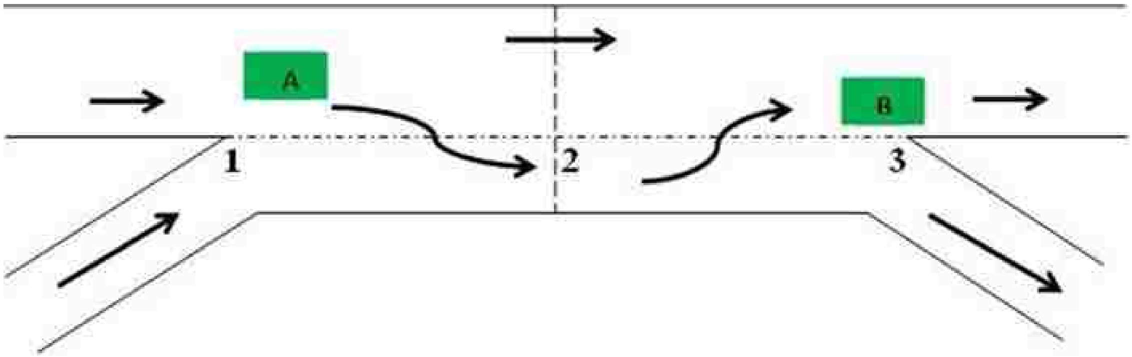

Figure 9: Multilevel structure within a freeway segment

In Figure 9, through vehicles are counted at sensor A and entering vehicles at section 1 are counted by sensor B in addition to through vehicles approaching the exit and gore areas. The two subsections 1-2 and 2-3 are located within a freeway segment between entrance at point 1 and exit at point 3. In the context of multilevel structure, the subsections form units of analysis and the freeway segments between points 1 and 2 forms a cluster. At the subsections, traffic characteristics and crashes are observed. The traffic characteristics are total traffic volumes and average vehicular speeds. At the cluster level (level 2) human factors and freeway geometric elements characteristics are observed including number of lanes, right shoulder, median shoulder and segment length. It should be noted that, with the exception of the total number of lanes, other geometric elements can be characteristic of the level 1 because they can be observed at the subsection level.

In multilevel language, characteristics formed at level 2 are called contextual variables and their effects on crash frequency observed at level 1 are called contextual effects (University of Bristol, 2015). For a freeway facility, it is expected that there are more segments and in each segment, two or three subsections are formed and sensors are spatially distributed across these subsections for counting volumes. In terms of segments, it is obvious that crashes which occur in the same segment are likely to have been generated by the same chain of causes attributed to local conditions such as intensity of lane changes as indicated by a weaving segment in Figure 9, segment overall geometry, and pavement rate of deterioration. However, depending on the type of a segment (weaving Vs non-weaving segment), crash occurrence across these segments are likely to be attributed to different causes as a function of local conditions in a particular segment.

The idea behind multilevel modeling is to aid a researcher in assessing the effects of higher level characteristics on the intercepts and coefficients at the lowest levels (Garson, 2013). This is possible by controlling for the effect of dependence on inferential statistics and at the same time be able to account for the cluster-specific effects. My focus in this study looks at the level 2 characteristics (geometric characteristics) and their effects on crash frequency while accounting for the cluster-specific effects. A unit diagram which represents the structure shown by Figure 10 is given below:

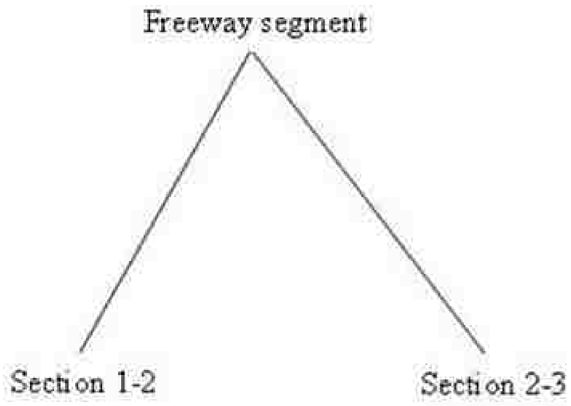

Figure 10: Unit diagram for the underlying structure of Figure 9

Information contained in a unit diagram can be expressed in a multilevel equation set up as follows:

$$E(y_{ij}|x_{ij}, U_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} \ldots + U_{oj} + U_{1j} x_{1ij} \quad (1)$$

From equation (1), β_0 is the overall average of cash frequency across all groups. The average crash frequency for cluster j (or segment j) is given as $\beta_0 + U_{oj}$ where U_{oj} denotes the cluster-effect residuals also known random effects (Scott, M.A *et al*, 2013). x_{2j} denote geometric element characteristics at the cluster level and β_2 is the effect of such characteristics on the average crash frequency. In my introduction, speficically in Figures 1 and 2, I indicated that two segments are likely to have different impacts on crash event occurrency due to differential characteristics which are local to the specific segments. Therefore, equation (1) captures such effect by allowing level 1 characteristics to vary across segments. This is shown by the term $U_{1j}x_{1ij}$ where in addition to the fixed effect β_1 of the variable x_{1ij}, the same variable is allowed to vary and its variance is captured by the random effect U_{1j}. Furthermore, equation (1) indicates that between-cluster variance varies as a function the variable with a random coefficient. This implies level 1 variance can be expressed in terms of variable x_{1ij} as follows:

$$var\left(U_{0j} + U_{1j}x_{ij}\right) = \sigma^2{}_{u0} + 2\sigma_{uo1}x_{ij} + \sigma^2{}_{u1}x^2{}_{ij} \qquad (2)$$

where σ_{u0} and σ_{u1} are the respective variances of the random effects U_{01} and U_{1j} and σ_{u01} is the covariance between random effects. Random effects are assumed to be normally distributed with mean zero and variances as shown above. Since they have a distribution and represent unobserved cluster-specific, they are estimated along with the fixed effects parameters.

The above set up introduces a detailed concept, in a single equation set up, behind multilevel models. Specific cases for count models in particular Poisson and Negative binomial regression models have extra terms which captures the effects of overdispersion in addition to the above concepts. Section 3.1 specifically deals with these models in a multilevel setting and sections 3.2 and 3.3 describe their specification and estimation in a matrix format.

3.1 Multilevel count model

3.1.1 General review

Occurrence of a crash has been determined to be a Bernoulli trial and a series of these trials can be described by the binomial distribution (Lord, Washington, and Ivan, 2005; Mulokozi & Teng, 2014). Researchers investigating the occurrence of crashes frequently use Poisson regression model (Navidi, 2011, Ross, 2010) because Poisson random variable may be used to approximate a binomial random variable. Based on Poisson model, the mean parameter is allowed to depend on regressors and such dependence is assumed to be parametrically exact with no other sources of random variation (Cameron & Trivedi, 2013; Agresti, 2002).

However, the Poisson process on which the standard model is based may fail due to the existence of unobserved heterogeneity across sites contributing additional randomness leading to the data exhibiting overdispersion (Mitra & Washington, 2007; Hauer, 2001; Nshankar & Sittikariya, 2014; Sittikariya, 2006; and Washington, Karlaftis, & Mannering, 2011). The extra-Poisson variation encountered due to unobserved heterogeneity is allowed by introducing in the model a multiplicative heterogeneity error term leading to mixed models such as negative binomial – a two parameter model flexible to account for extra-variation and which is obtained by integrating out the heterogeneity term in the Poisson-Gamma model.

When observed counts are considered as clusters, analytical approach should consider the possibility of dependence within clustered crash counts (Abdel-Aty & Huang, 2010; Karlaftis & Tarko, 1998) as well as the effects of higher levels on the regression coefficients at the lowest levels. Correlation within clusters may be due to variation being induced by common unobserved cluster-specific factors. Ignoring cluster-effects increases the likelihood in drawing conclusion based on unrealistic inferences because estimator standard errors are likely to be underestimated and the usual conditional mean is no longer correctly specified (Centre for Multilevel modeling, 2014; Cameron & Trivedi, 2013). Based on these concepts four count models are compared: two are standard Poisson and negative binomial regression models which do not account for cluster-effects while the other two are Mixed-effects Poisson and negative binomial regression models which in addition to fixed-effects parts account for the effects of randomness arising from heterogeneity and clustering. The mixed-effects models are estimated by Gauss-Hermite quadrature method because the distributions assumed for unobserved heterogeneity and cluster-effects are different and these have to be integrated out of the conditional mean (Stata, 2011).

3.1.2 Specification

The model specifies that y_i crashes for a freeway subsection i given $\boldsymbol{x}_i$ geometric and traffic characteristics are Poisson distributed with the likelihood function given as (Mulokozi *et al*, 2015):

$$f(y_i|\boldsymbol{x}_i) = \frac{e^{\mu_i}\mu^{y_i}{}_i}{y_i!} \tag{3}$$

The conditional mean crash frequencies is modeled as a function of the observed geometric and traffic characteristics known to influence the occurrence of crashes and it is given through the natural logarithm of the exponential mean function of the Poisson as:

$$\ln\big(E(y_i|\boldsymbol{x}_i)\big) = \ln(\mu_i) = \boldsymbol{x}'{}_i\boldsymbol{\beta} \tag{4}$$

The set of characteristics observed measures subsection heterogeneity in the mean crash frequency. However, still there is unexplained heterogeneity in the mean crash frequency because of unmeasured factors and to account for unobserved heterogeneity a multiplicative term

v_i is introduced in Model 4 (Cameron & Trivedi, 2013), (Veeraragavan & Dinu, 2011). Inclusion of a multiplicative unobserved heterogeneity results in the following function:

$$\ln\big(E(y_i|\boldsymbol{x}_i\,, v_i)\big) = \ln(\mu_i v_i) = \beta^* + x_{1i}'\boldsymbol{\beta} \tag{5}$$

where,

$$\beta^* = \beta_0 + \ln(v_i) \tag{6}$$

$v_i \sim \text{Gamma}\ (1, {\sigma_{v_i}}^2)$

From Equation 6 it is clear that a multiplicative heterogeneity is equivalent to a random intercept model, because v_i enters the model through the random intercept term. This implies that Model 5 comprises of an intercept which varies across independent crash counts and it can be interpreted as a one-level hierarchical model. However, our dataset has two levels. Level 1 has observational units as subsections defining the locations of sensors (FAST, 2014). Observed crash counts and their influencing geometric and traffic characteristics were recorded from these subsections. The subsections are nested within Level 2 which is made up of weaving and non-weaving segments treated as clusters. Our interest is to control for unobserved random effects of a segment type on the predicted average crash counts while controlling for geometric and traffic characteristics of the subsections. Therefore we include random cluster effects while controlling Level 1 factors and this is leading to a two-level hierarchical model for clustered data. For a series of M independent clusters (segments in this study) and conditional on the random cluster effects $\boldsymbol{U}_j$, a model with covariates at two levels can be written as (Diggle, Heagerty, Liang, & Zeg, 2002):

$$\ln\Big(E\big(y_{ij}\big|\boldsymbol{x}_{ij}\,, \boldsymbol{U}_j\big)\Big) = \ln(\mu_i) = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{U}_j \tag{7}$$

for subsections $i = 1, \dots \dots, n_j$, in segments $j = 1, \dots \dots \dots, M$. $\boldsymbol{X}_{ij}$ is the row vector of covariates for fixed effects with fixed effects regression coefficients $\boldsymbol{\beta}$; $\boldsymbol{Z}_{ij}$ is the row vector of covariates corresponding to the random effects and can be used to represent both random intercepts and random coefficients and y_{ij} represents the crash counts. The random effects U_j are multivariate normally distributed with mean **0** and variance-covariance matrix **Σ**.

The presence of unobserved heterogeneity term and cluster-random effects induce overdispersion in the conditional mean crash frequencies of the Poisson regression model leading to the variance exceeding the mean. Unlike Poisson regression model, negative binomial regression model is a two parameter model which is an alternative to overdispersed Poisson model since its variance is a function of the mean crash frequency. The negative binomial model assumes that crash counts are Poisson distributed and unobserved heterogeneity resulting from unobserved factors are gamma distributed. The distribution function of the negative binomial for cluster *j* , *j = 1,....., M* can be written as (Stata, 2011; Hilbe, 2011):

$$f\left(y_i|U_j\,,\alpha\right) = \frac{\Gamma(y_{ij}+r)}{\Gamma(y_{ij}+1)\Gamma(r)}\,{p^r}_{ij}\,(1-p_{ij})^{y_{ij}} \qquad (8)$$

$r = 1/\alpha\ , \quad p_{ij} = \frac{1}{1+\alpha\mu_{ij}}$

The conditional mean function for Model 8 is the same as indicated in Model 4. The marginal dispersion expressed as a function of both the dispersion parameters can be written as:

$$Var\left(y_{ij}\right) = \left[1+\{\exp(\sigma^2)\,(1+\alpha)-1\}E\left(y_{ij}\right)\right]E\left(y_{ij}\right) \qquad (9)$$

σ^2 in Equation 8 represents the vector of unique elements of the variance-covariance matrix **Σ** and α is the dispersion parameter. The values of both α and σ^2 reduces to alternative forms of count models. If the value of $\sigma^2 = 0$, the dispersion reduces to that of a standard negative binomial while if $\alpha = 0$, the dispersion reduces to that of a two-level random intercept Poisson model (Anders & Sophia, 2012). Furthermore, replacing the values of r and p_{ij} in Equation 8 above, gives the conditional log-likelihood function which can be written as:

$$l\,n\left(f\left(y_i|U_j\,,\alpha\right)\right) = exp\left[\sum_{i=1}^{n_{ij}}\left\{ln\left(\frac{\Gamma(y_{ij}+r)}{\Gamma(y_{ij}+1)\Gamma(r)}\right) - \frac{1}{\alpha}\;ln\left(1+\alpha\mu_{ij}\right) + y_{ij}ln\left(\frac{\alpha\mu_{ij}}{1+\alpha\mu_{ij}}\right)\right\}\right] \tag{10}$$

μ_{ij} is the linear predictor given by equation (5) which further assumes that random effects are multivariate normally distributed with zero mean and variance matrix Σ. The probability density function of the random effects is given as:

$$f_{U_j} = \frac{1}{\sqrt{(2\Pi)^K|\Sigma|}}\exp\left(-\frac{1}{2}\left(U_j\right)'\Sigma^{-1}\left(U_j\right)\right) \tag{11}$$

The likelihood contribution for the jth cluster is obtained by marginalizing the random effect out of the joint density function $f(y_j, U_j, \alpha)$. The resulting outcome in this operation is the marginal likelihood function given as:

$$\mathcal{L}_j(\beta, \Sigma, \alpha) = (2\pi)^{-\left(\frac{q}{2}\right)}\,|\Sigma|^{-\left(\frac{1}{2}\right)}\int \exp\left\{f\left(y_j\middle|U_j, \alpha\right) - \frac{U_j{}'\Sigma^{-1}U_j}{2}\right\}du \tag{12}$$

The log likelihood for the entire dataset is the sum of the contributions of the log-likelihood functions for the M individual clusters.

3.1.3 Estimation

Estimation of hierarchical models means obtaining maximum likelihood (ML) estimates which maximizes the likelihood of the given data. In the context of multilevel models and in particular non-linear model, the end results require a two-step procedure: (1) finding the likelihood by integration techniques and (2) maximizing the resulting likelihood function. The following text describes these two procedures by first explaining the parameters of interest.

3.1.3.0 Parameters of interest

As indicated by the LHS of equation (7), the parameters of interest are the regression coefficients contained in the matrix β, variance components in the matrix Σ where the diagonal elements are the variances of the random effects for the between-group variability and variances for the covariates which are allowed to vary, and the overdispersion parameter to characterize boundary conditions between appropriateness of the Poisson and negative binomial regression models. The elements in the matrix β constitute the fixed part of the model by estimating fixed parameter estimates. The variance parameters come from the fact that each segment has its own intercept and the cluster variability can be captured in each segment which is allowed to vary. For instance, consider the following model:

$$\log \mu_{ij} = \beta_0 + + \beta_1 \, enex_{ij} + u_{01j} + u_{1j} \, enex_{ij} \qquad (13)$$

Equation (13) implies that in addition to the between-segment variance captured by the cluster random effects u_{01j}, the type of segment variable (*enex*) is also allowed to vary and its residuals (random effects) are captured by u_{1j}. By allowing variability across segment types, then we term the variable "*enex*" a random parameter and in addition to its fixed parameter part estimated value $\widehat{\beta_{enex}}$, we also estimate the variance to characterize random part identified as ${\sigma_{u(enex)}}^2$. Such a matrix, where $\Sigma = \Omega_u$, can be specified as follows:

$$\begin{pmatrix} u_{01j} \\ u_{4j} \end{pmatrix} \sim MVN(0 \quad \Omega_u) \, , \quad \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \Omega_u = \begin{pmatrix} {\sigma_{u01}}^2 & 0 \\ 0 & {\sigma_{u1(enex)}}^2 \end{pmatrix} \qquad (14)$$

In equation (14), ${\sigma_{u01}}^2$ is the between-cluster variance while ${\sigma_{u1(enex)}}^2$ is the variance of the random parameter which in this case is the type of segment. The off diagonal elements of the matrix Ω_u are the covariances of the respective residuals. By assuming these parameters to be zero, the model introduces constraints on the variance-covariance structures and the matrix formed is called the variance components or the diagonal matrix. The size of the matrix Σ depends on the number of variables in the model and in this study, median shoulder, right shoulder and segment length are treated as random parameters in addition to their fixed parts. Another parameter of interest is the overdispersion parameter, α which helps to decide the appropriate model between Poisson and Negative binomial.

3.1.3.1 Estimating Parameters of interest

Estimation process entails obtaining the parameters of interest aforementioned such that the log likelihood function in equation (12) is maximized. In this case, the maximization is performed with respect to the model regression coefficients and the dispersion parameters which accounts for the cluster effects as well as the overdispersion. When such a condition to produce the maximum likelihood function is met, then the values of the parameters of interest achieved at this point are the optimum values which can be used for further inferences. However, discrete models such as count models have non-normal response at level 1 sampling model while higher levels involve multivariate normal assumptions (Raudenbush & Bryk, 2002), which in turn limits conventional estimation theory. To demosntrate this complexity consider the following system of equations where vehicular traffic volume is defined at level 1 while number of lanes is defined at the cluster-level which in this study, a cluster is a freeway segment:

Level 1: $\log \mu_{ij} = \beta_{0j} + + \beta_{1j}\,(volume)_{ij}$ (15)

Level 2: $\beta_{0j} = \gamma_{00} + + \gamma_{01}\,(lanes)_j + u_{0j}$ (16)

$\beta_{1j} = \gamma_{10} + + \gamma_{11}\,(lanes)_j + u_{1j}$ (17)

As it can be seen from the system of equations shown above, at level 1, our response variable represents counts which are modeled as the natural logarithm of crash frequency as a link function in addition to non-normal data. Since the regression coefficients β_{0j} and β_{1j} vary across clusters, a contextual variable – the number of lanes – is specified at level 2 to account for such variation. However, equations 16 and 17 models a continuous response in addition to the random effects u_{0j} and u_{1j} which represents deviations of the cluster regression coefficients from their overall crash frequency average. By substituting equations from level 2 into level 1 equation, the following combined equation results:

$$\ln(\mu_{ij}) = \gamma_{00} + \gamma_{10}\,(volume)_{ij} + \gamma_{01}\,(lanes)_j + \gamma_{11}\,(volume)_{ij}(lanes)_j + u_{0j} + u_{1j}\,(volume)_{ij} \tag{18}$$

Equation 15, is appropriate under the following assumptions:

$$E\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$Var\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = \Omega_u$$

$\tau_{10}\ and\ \tau_{01}$ are covariance parameters while $\tau_{00}\ and\ \tau_{11}$ are variance parameters. These parameters are the same as what is indicated in equation 14.

Equation 18 is a mixed-effect model with both fixed and random effects. To this point, it should be understood that fitting mixed-effects models require integrating out the random effects because they are unobserved. This implies that the resulting likelihood function is unconditional (also called marginal likelihood function) which does not involve U_j. This step involves the first stage towards obtaining likelihood function of the model. Using notations from equations 8 to 10, the following joint distribution function can be obtained:

$$g(Y, U/\omega) = f(Y/U, \omega) f_U \tag{19}$$

From Equation 19, $g(Y, U/\omega)$ is the joint distribution of the crash frequency and random effects vectors, $f(Y/U, \omega)$ is the probability distribution of the crash frequency at level 1 given the random effects and parameters of interest while f_U is the probability distribution of the random effects given the parameters of interest. The likelihood of the data given the parameters of interest ω is the marginal density of the crash frequency vector, Y. This is equivalent to integrating the joint function in equation over the space of the random effects and the following likelihood function results:

$$L(Y/\omega) = \int f(Y/U, \omega)\, f_U \tag{20}$$

The RHS of equation (20) which is equivalent to equation 10 for negative binomial model cannot be evaluated analytically and therefore numerical integration techniques are employed. One widely used modern method to directly estimate the integral required to calculate the log-likelihood is by using Gauss-Hermite quadrature. This is due to non-closed form of equation (20) where the prior distributions are not conjugate priors. For estimating model 20, the probability distribution of the random effects is approximated by a discrete distribution with q integrating quadrature points. This method permits likelihood ratio tests for comparing nested models and minimizing bias.

3.1.4 Inference

Inferences for the mixed effects model focuses on the fixed effect parameters and the covariance parameters which are contained in the model. To achieve this likelihood-based statistics are employed which include Wald statistic computed as the parameter estimate divided by its asymptotic standard error. The asymptotic standard errors are computed from the inverse of the second derivative matrix of the likelihood with respect to each of the covariance parameters. Wald statistics evaluates the significance of individual regression coefficients and it is given as:

$$Wald\ statistic = \frac{\hat{\beta}_j}{SE_{\widehat{\beta}_J}} \tag{21}$$

In case of nested models, a chi-square test statistic can be employed which compares full to a reduced model and a selected p-value is used as a cut off value for assessing significance of omitted model characteristics. If $\mathcal{L}_1$is the log-likelihood value associated with the full model and $\mathcal{L}_0$ is the log-likelihood value associated with the constrained or reduced model, the test statistic of the likelihood ratio test is given as:

$$LR - 2(\mathcal{L}_1 - \mathcal{L}_0) \tag{22}$$

If the constrained model is true, then the LR is approximately χ^2 distributed with $d_0 - d_1$ degrees of freedom associated with the reduced and constrained models respectively. Other evaluation techniques in addition to likelihood ratio tests are the Akaike and Bayesian Information criteria. These are based on log-likelihood function with adjustment for the number of parameters estimated and for the amount of data (Dobson, 2002). The Akaike Information Criterion is given as:

$$AIC = -2ln\mathcal{L} + 2k \tag{23}$$

while the Bayesian Information Criterion is given as:

$$BIC = -2ln\mathcal{L} + klnN \tag{24}$$

In both cases of equations 23 & 24, $ln\mathcal{L}$ is the maximized log-likelihood function of the model and k is the number of parameters estimated.

Assessing between Poisson and Negative binomial is based on the inference made on the dispersion parameter which quantifies the variability due to heterogeneity across freeway segments. If the dispersion parameter equals zero, the model reduces to the simpler Poisson model while if the dispersion is greater than zero the crash frequency is over-dispersed and negative binomial regression model can be employed for over-dispersed data. The dispersion parameter is also evaluated based on the likelihood ratio test statistics. Based on this approach I evaluated four models: traditional Poisson regression (PO) and Mixed-effects Poisson model (ME PO) as well as traditional negative binomial (NB) and Mixed-effects negative binomial (ME NB).

CHAPTER 4 METHODOLOGY: BAYESIAN SPATIAL MODEL

4.0 Introduction

Spatial data arise when observed units have locational component (LeSage, 1999). In this case, spatial dependence between the observations and spatial heterogeneity in the relationship being modeled are the problems expected. Spatial dependence is characterized by the Tobler's law (Tobler, 1979) which states that "Everything is related to everything else, but near things are more related than distant things". Two types of spatial dependence can occur: (1) positive spatial dependence are such that high values of a variable cluster in space, and (2) negative spatial dependence occurs when locations are surrounded by neighbors with very dissimilar values of the same variable.

Spatial dependence modeling requires an appropriate representation of spatial arrangement of observed areal units. In this case, relative spatial positions are represented by spatial weight matrices (Lee, 2011). These matrices can be prepared as either inverse distance weights matrices or binary contiguity weights matrices. This study adopts the latter type of matrices, binary contiguity weights matrices, which reflects the relative position in space of one unit of observation to other units by creating a matrix with ones and zeros whereby a matrix element is coded as one if two areal units share a common border and zero otherwise. In terms of spatial dependence, it is expected that neighboring units should exhibit a higher degree of spatial dependence than units located far apart. To demonstrate such coding processes consider the following arrangement of freeway segments represented here as a line diagram (Figure 11):

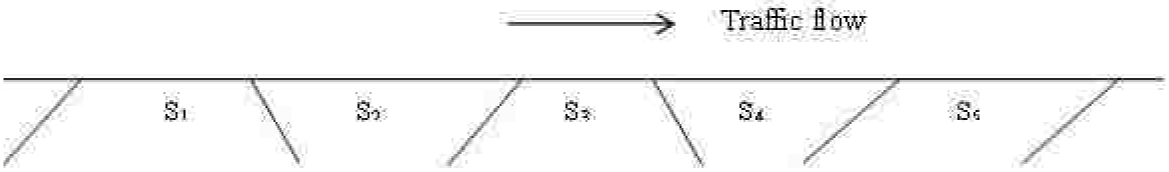

Figure 11: Line diagram representation of freeway segments for contiguity weight matrix

Based on the arrangement of segments shown in Figure 11, the following binary contiguity matrix can be constructed for segments s_1, s_2 and s_3:

$$W = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{25}$$

In the matrix above columns and rows are arranged to correspond with the arrangement of freeway segments as observed in the field. For instant, with the arrangement shown in Figure 11, freeway segment s1 corresponds to the first element in the matrix identified by the first column and first row. The above matrix controls spatial dependence structure of the random effects relating to geographically adjacent areas to be highly correlated while non-contiguous segments are conditionally independent given the values of the remaining random effects. The spatial weight matrix is then incorporated in Conditional Autoregressive models as explained in the section 4.1.2 of spatial model specification.

Another building block for appropriate using spatial models is the concept of Bayesian modeling techniques which is based on Bayes' theorem whereby posterior means are computed as the function of the data likelihood and priors. The model naturally identifies hierarchy since priors are being used to get posterior means (Dobson *et al*, 2008). Section 4.1 covers these concepts in a context of spatial models.

4.1 Bayesian Spatial Models for crash frequency

4.1.1 General review

A spatial phenomenon by which crash frequency values for contiguous roadway segments tend to be more similar or dissimilar constitute the concept of spatial autocorrelation (Flahaut, Mouchart, Martin, & Thoma, 2003; Hepple, 2009). Similar values of crash frequency across contiguous locations lead to positive autocorrelation and this occurs when the level of co-variation is higher than expected. Dissimilar values on the other hand result in negative autocorrelation when the level of covariation is such that higher crash frequency values are contiguous with low crash frequency values. The absence of spatial autocorrelation implies lack of significant positive or negative autocorrelation.

The distribution of the crash frequency in the context of a Generalized Linear Model can be described by a Poisson model (Noland & Quddus, 2004; Quddus, 2008) which is a member of the exponential family of distributions (Agresti, 2002; Dobson, 2002). Poisson model transforms the mean of the crash frequency to the natural parameter of a Poisson GLM. Such a transformation leads to a canonical link which facilitates modeling the natural logarithm of the mean of the crash frequency. The existence of spatial pattern in crash frequency can be modeled by a set of influencing factors including geometric elements of the freeways, traffic characteristics, and environmental factors as well as human factors and random effects which accounts for the possible effects of over-dispersion.

However, it is likely that not all the accounted factors in the modeling process fully explain spatial pattern in the crash frequency. In this case, Conditional Autoregressive Models (Aguero-Valverde, 2014; Aguero-Valverde, 2013) are specified to account for remaining spatial effects leading to residual autocorrelation in the crash frequency. CAR models contain a precision matrix to control the spatial autocorrelation structure of the random effects based on the weight matrix. Contiguity of freeway segments can be specified in the model by a binary coding where a code equals to 1 if the freeway segments share a common border and is zero otherwise. A spatial autocorrelation parameter with variance equals τ^2 is used to indicate the amount of autocorrelation in the crash frequency. If this value is significantly different from 0, it implies the existence of spatial autocorrelation of crash frequency for contiguous freeway segments.

Estimation of the unknown parameters for the aforementioned models is done in a Bayesian framework from which the unknown parameters are set to reflect prior knowledge. Specifically, diffuse normal priors are assumed for regression coefficients while uniform priors are adopted for random effects. Furthermore, the spatial autocorrelation parameter is assumed to have independent priors. Inference is based on Markov Chain Monte Carlo (MCMC) simulation using Gibbs and Metropolis steps as sampling techniques for posterior means. To ensure valid inference, Markov Chain convergence to the target densities, a specified number of samples is applied as a burn-in and thinning is adopted to reduce autocorrelation of neighboring samples.

This study investigates the existence of spatial autocorrelation on contiguous freeway segments with ramps as natural delineators while controlling for traffic and geometric characteristics observed using two count models: Non-spatial and Spatial Poisson models. It should be noted that the existence of spatial autocorrelation is an indication of the presence of unobserved factors unaccounted for which are manifested through the residual spatial autocorrelation. To this point it is hypothesized that crash frequency observed in contiguous freeway segments exhibit spatial phenomenon leading to spatial autocorrelation and a value of the spatial autocorrelation parameter significantly away from 0 is an indication of the existence of spatial phenomenon across adjacent segments. Section 3.0 dealt with general investigation of an existence of unobserved heterogeneity. Section 4.1 focuses on the special autocorrelation as a special case of unobserved heterogeneity in freeway segments causing crash frequency.

4.1.2 Global Moran's I Index

Identifying the existence of second order spatial effects (spatial autocorrelation) requires the use of spatial statistics (ESRI, 2013) before incorporating spatial effects terms in regression modeling process. Spatial statistics help to identify patterns of crash frequency across freeway segments. This study uses Moran's I statistics, an index which incorporates spatial concepts to reflect the existence of spatial autocorrelation based on the segment locations and corresponding crash frequency. The conceptual models employ the zone of indifference to conceptualize spatial relationship and the index is calculated using ArcGIS software as follows:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i}^{n} {z_i}^2} \quad (26)$$

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \quad (27)$$

From equations (26) and (27), z_i is the deviations of crash frequency for segment *i* from its mean crash frequency, $w_{i,j}$ is the spatial weight between segment i and *j*, n is the total number of segments, and S_o is the aggregate of all the spatial weights. The z-statistic for the index is given as follows:

$$Z_I = \frac{I - E[I]}{\sqrt{V[I]}} \quad (28)$$

where $E[1] = \frac{-1}{n-1}$ is the expected index and its variance is given as:

$$V[I] = E[I^2] - E[I]^2$$

Interpretation of the existence of spatial autocorrelation given by the index is based on the values of the z-scores and the p-value. This study uses a cut off value for p-value equals to 5% significance. In this case, a statistically significant p-value with a positive z-score implies that spatial distribution of high crash frequencies and low crash frequencies is more spatially clustered than would be expected by the existence of random process. On the other hand a statistically significant p-value with a negative z-score indicates the existence of spatially dispersed high crash frequencies and low crash frequencies in comparison to random process.

4.1.3 Specification of hierarchical Bayesian model

Let y_i denote the number of crashes observed in a freeway segment for the i^{th} covariate pattern. Let the expected value of y_i depend on the explanatory variables x_i. The Poisson generalized linear model (Guo, Wang, & Abdel-Aty, 2010) with the natural link function in the logarithmic function can be specified as:

$$ln(\mu_i) = x_i{}'\beta, \qquad i = 1, \ldots., n \tag{29}$$

μ_i denotes the expected values of the crash frequency for segment i, $x_i{}'$ is the matrix of observed influencing factors including an intercept, and β is the matrix of regression coefficients which quantifies the impact of covariates on the expected crash frequency.

Equation 1 can be used to model spatial pattern in the crash frequency across freeway segments via a matrix of the covariates which in this case are the geometric and traffic characteristics observed on the freeways. However, the observed crash frequency for Poisson model exhibit over-dispersion and to capture this effect, equation 1 can be extended to include random effects, $\emptyset_i$ to account for the possible effects of over-dispersion:

$$ln(\mu_i) = x_i{}'\beta + \emptyset_i, \qquad i = 1, \ldots., n \tag{30}$$

Under Bayesian modeling frame work, prior distributions for the unknown parameters are set to reflect prior knowledge about the parameters of interest (Guo, Wang, & Abdel-Aty, 2010). In this case an independent Gaussian prior (diffuse normal priors) is assumed for each regression coefficient, $\beta_j \sim N(m_j, \nu_j)$ with mean, m_j equals 0 and variance, ν_j necessarily large. Uniform priors are assumed for random effects, $\emptyset_i \sim U(0, \sigma^2)$ where $\sigma^2 \sim U(0,\ M_\sigma)$ with large variance, M_σ.

It is further assumed that there exist second order spatial effects (Bolstad, 2005; Gelman, Carlin, Stern, & Dunson, 2014) unaccounted for by the covariates and specify a Conditional Autoregressive Priors (Lee, 2011; Kery, 2010):

$$\emptyset_k | \emptyset_{-k} \sim N\left(\frac{\rho \sum_{i=1}^{n} w_{ki} \emptyset_i}{\rho \sum_{i=1}^{n} w_{ki} + 1 - p}, \frac{\tau^2}{\rho \sum_{i=1}^{n} w_{ki} + 1 - \rho}\right) \tag{31}$$

Equation 3 is a special case of the Gaussian Markov random field which contains a precision matrix to control the spatial autocorrelation structure of the random effects based on the weight matrix W. Contiguity of freeway segments can be specified in the model by a binary coding where $w_{kj} = 1$ if the freeway segments share a common border and is zero otherwise. ρ is a spatial autocorrelation parameter with variance equals τ^2. If the value of ρ is significantly different from 0, it implies the existence of spatial autocorrelation of crash frequency for contiguous freeway segments. Both ρ and its variance parameter, τ^2 have an independent prior specified as follows:

Spatial autocorrelation: $\rho \sim U\ (0,1)$;

$$\tau^2 \sim U\ (0, M_\tau) \tag{32}$$

Decision to adopting equation 31 is based on the appealing fact that (Kery, 2010) conducted a comparative research and identified that the random effects modeled by a conditional autoregressive (CAR) prior distribution specified by equation 31 is the best because it produces consistently good results across the range of spatial correlation scenarios considered. It also represents a range of strong and weak spatial correlation structures with a single set of random effects which is beyond the models proposed.

Inference for the above models is based on Markov Chain Monte Carlo (MCMC) simulation (Kery, 2010; Lee, 2011; Dobson, 2002) using a combination of sampling techniques. The variance parameters are Gibbs sampled from their full conditional truncated inverse gamma distributions, while the remaining parameters are updated using Metropolis steps. An important key part of the analysis based on sampling techniques is to be able to make valid inferences. This is possible by monitoring markov chain convergence to the target densities. To ensure the markov chain lies within the stable area of high likelihood we apply a burn-in of 20,000 samples to ensure that the samples drawn from the chains approximate the posterior distribution. A thinning equal to 10 is applied to reduce autocorrelation of neighboring samples (Dobson, 2002). The results of convergence are monitored for stable posterior distributions based on trace plots and posterior densities of covariates.

To estimate the specified models CARBayes package is applied in an R software environment (Lee D. , 2014) and WINBUGS version 1.4.3. Choosing the most parsimonious model is based on the Deviance information criterion (DIC), which is a generalization of Akaike Information Criterion (AIC) for Bayesian models. Evaluation of the significance of estimated parameters is based on 95% credible intervals.

CHAPTER 5 DATA COLLECTION AND SUMMARY

5.0 Multilevel count model data

5.0.1 Data collection approach

Data used in this study comprised of traffic and geometric characteristics as well as crash frequency from the entire freeway network located in the Las Vegas area of Nevada. Traffic characteristics included average speed and traffic volumes recorded by inductive loop detector and were downloaded from the website managed by the Freeway and Arterial System of Transportation (FAST) division which is managed by Regional Transportation Commission of Southern Nevada (FAST, 2014). Clustered data were observed at two levels: Level 1 which is considered as the lowest level consisted of short stretches of freeways where every sensor installed within a given stretch records traffic characteristics and the variables at this level constituted only traffic volumes and average speed. Traffic volumes as well as speeds are recorded in windows of 15 minutes and to get the total volume, data were aggregated for every month and then summed for 12 months to obtain yearly volumes. However, speed variable values were processed as an average.

To get data at Level 2, natural delineation of the freeways were used, which consisted of freeway segments delineated by entrance and exit ramps and this resulted in weaving and non-weaving segments. Data at this level comprised of geometric characteristics which included the right shoulder, median shoulder, base length of the segments, type of segments, and total number of lanes. To obtain these segments, ArcGIS (ESRI, 2013) were used to create a GIS shapefile which in turn were overlaid on base map for visualization purposes for geometric characteristics. The length of our segments, known as the base length (Roess, Prassas, & McShane, 2011; TRB, 2010), was taken as the distance between the gore points of the entrance and exit ramps as indicated on Figure 12. As indicated in figure 10, EN-EX segments were placed under one group of weaving segments and coded accordingly in our dataset to distinguish it with non-weaving segments denoted as EX-EN segments. The configuration aforementioned helps to identify the variation between weaving and non-weaving segments. Furthermore, observations revealed the existence of relatively short segments. Therefore a categorical variable was created to compare segments with base length less than 677 meters compared to segments with length greater than 676 meters. This variable comprised of segments with short base length as indicated in Table 1.

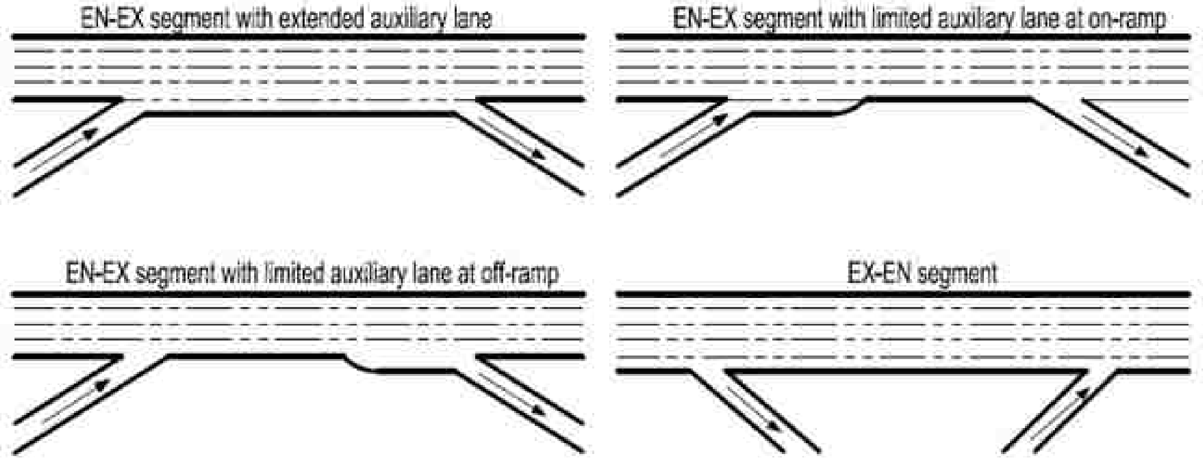
EN-EX segment with extended auxiliary lane
EN-EX segment with limited auxiliary lane at on-ramp
EN-EX segment with limited auxiliary lane at off-ramp
EX-EN segment

Figure 12 Segments used at Level 2 with base length definition

Using visual aid from both ArcGIS and google earth software, geometric characteristics were collected and their characteristics summarized as indicated in Table 1 in section 5.0.2

5.0.2 Summary statistics

Table 1 Descriptive Statistics for Multilevel model

Variable	**Scale**	**Mean**	**Std. Dev**	**Min.**	**Max.**
Crash frequency	Number	3.9846	5.0557	0	48
Log of traffic volume	Natural log	7.2333	0.2361	6.4503	8.1216
Width of right shoulder	Meter	3.6301	1.2562	0.9100	9.1800
Width of Median shoulder	Meter	2.9947	1.1720	0.8300	8.6700
Segments length	Categorical	0.2385	0.4270	0	1
Average vehicle speed	Miles per hour	75.1039	5.8211	47	81
Weaving segments	Categorical	0.4231	0.4950	0	1
Through lanes	Number	3.4577	0.8532	2	7

It should be noted that to maintain data structure for the proposed model, traffic volumes and average speed varied both at Level 1 and 2 while geometric characteristics only varied across the weaving and non-weaving segments. Number of crashes in each freeway stretch (at Level 1) were visualized and collected from the website managed by the Freeway and Arterial System of Transportation (FAST) division and linked with the traffic and geometric characteristics. As crash data indicates in Table 1, there is a greater variability within the crash frequency which indicates dispersion amounting to 1.268793 (5.055651/3.984615).

The average mean of the width of right shoulder was observed to be 3.63 meters with variability of approximately 1.2 meters while the width of the median was equal to 2.99 meters with a standard deviation of 1.1 meters. Descriptive statistics also indicates that there are approximately 24 percent of the freeway segments which have a relatively short base length. Forty two percent of our segments at level two comprised of weaving segments while 58% are non-weaving segment. Combining all of the data, we observed 260 freeway stretches where sensors are installed with a minimum number of two stretches and a maximum of four within segments at Level 2 and the final modeling tasks retained only significant variables as explained in the following section.

5.0.3 Graphical summaries

In addition to the summary statistics described above, graphical summaries can also help as a visualization tool of the observed relationship.

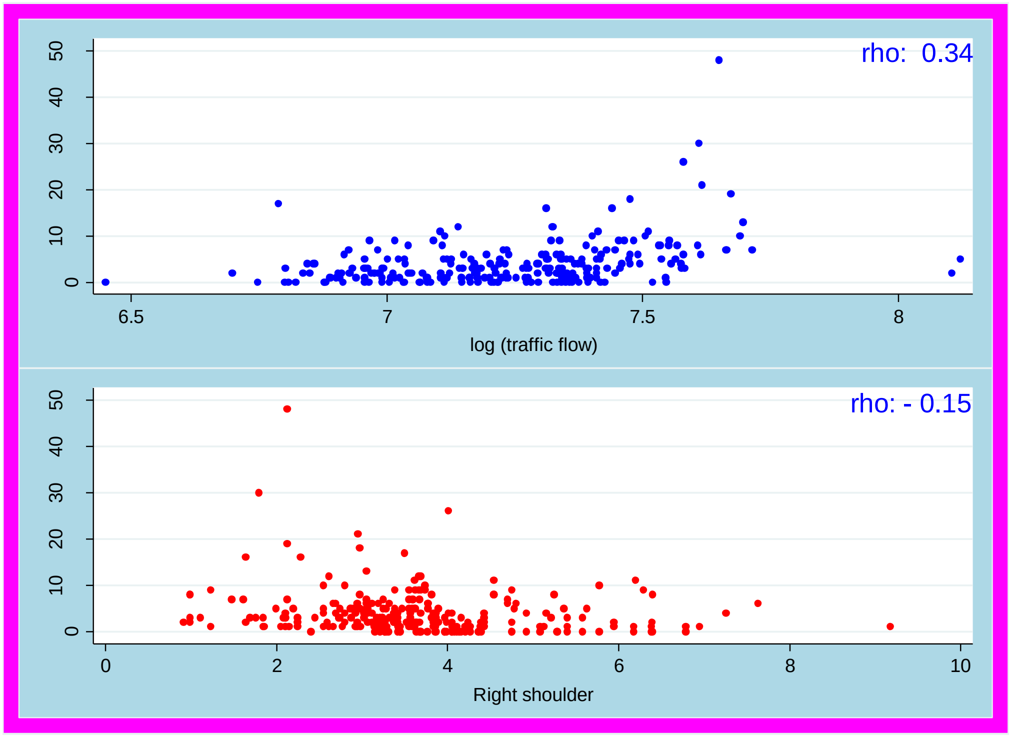

Figure 13: Relationship of crash frequency, right shoulder and logarithm of traffic flow

Figure 13 indicates the observed relationship of crash frequency against the right shoulder and logarithm of traffic flow plotted in a scatter plot. The upper panel of the graph displays distribution of crash frequency across the freeway segments. It is clear from the plot that segments with traffic volume ranges from approximately 6.7 to 7.7 in logarithmic scale experienced high crash frequency with an average of 7.3. Furthermore, the plot indicates less variability of crashes around the average. The bottom panel displays the relationship of crash frequency and the right shoulder in meters. There are concentrations of crashes for segments with narrow width of the right shoulder ranging from 2m to 4m, a result which shows the existence of negative impacts for narrow shoulders. The upper right corner of the two plots also displays the strength of relationship between the aforementioned geometric and traffic features with the level of crashes. It is clear that crash frequency is positively correlated with traffic levels while there are observed negative correlation between correlation for the right shoulder and crash frequency.

Figure 14 below is a box plot which indicates the distribution of crashes across the number of through lanes. The maximum number of lanes observed to have crashes occurred was seven although, only one segment was observed and this is not included in the box plot shown below. For the remaining segments, the plot indicates differential impacts of the number of lanes on crash occurrence across the segments. With the exception of segments with two and six lanes, there was an increasing trend in the mean crash frequency as the number of through lanes increased from three to five lanes. Crash frequency distribution behavior exhibited across segments with two and six lanes indicates the existence of other factors leading to crashes.

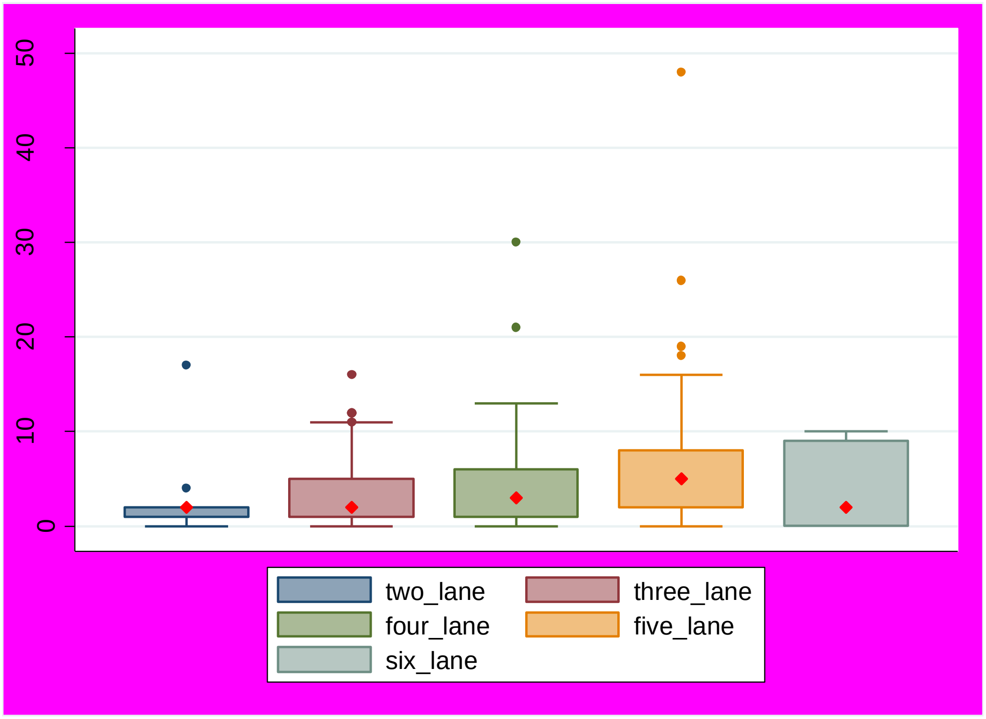

Figure 14: Distribution of crash frequency across the number of through lanes

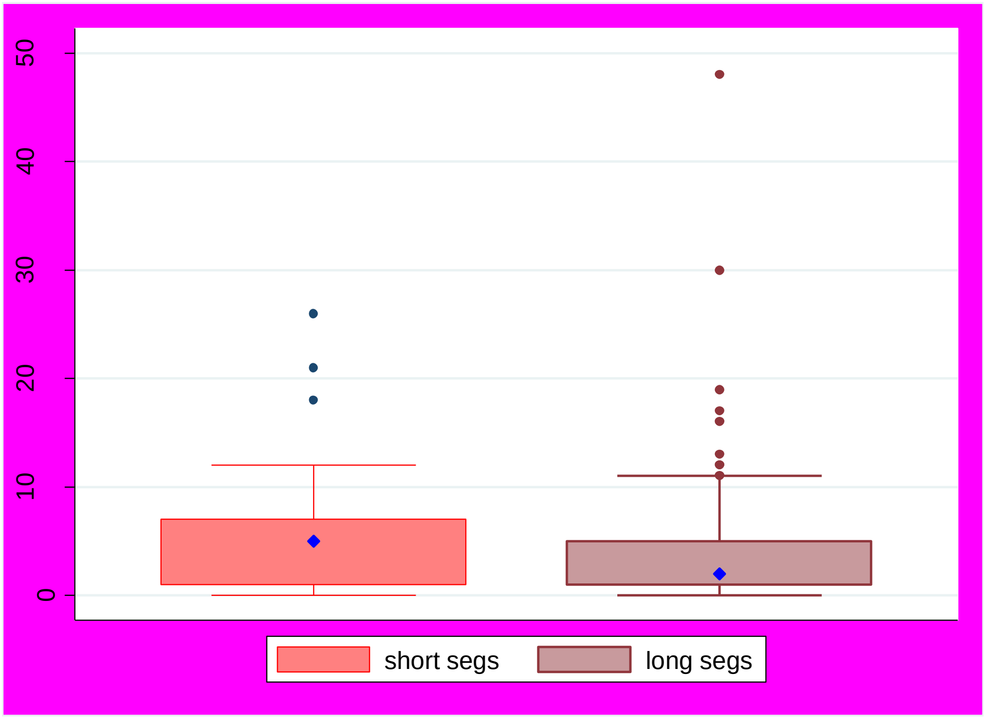

Figure 15: Distribution of crashes across segment length

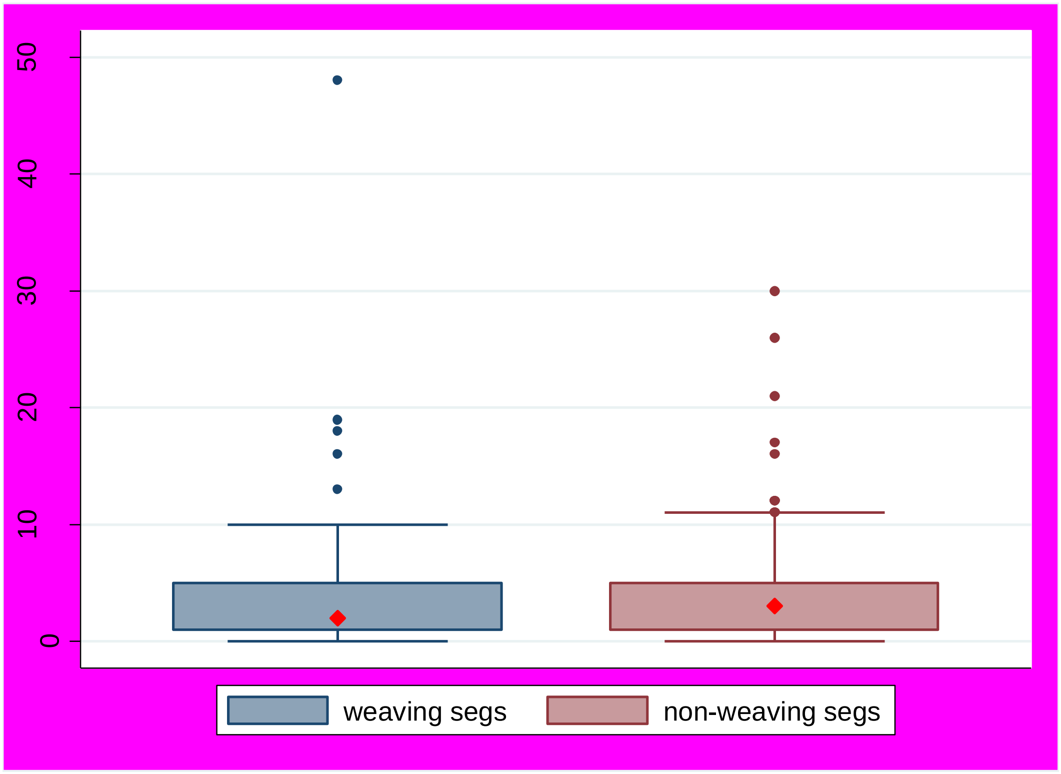

Figure 16: Distribution of crashes across weaving and non-weaving segments

Figure 15 shows the distribution of crash counts across segment lengths. Due to the nonlinear nature observed for relationship between crashes and segment lengths, the distribution is divided into short and long segments. The short segments considered in this case are segments with length equal or less than 0.5 mile while segments with length greater to 0.5 mile were considered as long segments. From the figure it can be seen that short segments had the highest mean crash frequency when compared to long segments. However, this observation cannot be conclusive because many factors interact together to generate the observed crash frequency. In this a multivariate model to be estimated and discussed in chapter 6 may reveal different trends of influencing factors.

Figure 16 above is a box plot which indicates the distribution of crashes across weaving and non-weaving segments. The comparative box plots shows that weaving segments had the lowest mean crash frequency compared to non-weaving segments while there was high mean crash frequency for short segments compared to long segments. However, a closer look of the weaving and long segments indicated the existence of few segments with extreme values of crash frequency although in general mean crash frequency was low in these areas. These observations may indicate the existence of other crash generating factors than the geometry under investigation.

From the above summary and graphical results, it is evident that the occurrence of a crash is not influenced by the observed factors only. Other hidden factors are contributing to the observed variability in crashes. This can easily be seen from different graphical displays particularly the existence of random behavior for some of the geometric factors such as the number of lanes and the extreme values of crash frequency observed in the case of weaving and long segments contrary to what would be intuitively expected. To spot out these differences causing crash frequency variability requires an advanced model to effectively separate variability across segment and within individual influencing factors.

5.1 Bayesian spatial model data

This section requires dataset with a structure focusing on the investigation of a special case of unobserved heterogeneity: spatial dependence of crash frequency for contiguous freeway segments. The model inputs are traffic and geometric characteristics from contiguous freeway segments extracted from loop detectors managed by FAST. Freeway segments which shares a common border identified as natural delineation between entrance and exit were considered.

Since the purpose is to identify the existence of spatial dependence, contiguous freeway segments with missing traffic characteristics were removed from the study and retained only segments with all information required.

Based on the aforementioned criteria, a total of 36 Segments were selected for study. Using ArcMap, a polygon shapefile were created for all segments under study with visual aid from a base map as a tracing tool. Furthermore, sensor codes with their locations were observed from Google maps and matched with the created GIS shapefile of freeway segments and traffic characteristics which included vehicular speed and traffic volumes were extracted for each sensor located on those segments.

Geometric characteristics were obtained by changing a GIS shapefile to KMZ and overlay the resulting KMZ file on Google earth map for visual aid. Number of lanes, median shoulder and right shoulder were observed and measured from the overlaid KMZ file as shown on Table 2 which shows summarized data.

Table 2: Descriptive Statistics for Bayesian spatial model

Variable	Scale	Mean	Std. Dev.	Min	Max
crashes	Number	57.7778	86.5801	0	411
Centered width of Log of right shoulder	Natural logarithm	0.9969	0.3323	-0.0163	1.6832
Log of base length of segment	Mile	0.6766	0.4409	0.2019	2.5727
Number of lanes	Number	4.3056	1.2833	2	7
Number of lanes * Log of base length of segment	Mile	2.7146	1.5166	0.8077	7.7180
Weaving segment	Categorical	0.4167	0.5000	0	1

CHAPTER 6 MULTILEVEL COUNT MODEL RESULTS

6.0 Model structural form assessment

Multilevel models comprise four count models: two are traditional Poisson and Negative binomial models where random effects are not included and two other models included random effects terms to account for the possible existence of segment-specific and individual-factors effects and both models were estimated based on total of eight factors.

Traffic characteristics include vehicular speed and traffic flow while geometric characteristics include total number of lanes, widths of the right and median shoulders, segment length as well as the type of a segment. One of the distribution assumptions of count models, a structural relationship function, is that a factor is linearly related to the natural logarithm of the expected crash frequency or exponentially related to the expected crash frequency.

Furthermore, measurement scale may also obscure the significance of a variable under investigation or reduce its levels of impact. Issues of measurement scales in this study indicated the need for model structural form assessment based on different function forms of the influencing factors. As indicated in Table 3, the variable involving traffic flow and segment length were found highly significant in the preliminary results with a z-statistic of 10.43 for traffic flow and -2.62 for segment length. These values of statistics are an indication of the significance in influencing crash occurrence. However, their impacts on the crash frequency occurrence are approximately zero for both of them.

Table 3: Model set #1 results based on original explanatory variables

(Dep. var = crash frequency)	**PO**		**ME PO**		**NB**		**ME NB**	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Traffic volume	0.000	10.43	0.000	2.18	0.000	5.57	0.000	2.95
Right shoulder	-0.193	-7.13	-0.085	-1.93	-0.176	-3.50	-0.146	-2.70
Seg. base length	-0.000	-2.62	-0.000	-1.88	-0.000	-1.58	-0.000	-1.97
Weaving segments (enex seg.)	-0.190	-2.65	-0.341	-2.47	-0.179	-1.29	-0.289	-1.83
Intercept	-1.998	-14.76	1.604	6.37	1.592	5.35	1.754	5.76
Covariance parameters								
Variance - Intercept (Std. err.)				0.309 (0.172)				0.206 (0.105)
Variance - enex seg. (Std. err.)				0.412 (0.099)				0.193 (0.187)
Model fit criteria								
lnL		-831.014		-655.203		-620.319		-614.043
AIC		1672.028		1324.407		1252.64		1244.09
BIC		1689.831		1349.331		1274.00		1272.57
Alpha (α)						0.667		0.370
LR test for α = 0						421.39 (0.000)		-4.06 (0.000)
LR test Vs PO & NB (p-value)				351.62 (0.000)				15.55 (0.0002)

In order to describe properly data structure, different functional forms were tried by transforming the traffic flow and segment length to account for the problems of measurement scales. As shown in Figure 16, the relationship of crash frequency and traffic flow in 10th millions of vehicles grows faster for small number of flow compared to the trending observed when the log of traffic flow is related to the crash frequency. Comparing these relationships, the logarithmic form of traffic flow is more highly related to the crashes than traffic flow in 10th mill. of vehicles. This indicates that crashes may likely be linear in logarithmic forms with traffic flow.

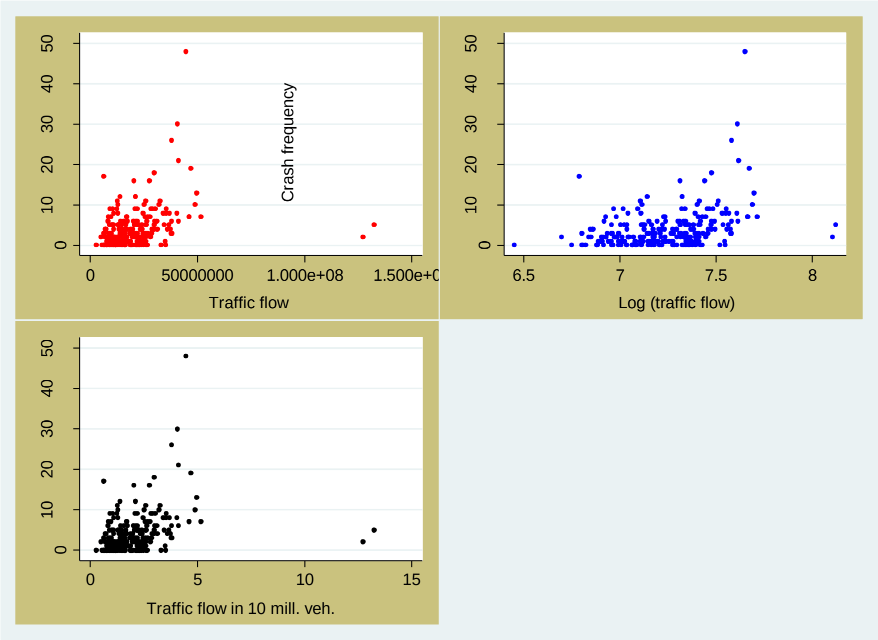

Figure 17: Scatter plot of crashes and traffic flow indifferent functional forms

To avoid problems observed by modeling using traffic volume and segment lengths in their original forms, six model sets (see the appendix) where estimated with different functional forms of the segment length and traffic flow. For each model set, four count models were estimated, two of them without random effects terms and the other two included the random effects terms. Table 4 shows different functional forms of the traffic flow and segment length and their combination in the estimated model sets.

Table 4: Functional forms

Model set #	**Function forms ad their combination in the model sets**
Model set 1	Traffic flow, segment length
Model set 2	Traffic flow, log (segment length)
Model set 3	Log (Traffic flow) , segment length
Model set 4	Log (Traffic flow), log (segment length)
Model set 5	Traffic flow in 10mil. Vehicles, segment length
Model set 6	Traffic volume in 10mil. Vehicles, log(segment length)

Note: See the appendix for complete functional forms specifications.................................

Along with the functional forms displayed on Table 4, other variables included in the proposed model sets were type of a segment, median shoulder, number of lanes and vehicular speed in their transformed forms. Through a successive processing of modeling and account for the combined effects of variables to the expected crash frequency, the Wald and overall significance tests were conducted to come up with an initial set of significant variables. The Wald tests were applied for testing overall significance of explanatory variables on models with random effects while the likelihood ratio test were used to test overall significance of the regression models on count models without random effects. Based on these tests, only variables which revealed a significant combined effect on the expected crash frequency were retained in the model sets.

The six model sets estimated show the effects of different functional forms on the regression coefficients and their levels of impacts on the crash frequency. To select the appropriate functional forms which could describe the data set structure observed, selection criteria based on Akaike Information and Bayesian Information criteria were applied with the requirement that, a model with the smallest information criterion is considered the appropriate model. The information criteria are further designed to penalize factors which do not contribute to the impact on the expected crash frequency but also they account for the model complexity. Tables 5 and 6 show assessment criteria used in evaluating and selecting the appropriate function forms.

Table 5: Selection criteria for the six model sets

Model sets	**Poisson Model**		**Negative Binomial model**	
	AIC	BIC	AIC	BIC
Model set 1	1672.03	1689.83	1252.64	1274.00
Model set 2	1673.50	1691.30	1252.89	1274.25
Model set 3	1594.00	1611.80	1248.99	1270.35
Model set 4	1594.36	1612.17	1249.11	1270.48
Model set 5	1672.03	1689.83	1252.64	1274.00
Model set 6	1673.50	1691.3	1252.89	1274.25

Note: * indicates the appropriate model selected

Table 6: Selection criteria for the six model sets

Model sets	**Mixed-effects Poisson model**		**Mixed-effects negative binomial model**	
	AIC	BIC	AIC	BIC
Model set 1	1324.41	1349.33	1244.09	1272.57
Model set 2	1324.61	1329.54	1243.94	1272.42
Model set 3	1321.09	1346.01	1237.80	1266.29
Model set 4	1321.21	1346.13	1237.53	1266.01
Model set 5	1324.41	1349.33	1244.09	1272.57
Model set 6	1324.61	1349.54	1243.94	1272.42

Note: * indicates the appropriate model selected

As shown on Tables 5 and 6, model set 3 for the traditional model had small values compared to the other model sets. As shown in Table 3 for model set 3, transformation was applied on traffic flow in logarithmic form while the segment length was modeled in its original form. The mixed effects picked up models from different model sets (Table 6). These involved the combined functional form in logarithm form for segment length and two forms of transformation on traffic flow: the log of traffic flow and traffic flow in 10^{th} millions of vehicles. Having these values describing transformed forms, the next section discriminates further the selected models based on count model distributional assumption of the dispersion parameter, likelihood ratio tests for nested models as well as Akaike and Bayesian Information criteria which accounts for model complexity in terms of the functional form which passed the test in section 6.0.

6.1 Final model selection and interpretation

The final models obtained from functional form assessment are summarized in Table 6 and these include Poisson model (PO), mixed-effects Poisson model (ME PO), negative binomial model (NB), and mixed-effects negative binomial model (ME NB) for variables found significant based on the individual statistic tests and the overall tests. The models are evaluated based on overdispersion effects and by considering nested and non-nested structures (Cameron & Trivedi, 2013). Dispersion parameter was used to account for overdispersion between Poisson (PO) and negative binomial (NB). Likelihood ratio test criterion was used for nested models (NB and ME NB) while Information criteria (AIC and BIC) were used for non-nested models (ME PO and ME NB). Based on model fit criteria, the value of α = 0.626 and its test resulted in a chi square statistic of 340.43 with p-value equals 0.000. This implies that the data at hand could be modeled by negative binomial model (NB). However, appropriateness of negative binomial (NB) can only be possible for the fixed-effects part only and observed data was clustered with random parameters. This required extending the traditional negative binomial to accommodate random parameters and therefore it was appropriate to compare NB with ME NB. The likelihood ratio test resulted in a chi square of 15.55 with a p-value equal to 0.000 which implies that the appropriate model was a mixed-effects negative binomial. Mixed-effects Poisson model was also found appropriate when compared to Poisson (PO) model as shown by the likelihood-ratio test with chi square of 270.23 and a p-value of 0.000. Under these results it was appropriate to rule out negative binomial (NB) and Poisson (PO) and considered the mixed-effects models (ME PO and ME NB).

Comparison of mixed-effects Poisson model (ME PO) and mixed-effects negative binomial model (ME NB) based on the information criteria. Two criteria were used for comparison: Akaike information criterion (AIC) and Bayesian Information criterion (BIC). The requirement under these criteria statistics is that a model with low value is considered to have a good-fit under the data at hand. As the results indicated on Table 7, the mixed-effects negative binomial had the lowest values for both AIC and BIC values. Under these results, mixed-effects Poisson model was eliminated and the final model according to the selection criteria was the mixed-effects negative binomial model (ME NB). The final model also had a significant heterogeneity parameter with a value equal to 0.353 which is statistically significant with p-value equal to 0.000.

6.1.2 Model interpretation

Table 7 below indicates the final model selected which was the mixed-effects negative binomial model. These results comprised of variables with fixed effects as well as random effects parameters. Fixed effects parameters included explanatory variables at the lower level and the second level. As aforementioned in the methodology section, a freeway segment in this case is treated as a cluster while the small sections dividing the cluster occupy level one. Variables observed at level one was the counts of crashes, traffic flows, and vehicular speed. Only the traffic flows modeled as function of logarithm were significant. Explanatory variables at the segment level are also called contextual variable and vary across the freeway segments.

Table 7 Final models selected

(Dep. var = crash frequency)	PO		ME PO		NB		ME NB	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Log (Traffic volume)	1.723	12.76	0.727	2.90	1.623	6.30	1.291	4.30
Right shoulder	-0.172	-6.43	-0.091	-2.11	-0.180	-3.61	-0.152	-2.86
Base length	-0.000	-1.85	-0.546	-1.86	-0.000	-1.48		
Log (Base length)							-0.690	-1.88
Weaving segments (enex seg.)	-0.365	-3.30	-0.361	-2.63	-0.183	-1.31	-0.324	-2.11
Intercept	-10.410	-10.11	-3.472	-1.90	-9.531	-4.92	-5.477	-2.07
Covariance parameters								
Variance - Intercept (Std. err.)				0.318 (0.166)				0.191 (0.173)
Variance - enex seg. (Std. err.)				0.357 (0.090)				0.179 (0.088)
Model fit criteria								
lnL		-792.000		-653.542		-618.500		-610.76
AIC		1594.000		1321.08		1248.99		1237.53
BIC		1611.801		1346.01		1270.36		1266.01
Alpha (α)						0.636		0.362
LR test for α = 0						347.01 (0.000)		-4.42 (0.000)
LR test Vs PO & NB (p-value)				276.91 (0.000)				15.58 (0.000)

Note: $\log \mu_{ij} = \beta_0 + \beta_1 log(volume)_{ij} + \beta_2\, rshoulder_j + \beta_3 slength_{ij} + \beta_4\, enex_{ij} + u_{01j} + u_{4j}\, enex_{ij}$ (33)

$$\begin{pmatrix} u_{01j} \\ u_{4j} \end{pmatrix} \sim MVN(0 \quad \Omega_u)\,, \quad \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Omega_u = \begin{pmatrix} {\sigma_{u01}}^2 & 0 \\ 0 & {\sigma_{u4}}^2 \end{pmatrix}$$

Var $(u_{01j} + u_{4j}\, enex_{ij}) = {\sigma_{u01}}^2 + {\sigma_{u4}}^2\, (enex_{ij})^2$ (34)

Variables: ***log (volume)*****:** log of traffic volume; ***rshoulder*****:** right shoulder; ***slength***: short base length; ***enex:*** weaving segments;
PO: Poisson model; **NB**: Negative Binomial model; **ME PO**: Mixed-effects Poisson model; **ME NB**: Mixed-effects Negative Binomial

Table 7 revealed interesting results which shows the importance of including random terms in modeling traffic safety. Of particular importance are the changes in the regression coefficients across models with and without random effects. In should be understood that random effects represents a group of influencing factors which are believed to generate the underlying mechanisms of crashes on the road network. These factors may be known but cannot be collected or they may be unknown. The random effect represents the overall effects and they are summarized by the distributional parameters in this case the random effects are summarized by their variances.

It can be seen that there is a consistence revealed by the direction of changes in the regression coefficients. Specifically models without random effects are shown to have high regression coefficients compared to those models with random effects. This is an indication of biasness in the quantities of the regression coefficients for models which did not include random effects. Basing our decision on these models is likely to lead to making unrealistic judgement of our safety levels of the freeway segments.

On the hand, the models with random effects are likely to generate realistic quantification of influencing factors because according to the model estimation theory, the group effects which are random effects are factored out and summarized as variances. This implies that the estimated effects of the explanatory variables involve on those factors retained in the model. These effects can be revealed by the changing of regression coefficients as one moves from using a convention count model to a mixed effect model which accounts for all the factors which were not included in the model. For all models which included random effects, their regression coefficients were found small compared to those from the conventional count models.

6.1.1 Fixed effects coefficients interpretation

After accounting for the effects of all factors which are either unknown or cannot be collected easily and randomness within a single explanatory variable such as the type of a segments, the first part of the mixed-effects model displays results indicating quantification of the influencing factors of the observed model on the crash frequency occurrence. These results revealed fixed effects of factors on the occurrence of crash frequency and by definition a fixed effect refers to the assumption about the influencing factors. By fixed effects in this case means that the estimated regression coefficients of explanatory variables are assumed to be constant across the clusters, that is the freeway segments.

As shown on Table 7, all geometric variables retained in the model had negative effects on the crash frequency occurrence while traffic characteristics reflected in the traffic flow displayed positive effect. Results indicating the effects of traffic flow on crash frequency had a nonlinear relationship and with the application of log on traffic flow, the effect is of constant elasticity effect. Specifically, the results indicate that for a one percent increase in traffic flow, there was an increase of 1.291% increase in the crash frequency. This means that more vehicular traffic increased crash counts within the clustered subsections with the same base segment length and segment type. This is intuitively true because with an increase in the number of vehicles using the facility gaps between vehicles is reduced and therefore the likelihood of drivers to be involved in crashes is high. It is also likely that with a congested facility resulting from an increased number of vehicles, drivers are likely to be influenced in their driving behavior leading to maneuverability which in turn is likely to result in more crashes.

The coefficient on the right shoulder was found negative which indicates that wider shoulders reduced the log of crash counts between segments of the same segment length, traffic flow as well as of the same type of segment. This can be explained by the fact that with the wider shoulders, there would be enough area for drivers to maneuver to avoid crashes. In addition, wider shoulder also tends to improve safety by providing a stable, clear recovery area for drivers who have left the travel lane.

Segments with relatively long base length reduced the crash frequency for segments with the same levels of traffic flow, shoulder width and of the same type. These are expected results because when compared to relatively short segments (length less than 0.5 mile) it is obvious that there will be an increase in the crash counts due to complex driving environment which may contribute to the occurrence of crashes. With respect to short segments, Liu *et at* (2010) found the same results for freeways with closely spaced entrance and exit ramps.

Weaving segments (EN-EX type) reduced the log of crash frequency compared with non-weaving segments for the same fixed values of the other roadway and traffic characteristics included in the model. The results may be explained by the existence of speed-change lanes where drivers have time to make decisions whether to accelerate and merge with the through traffic or decelerate to diverge from the main facility. This fact helps them to avoid risk hazards which may be encountered because more time will be available while drivers are still on speed change lanes before taking an action of either merging or diverging.

6.1.2 Random effects coefficients interpretation

The second part of the model as shown in Table 7 is the random part of the model. The random part accounts for unobserved variability in crash frequency arising from the unobserved factors believed to influence crash occurrences on the network. Furthermore, in addition to the group effects which summarize all factors effects not modeled, variability can also arise when there are individual differences within a specific explanatory variable. The formal variability is summarized by the model intercept which is allowed to vary to account for effects arising from the combined effects of unobserved factors. The latter variability accounts for the individual factors differences across the freeway segments. For instance, segment lengths vary from one segment to another and its corresponding effects will also vary. The proportional of influence which leads to excessive variability is accounted for by allowing the regression coefficients on that variable to have a random effect and is summarized by its variance.

With respect to this study, the focus was to investigate the existence of variations between – segments (measured by the slope variance) and variation in crash frequency between the segment regression lines. Only the segment type was found to have a significant variation in influencing crash frequency. The variance for the intercept was equal to 0.139 with standard error of 0.165. This implies that approximately 13.9% of variation can be attributed at the segments level (EN-EX Vs EX-EX segment types).

These results revealed a clear difference across the freeway segments. Specific to these results is the differential impact of factors attributed to local environment and operational effects in the mechanism of crash occurrence. Factors such as aging, mental ability to process information while driving, differences in reaction time to stimuli encountered all can be attributed to the segment level. It should be understood that factors at the segment level are likely to be correlated if they are attributes to a road use.

To better understand which segment predicted either above or below the average crash counts, it is appropriate to plot the segment ranked residuals with 95% confidence interval. It should be understood that these residuals represent segment departures from the overall mean of crash counts and therefore a segment with confident intervals which do not overlap the line representing the mean crash count across the segments is said to differ significantly from the average at the 5% level. As shown in Figure 17, at the right-hand side of the plot there are clusters of segments with above average crash counts and these were found to be of both types (EN-EX and EX-EN segment types). The graph can help in network scanning to unlock areas with unobserved factors leading to crash.

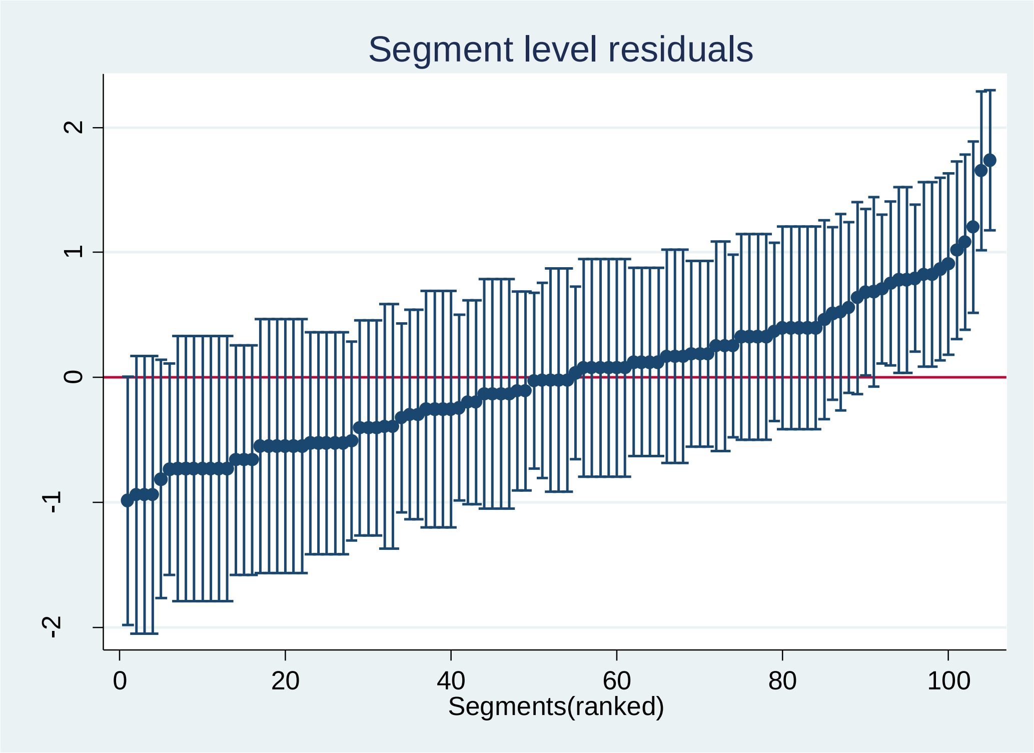

Figure 18 Variation in crash counts of clustered sections within each segment type

Since the major sources of heterogeneity in crash frequency are attributed to the influencing factors, it is appropriate to associate these sources as a function of explanatory variables. This can be achieved when the variation across the segments can be investigated as a function of segment type, an explanatory variable which is found significant in its variation across the freeway segments. Using Equation 3 on the note of table 7, it was found that EN-EX segment had the highest between segment variance (5.77%) compared to EX-EN segments although the fixed effects results indicated that EN-EX segments reduced crash counts. On one hand this result indicates the existence of underlying process generating unobserved factors linked to crash occurrence but on the other hand the existence of more geometric feature on weaving segments relative to non-weaving segments is likely to contribute to possible heterogeneity in crash frequency.

6.2 Accounting for Regression to the mean effect

The results discussed provide an opportunity to understand and quantify the effects of a combined effect of factors defining a freeway segments. These factors are observed to be geometric, traffic, human, and environmental factors. Within a typical segment these combined traits define a reference population on which inference can be done about the population of areas with the same characteristics. To appropriately characterize the combined effect on the crash frequency, expected values are desired rather than an individual effect of factors such as the shoulder.

Since we are dealing with a reference population, on average the expected crash frequency of that population is unknown but observed crash counts for a segment has shown to regress towards this mean leading to the regression to the mean bias in improving safety levels of these areas. To extend on the aforementioned discussion of interpretation of individual factor results, this section gives a background first on the concepts of regression to the mean and Empirical Bayes approach to solving regression to the mean bias. This is because the estimated results on Table 7 are based on what was observed after accounting for site heterogeneity and therefore cannot reflect the actual safety levels of a site.

6.2.1 Regression to the mean

Crash frequency occurrence in most locations has exhibited a tendency to fluctuate around unknown expected crash frequency for groups of locations with the same characteristics. Such tendency observed in crashes counted across groups of sites which are characterized by the same factors influencing crash occurrence, introduces regression to the mean bias into crash estimation and analysis. In the event that if sites are selected for treatment based on observed high crash frequency, then further problems leading to selection bias are likely to be encountered (Hauer, 1997; Sharma, 2006). Based on the effect of the regression to the mean, actual counts observed at a location in a given period do not reflect the actual expected crash counts.

Furthermore, estimated expected crash counts from the predicted models as one shown in Table 6 of this study reflects the expected counts of a given site based on what was observed and included in modeling process. It is based on these reasons, the true estimates of the expected crash frequency has to be adjusted to reflect actual counts observed as well the estimated expected crash frequency predicted by regression analysis techniques. Empirical Bayes approaches are applied to account for regression to the mean effect as explained in the following section.

6.2.2 Empirical Bayes Approach

Empirical Bayes methods are used to predict the expected crash frequency of a given site from a group of sites with similar characteristics by combining the actual observed crash counts and estimated expected crash counts obtained from the regression analysis (AASHTO, 2010). As shown in Table 6, this study estimated the expected crash counts of sites based on traffic and geometric characteristics. The following is the function in natural log form of the estimated crash frequency based on the fixed effect part only:

$$ln\mu = -5.477 + 1.291 * \log(flow) - 0.152 * shoulder - 0.69 * \log(length) - 0.324 * enex \quad (35)$$

Based on equation (35), expected crash frequency for each included in the model can be obtained. It should be understood that each set of combined characteristics determines an estimated expected crash frequency for a population of sites with similar characteristics from which each individual site's observed crash frequency has a tendency to regress to the actual expected crash counts in that group. To account for this effect, Empirical Bayes techniques require that, the estimated crash frequency of a site in a group of sites with similar characteristics is weighted by a linear combination of the observed and estimated crash counts by the following relationship:

$$E(\mu|\mu_s) = \alpha E(\mu) + (1 - \alpha)\mu_s \quad (36)$$

$$\alpha = \frac{1}{1 + \frac{var(\mu)}{E(\mu)}} \quad (37)$$

From equations (36) and (37), $E(\mu)$ is the expected crash counts obtained from equation (1) for every site, α is the weight used to determine proportions of crash counts and estimated crash counts in order to determine the actual counts expected.

6.2.3 Predicting site safety

The combined values of the estimated expected site crash frequency from the observed information and the actual crash counts observed in an analysis period characterizes safety levels of a site. This has been shown by the application of empirical Bayes approach. The importance of this method is that it draws information from what is being observed and partly from what is left out of the model reflected in the actual number of crashes observed at a site. This technique was applied on this study to estimate the expected crash frequency of a site because as indicated by the results on Table 7, site characterization cannot be determined based on what was observed only but also on the actual counts.

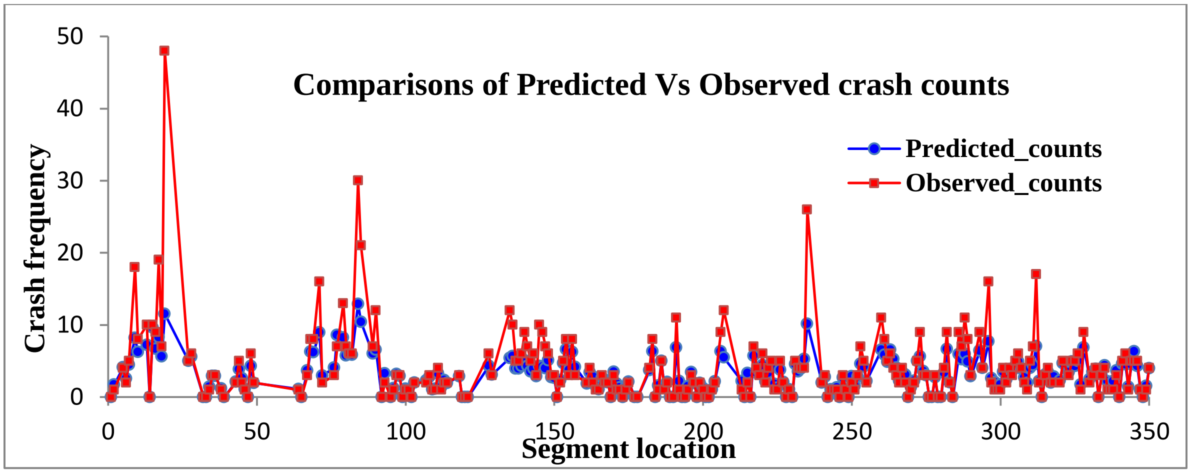

Figure 19: Comparison of observed Vs predicted crash frequencies

Figure 18 above shows the actual crash counts observed at a given site and expected crash frequency predicted by empirical Bayes approach which combines estimated information from the regression analysis and actual observed crash frequency. Based on these results it is clear that both observed crash counts and estimated expected crash frequency cannot be a good estimator of the actual expected levels of a site. As the figure shows, the actual number of counts observed at the site what found to be significant compared to the actual expected safety levels of a site given the observed crash counts.

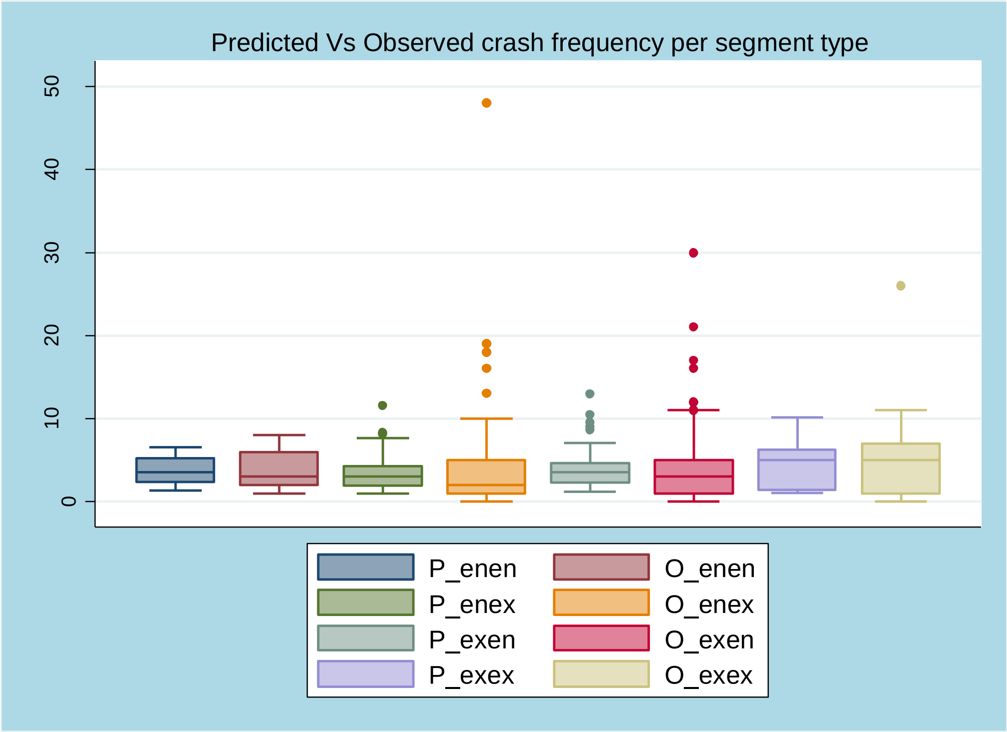

Figure 20: Predicted Vs Observed crash frequencies

The type of a segment also indicated variations between the predicted and observed crash counts. Figure 19 shows predicted crash counts paired with its corresponding observed crash counts for the same type of a segment. Starting from the left side of the box plot, the first box plots indicates freeway segments with entrance on both terminals of the segments, that is EN-EN. Approximately the estimated expected crash counts were found to be higher compared to the observed crash counts. The same trend was also observed in case of the weaving segments denoted as EN-EX which had the highest estimated crash counts compared to what was actually observed as well EX-EN. However, EX-EX segments had a different trend with approximately equal number of crash counts for both observed and predicted.

Discussion about the fixed effects results revealed more variation in the weaving segments compared to non-weaving segments. It was also shown that weaving segments reduced crash frequency compared to the non-weaving segments. By comparison with the results as indicated from Figure 19, the same trend can be depicted on average. However, other segments were compared in general while the box plot gives the details for every segment.

6.3 Summary

The purpose of chapter 6 was to predict safety of a site which in this case is the freeway segment. To accomplish this important step, it was important to account for a number of factors including assessing the appropriate structural functional forms of the safety predicting model, selecting parsimonious model, accounting for variability at the site and at the individual level of factors influencing site safety, and accounting for the regression to the mean bias.

The aforementioned issues are important towards characterizing site safety but also towards an important step of decision making to improve site safety through the appropriate selection of treatment methods and strategies. Structural form of the model was important because of violation of the distributional structural form across the factors and crash counts occurrence mechanisms. Specifically, structural form assessment involved covariate transformation to a function appropriate to realize its importance in influencing the mechanism of crash occurrence. This involved moving from original form of traffic flow and segment length to using natural logarithms.

Selecting the appropriate count model in association to the data structure involves assessment of the relative importance of random effects and issues of overdispersion arising through the use counts models. Overdispersion limits the applicability of traditional poison model in the event that an ancillary parameter is significantly leading to a two parameter model which is the negative binomial. Further selection is important when factors influencing crash occurrence cannot be included in the statistical model because they are unknown or cannot be collected. This helped to separate the appropriate models into two parts involving fixed effects and random effects where the latter accounts for unobserved heterogeneity.

Accounting for unobserved variability arising from unobserved factors is an important step towards gaining actual and realistic results of estimating the expected crash frequency of a site. However, due to the effects of regression to the mean, it is important to characterize site safety by accounting for both the predicted site safety as well as the actual counts observed at a given site. This brought in the use empirical Bayes approach which combines what was an estimate as an expected crash frequency and what was actually observed at the same site. The ultimate goal was to characterize the site in terms of the actual expected crash frequency given the observed counts at the site.

The use of predicted site safety levels is in aiding an engineer and decision maker to rank site according to their safety levels and prioritize which site to be treated. Furthermore, future evaluation can also use these results as before situation and evaluate the impact of treatment provided at the site in the after period by comparing site levels before and after the treatment has been applied at the site.

Issues of unobserved heterogeneity may also lead to another problem which is likely to violate classical statistical model when crash frequency observed from freeway segments sharing an arbitrary border exhibit correlation in space. Spatial autocorrelation has been observed to influence estimated results thereby introducing biasness into the regression coefficients and decision making based on these results are likely to lead to erroneous decisions. Chapter seven discusses the concepts and effects of spatial autocorrelation, also known as second order degree effects, in terms of quantitative results estimated.

CHAPTER 7 BAYESIAN SPATIAL MODEL RESULTS

7.1 Spatial autocorrelation

Spatial autocorrelation as aforementioned may be generated by spatial processes in which operation condition may likely cause freeway segments to influence each other in crash frequency occurrence by contagion. It may also arise due to misspecification of the model, an event which leaves spatially autocorrelated patterned information in the model residuals. Under these circumstances, it is important to test the existence of spatially autocorrelated crash frequency before making decisions to incorporated random effects in the modeling processes.

Spatial autocorrelation can be of two natures: Positive spatial autocorrelation is when similar values cluster together in a map. Similarity in this case means that high values of crash frequency cluster near high values of other crash frequencies or low values do cluster near other low values of the crash frequency. On the other hand, negative spatial autocorrelation is when dissimilar values cluster together in a map and by being dissimilar is an indication that clustering of crash frequency near to each other in space occur between high and low values of crash frequencies.

7.1.1 Global Moran's I Index

Based on the concepts mentioned above concerning spatial autocorrelation, this study tested the existence of spatial autocorrelation for contagious freeway segments using a Global Moran's I Index. An index is based on the hypothesis that crash frequencies observed on the freeway segments are randomly distributed and when its p-value is highly significant, the hypothesis can be reject. At this point either the crash counts exhibit positive or negative autocorrelation.

The test was run on ArcMap for the digitized freeway segments with crash frequency attributes and the results indicated Global Moran’s I Index equal to 0.162 with z-score = 2.06 and p-value = 0.04. It should be understood that a significant p-value with a positive z-statistic is an indication of the existence of spatial autocorrelation which implies that high crash frequency values tend to cluster near other high values of crash frequency.

The existence of the dependence in crash frequency across contagious freeway segments has an implication in the statistical modeling of crash frequencies. Specifically, this focuses on the assumptions of independence guiding distributional assumptions of statistical models. Under these models it is assumed that crash frequency observed on the freeway segments are randomly distributed across those segments. However, under these results it is evidence that when modeling freeway segments abutting each other, it is likely that spatial autocorrelation to occur. When terms recognizing such existence are not incorporated in the modeling process, the results may be biased because standard errors are likely to be inflated leading to wrong inferential statistics. Section 7.3 focuses on building a regression model in a Bayesian framework which incorporates terms modeling spatial autocorrelation based on Conditional Autoregressive (CAR) Models.

7.1.2 Local Moran’s I index

The above test conducted above shows a Global statistics which indicates in general the existence of spatially correlated crash frequencies for contagious freeway segments. To identify locations within a network where clusters are located, a local statistic can be of help. Getis-Ord

General G, is a local statistic index which measures the degree of clustering for either high values or low values of crash frequencies. Based on the criteria, a positive value of the index indicates that segments with high crash frequencies are close to segments with high crash frequency and vice versa. This implies that the segment under study is a cluster. A negative value of the Index indicates the existence of dissimilar values and they are identified as outliers. Figure 20 below displays those segment found significant with clustering behaviors.

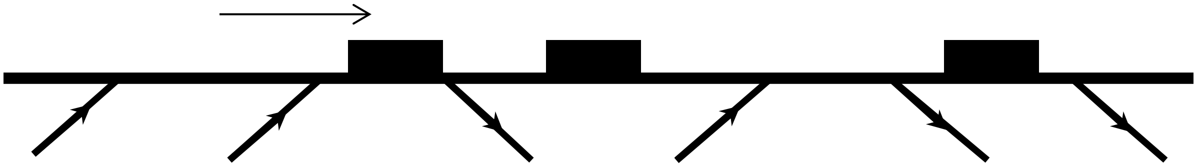

Figure 21: Identified freeway segments with crash frequency clusters

The notation indicated as "HH" in Figure 20 shows those segments identified to have high values (H) of crash frequency and surrounded by segments with high values (H) of crash frequency. Traffic flow is moving towards right from the left side and it can be evident that both types have identified as having clusters. However, by the concept and assumption used in this study, freeway weaving segments are likely to generate more vehicles to the proximate segments because of geometry. This means that at the on-ramp of a freeway, more vehicles are expected to enter the freeway and combining with through vehicles entering that segments, it is expected that more vehicle will occupy the segment closest to the weaving segments which in turn increases the likelihood of a crash to occur.

On the other hand, spill over caused by shock wave can also be a source of secondary crashes when congestion exists in the vicinity of the non-weaving segments. Based on these possible occurrences of both primary and secondary crashes, each segment is likely to affect each other segments with high impact expected to be seen for freeway segments in close proximity.

7.2 Modeling crash counts with spatial effects

Identification of the existence of spatial autocorrelation is an evidence of violation of the independent assumption under classical modeling. This means a spatial model has to be used in fitting crash counts with its associated influencing factors. Conditional autoregressive models are normally used and incorporate spatial effects autocorrelation parameters and spatial weight which characterizes spatial dependent for contagious freeway segments to predict expected levels of crash counts for a given site. The following subsection describes modeling procedures, assessment of the modeling results and their interpretation.

7.2.1 Model comparison and assessment

CARBayes package version 4.0 and WINBUGS version 1.4.3 were used in estimating the two models as shown on Table 1 below. To reduce autocorrelation of samples from the posterior distribution, the sequence was thinned by keeping every 10th simulation draw from each sequence. Furthermore the first 20,000 samples were discarded and concentrate on the last 80,000 samples to be able to diminish the influences of early iterations and achieve the target distribution. This implies that the final results are summarized from 8,000 drawn samples. To ensure that the chain's stationary distribution approximates the target distribution, the chain was monitored based on trace plots, historical plots of chain process as well as density plots of posterior means of the covariates and autocorrelation term. The final results include the trace plots for only the autocorrelation term as shown on Figure 21 below.

Based on the model fit criteria, the Spatial GLM Poisson model had a deviance Information criterion (DIC) equal to 262.12 which is small compared to a Non-spatial GLM model. This implies that the spatial model exhibited better fit to the data and therefore interpretation of the results is based on the spatial GLM Poisson model.

The final results of a Spatial GLM Poisson model contain posterior means of covariates and autocorrelation term. The significance of these terms is based on 95% credible intervals. When the 95% credible intervals includes zero, the corresponding factor is not significant at the 95% level and vice versa.

7.2.2 Model results and general overview

Table 8 below shows the estimated models which include a non-spatial GLM model in which spatial random effects are not included and a spatial GLM Poisson model where spatial effects are explicitly modeled by using a Conditional Autoregressive (CAR) model.

Table 8: Estimated Posterior means of covariates

Covariates	**Estimates (95% credible interval)**	
Dep. Var. = crash frequency	**Non-spatial GLM Poisson**	**Spatial GLM Poisson model**
Intercept	7.45 (6.78,8.06)	4.06 (0.40,7.12)
No. of lanes	-0.57 (-0.68,-0.45)	- 0.21 (-0.98,0.64)
Segment Length in mile	-8.03 (-8.10,-7.15)	- 4.33 (-9.13,-0.60)
No. of lanes * Seg. Length	1.99 (1.80,2.19)	1.10 (0.12,2.52)
Log (right shoulder)	-1.08 (-1.25,-0.92)	
Weaving segment	-0.03 (-0.14,0.07)	
τ^2		1.07 (0.63,1.94)
ρ		0.49 (0.08, 0.85)
DIC	1567.6	262.12

Comparing the two models it is clear that one can unlock some important features of the models. The first is the absolute magnitude of the regression coefficients. It can be observed that regression coefficients of a non-spatial model are large in absolute values compared to a spatial GLM model. This clearly indicates the existence of spatially autocorrelated features causing inflation of the regression coefficients. It should be understood that spatial correlation also represents spatial heterogeneity arising from the existence of unobserved factors which vary spatially for contiguous freeway segments. When these effects are not explicitly factored out of the model, it is likely for the model to suffer misspecification which leaves out patterned information in the model residuals.

7.2.3 Model assessment

Based on the results on Table 8, approximately 49% of crash frequencies across contiguous freeway segments are autocorrelated with a variance equals to 1.07. This result supports the aforementioned hypothesized situation that, there are spatial correlations of underlying processes generating crashes and these are likely to propagate across the adjacent segments. Most of the research activities analyze crash events on freeways based on the assumption that crash frequency observed on freeway segments are independent. This results lead to biased estimates if spatial effects are not included in the modeling processes.

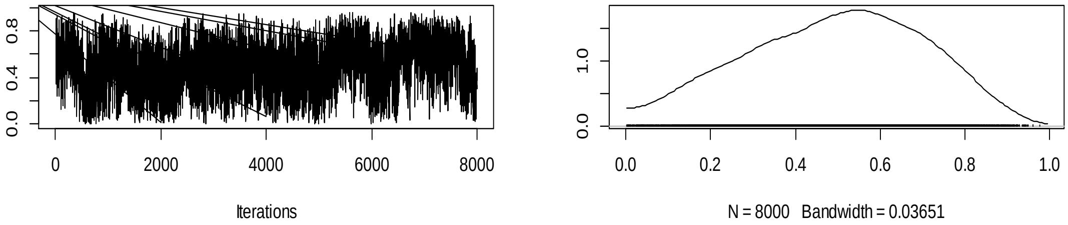

Figure 22: Posterior distribution of spatial correlation parameter

7.2.4 Model parameters estimated

The following discussion is based on the incident rate ratios which are the exponentiated posterior means. In addition to the aforementioned findings, we also investigated the impact of geometric elements on the crash frequency after controlling for the existence of spatial effects. As the results showed, only the number of lanes was not significant compared to the segment lengths. However, since the interaction term was statistically significant and it was created from the main effects – number of lanes and segment length -, number of lanes is retained in the model to better interpret the impacts of interactions across the segment length as the number of facility lanes changes in addition to the main effects.

7.2.4.1 Marginal effects

The impact of the number of lanes and segment lengths are better interpreted based on their marginal impacts as shown on figure 20 below including its corresponding marginal effect function. Including an interaction term is based on the fact that the influence of the longitudinal space depends on the transverse space available to accommodate the number of vehicles available. The results displayed on Table 7 (spatial GLM model) can be reproduced in terms of the regression equation as follows:

$$ln\hat{\mu} = 4.06 - 0.21 * lanes - 4.33 * length + 1.1(length * lanes) \quad (38)$$

Based on calculus techniques, the marginal effects of an increase in segment length can be computed using partial derivatives. Since the variables involved are regarded as continuous variables, the marginal effects in this case measure the instantaneous rate of change of one variable as the other factor is increased by a small amount. Applying calculus principle, instantaneous rate of change of the crash frequency with respect to the segment length is equal to the slope of the regression function displayed in equation 38 which in turn equal to the partial derivative of the function at the given length. Based on these principles, the following equation can be obtained:

$$\frac{\partial(\hat{\mu})}{\partial(length)} = (1.1 * lanes - 4.33)\exp\{4.06 - 0.21 * lanes - 4.33 * length + 1.1(length * lanes)\} \quad (39)$$

Equation 39 can be used to predict on instantaneous basis the estimated expected crash frequency at an instantaneous location along the freeway segments as vehicles moves for a given segments with fixed number of lanes. This is an important equation since it may be possible to unlock potential locations where crash counts are likely to cluster (that is high values) as one travels longitudinally for a fixed number of travel lanes.

As shown on Figure 20, the expected crash frequency increases with longitudinal space for all segments with number of travel lanes observed. It should be understood that we interpret number of lanes as representing width of freeway in a transversal dimension. Possible explanation for this trend is that an increase in segment length provides an opportunity for drivers to maneuver because spaces for such behavior are available. Maneuverability may include lane changes and speeding behavior which are likely to results in crashes.

Although, all the segments showed approximately the same trend in influencing crash occurrence, a closer look at the displayed plots indicates differences in the intensity of impacts across segments with different number of lanes. For the same segment length, the marginal impacts of segments with two and three lanes is high compared to those segments with number of lanes greater than three.

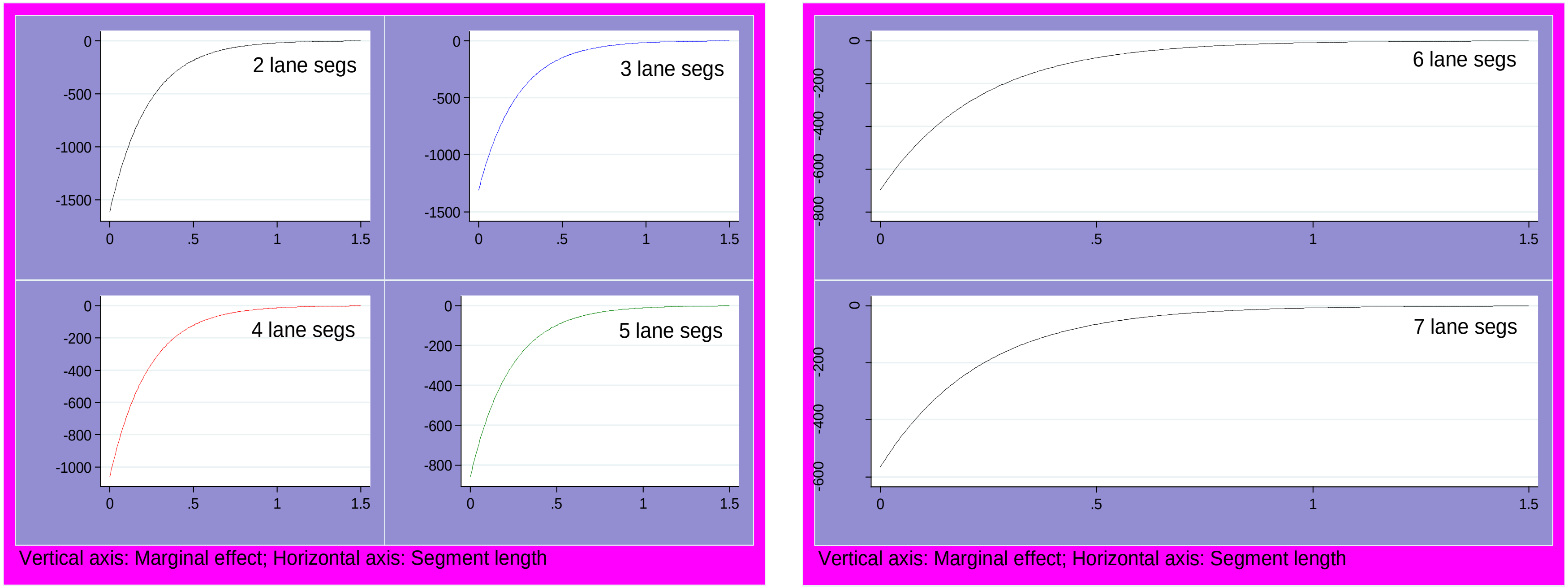

Figure 23: Marginal effects of number of lanes on crash frequency

Note:

$$\frac{\partial(\hat{\mu})}{\partial(length)} = (1.1 * lanes - 4.33)\exp\{4.06 - 0.21 * lanes - 4.33 * length\}$$

Marginal effects across different number of lanes for an arbitrary segment of length of 1.5 mile

No. of lanes	2	3	4	5	6	7
Marg. Effect	-185.14	-150.07	-121.65	-98.61	-79.93	-64.79

The above discussion on the plots can be clearly seen on the inset table displayed below Figure 22. A fixed length of 1.5 mile was used to compute the marginal impact across segments of different number of lanes. From the results on the table, segments with more lanes had little impact on reducing the expected frequency. A sensitivity analysis was conducted on different segment lengths and the results were approximately of the same magnitude as the ones displayed in the inset table on Figure 22.

7.3 Summary

The results discussed above on spatial models are mostly centered at solving issues related to spatial autocorrelation in crash frequency data. When values of crash counts are correlated in space for abutting freeway segments, an independent assumption from which classical count models depend is violated and the results obtained are likely to be biased and inconsistent. The existence of spatial autocorrelation can be investigated using an index known as Moran's I Index, a spatial statistical index, which identifies the existence of dependence in values of crash frequency for contagious freeway segments.

The identification of the existence of spatial autocorrelation necessitates inclusion of spatial effects to capture all combined effects which are believed to influence crashes in the same trend but cannot be observed. The included spatial effect terms are modeled by spatial models known as Conditional Autoregressive (CAR) models which include spatial weights defining spatial relationship of contagious segments and spatial autocorrelation terms for capturing the combined effects of unobserved factors identified by a spatial index aforementioned.

Under this section it is also possible to further identify clusters of crash frequency based on Anselin Moran's Index. This is an index used to identify the strength of spatial association and in this case it is a locally based index compared to the Global Moran's index discussed earlier. Clusters as shown on Figure 18 imply that high values of crash frequency in an identified site are surrounded by the freeway segments with high values of crash frequency. In terms of spatial statistics terms this is known as positive spatial autocorrelation implying that there are spatial mechanisms generating crash counts trending in the same direction.

CHAPTER 8 CONCLUSIONS AND RECOMMENDATION

8.0 Multilevel modeling

8.0.1 General model purpose and summarized findings

The purpose of the study in chapter 6 was to analyze the impact of geometric and traffic flow characteristics on freeway crashes while accounting for the effect of unobserved factors, also known as unobserved heterogeneity, which are likely to influence crash occurrence. The investigation was also motivated by the fact that the distribution of crashes in space is not limited to only the influence areas of the divergence and convergence segments as well as weaving segments. Areas beyond the influence areas were observed to have crashes occurred and by including these areas it was possible to cluster data within the weaving and non-weaving segments to quantify the variability of unobserved factors through the variance of random parameters reflected in the group-effect heterogeneity and individual-effect heterogeneity using multilevel count models.

The group-effect heterogeneity considers variability at the cluster level and in this research a cluster represents a freeway segment. All factors believed to have generated crashes but are either not known or cannot be easily available to be modeled are reflected in the modeling process by allowing the model intercept to vary across the freeway segments. By doing this, it is possible to capture the effect of unobserved factors which in this case are referred to as group-effect heterogeneity. To realize the impact of factors observed and included in the model, these group-effects are integrated out and summarized by producing a variance which is an indication of their existence and are quantified by the variance parameter.

Individual–effect heterogeneity considers the impact of individual variables variability. In this case the main idea is that a single factor such as the number of lanes on the freeway cannot have the same impact across freeway segments. To clarify these concerns consider two segments with one having four lanes and another one having 2 lanes. It is evident that a segment with four lanes is likely to have a different impact on crash occurrence experience when compared to a freeway segment with two lanes. The purpose of accounting for individual-effect is to realize such difference within a factor by allowing the slope which is a regression coefficient on that factor to vary across the freeway segments. The differences obtained are summarized and a variance is an output which represents individual effect-heterogeneity.

The aforementioned concepts are possible if we consider a freeway segment as a cluster and that cluster is being divided into small sections, the purpose being to capture such variability at a micro-level. Information are then collected at the small section level which in this study included traffic characteristics from installed detectors on freeway network. At the cluster level, geometric characteristics are collected and these included the number of lanes, median and right shoulder, segment length and the type of a segment. Model results contains two parts: the fixed effect part which shows the impact of geometric and traffic characteristics and the random part which gives summaries, in terms of variances, of all factors which were not included in the model but may have an impact on crashes and variability reflected in the differences within a single factor.

The results of the fixed-effects part indicated that more vehicles increased crash frequency while wider right shoulder, segment length decreased crash frequency during the analysis period. It was also revealed that weaving segments decreased crash frequency compared to non-weaving segments.

The results of the random-effect part indicated that 13.9% of the variation in crash frequency is unaccounted for which is an indication of the existence of unobserved factors influencing the occurrence of crashes. This variance represents group-effect heterogeneity explained earlier where all factors attributed to a given segment, here referred to as a cluster, are integrated out and summarized by outputting a variance. Although what specifically contributes to having a variance of 13.9% is not known, this results represents important alert to safety analyst professionals that there are more information leading to crashes but are either not known or cannot be easily obtained at the moment. Further research with different designs and modeling approach are required on continual basis to learn more and understand more factors leading to the occurrence of crashes.

It was further revealed that weaving segments (EN-EX) had the highest between segment variance compared to non-weaving segments. According to the earlier explanation this represents individual-effect heterogeneity and the model results indicated that 19.6% of the variation is between the weaving segments. Although other factors did not have a significant variance, this result narrows down location on freeway network where variability in crash occurrence is expected compared to other locations namely non-weaving segments.

8.0.2 Recommendation for application

An important step is the guidance on how findings obtained from this study can be used to improve road safety. One application is that, the model framework gives an engineer, the cluster of the actual segments with above average crash frequency (figure 14). This is possible because the model setting includes random effects part and it is under this part all factors generating crashes but are unobserved are available for every freeway segment. These group-effects are ranked and drawn on a caterpillar plot and a segment with the above or below average crash frequency can be identified. Spotting out these segments is a way of screening the network to find locations with need for further research and improvement based on the existence of unobserved factors. It is also intuitive that as we narrow down our investigation and improvement more problematic locations, we are exercising efficient allocation of resources in the sense that improvement resource can be allocated to areas where there is a need.

The second application of the model concerns before-after studies in road safety of actual roadway sections where the geometric elements were changed. This study considered the analysis period to be 2013 and therefore the findings provide an estimated model of what safety levels was if changes in geometric elements were made prior to the analysis period. This comprises an after study results. Data prior to the analysis period can be used to obtain a prediction model and through extrapolation, the model predicts what would have been the safety levels in the absence of changes in the geometric elements (Hauer, 1997). The difference of safety levels before and after studies indicates improvement. This was the essence of incorporating empirical Bayes approach which in return helps getting more realistic results to characterize safety levels of freeway segments.

The sign and direction of improvement is a function of specific changes made. For instance, an additional lane on the main facility between on-and off-ramps is likely to indicate better positive safety levels due to an added freeway capacity expected. Negative safety levels may result from narrowing a geometric element such as a shoulder for the purpose of adding high occupancy vehicle lanes. It is therefore important to conduct before-after studies to quantify actual levels of safety once changes are made in a network.

8.0.3 Conclusion

Incorporating items which accounts for unobserved factors in count models has recently been of great interest to transportation practitioners (Mannering & Bhat, 2014). By allowing parameters to vary across segments it is possible to capture and quantify unobserved factors. Ignoring these factors results in biased coefficients in a multilevel settings because the estimate of the standard errors will be wrong (Bristol, 2014). Unobserved factors are the results of a number of underlying processes both in space and time. Spatial phenomena behavior across freeway segments can results in spatial effects (both first and second order effects) leading to variation in the mean crash frequency of the process in space of spatial autocorrelation in the process. Spatial autocorrelation is the tendency for deviations in values of the process from its mean to follow each other in neighboring sites (Anders & Sophia, 2012). A natural extension of this study is to include spatial effects in modeling crash frequency in addition to geometric features.

Previous research has shown to improve on applying methodologies which helps to understand factors contributing to crash occurrence while leaving out those factors unavailable or cannot be corrected at the moment. However, the main purpose of all of these approaches is trying to explain variability in the crash frequency through the variability of influencing factors. It is under this approach that current research hasn't documented at a micro-level all sources of variability in crash frequency. Literature reviews indicate the existence of limitations in account for the possible sources at all levels of the model. This implies that data structure which is currently being employed in multilevel modeling do not entirely capture these sources leading to unobserved heterogeneity all levels.

These important methodological barriers which remain in the statistical analysis of crash data was the essence of my research. To address issues of unobserved heterogeneity, one needs to capture more variation at a small level within the road network in order to account for the most possible source of variation. However, research already conducted had limitations in data structure set up.

For instance, (Venkataraman, *et al*, 2011) used segments at the interchange and noninterchange levels but accounted only variation within the explanatory variables by estimating random parameters which correpond to individual-effets heterogeneity in this research. Furthermore, Interchange segment were defined by the farthest merge and diverge ramp limits for each direction which implies that a segment contained more than one interchange. Noninterchange segments were defined as a cintinuous travel segment between two interchanges. By including only parameters to capture differences across individual factors leaves out variability leading to group-effect heterogeneity and it is at this point my research differes with the current practice. Specifically, my research is at the a micro-level where a freeway segment is further subdivided into small segments to aid in capturing more variations across the segments and within the sub-sections.

Furthermore, (Anastasopoulos *et al*, 2009) used segments with homogeneous characteristics but accounted for random parameters only. However, even if a segment is homogeneous, there are still differences in terms of unobserved factors across these segments which need to be accounted for but were not captured. This could be possible by allowing an intercept to vary across the segments. Other research activity, such as (Dinu *et al*, 2011) divided the freeway into homogeneous sections based on traffic volume, carriage way width, and shoulder width but only random parameters were estimated which in turn leaves variability at the segement level un accounted for. Compared to the se research activities, my research has addressed possible variability in crash frequency at a micro-level by considering a freeway segment divided into small sections. Furthermore, variability at all levels are captured by introducing group-effects heterogeneity and individual-effects heterogeneity.

8.1 Bayesian spatial modeling

8.1.1 General model purpose and summarized findings

The purpose of study reflected in the results represented in chapter 7 was to account for the effects of spatial autocorrelation, also known as second order spatial effects. Spatial autocorrelation traditionally compares values of a variable assigned to areas and determines if the distribution of the variable is random, clustered, or uniform. It is one of the components leading to unobserved heterogeneity being referred to as spatial heterogeneity. If these effects are not included in quantifying the impacts of influencing factors, the results obtained are likely to be biased.

This study investigated the existence of second order spatial effects for contiguous freeway segments. This was motivated by the fact that there are spatial correlations of underlying processes generating crashes and these are likely to propagate across the adjacent segments. Furthermore, there are interdependences across freeway characteristics in influencing crash occurrence which means the effect of one of the geometric elements depend on other geometric element. Including interaction terms in modeling reduces the impact of unobserved heterogeneity because it accounts for effect modification resulting from modification variables.

The existence of spatial autocorrelation was tested by computing a Global Moran’s Index which determines the possible existence of spatial pattern in values of the crash frequency assigned to contagious freeway segments. Positive autocorrelation was found which implies that there exists crash frequency clusters for abutting freeway segments trending in approximately the same direction. By the same directions it means, high values of crash frequency are likely to be found near other high values of crash frequency for freeway segments which share border.

Furthermore, a local index was also found to understand and located locations on a network where such behavior is highly significant. The results showed at a local level, the existence of these clusters and an interesting part at this point was to know which type of segments which high values borders the other. Under these result, all segment type was identified to abut each other which means no particular segment is safer than the other.

Based on the main effects of the spatial GLM model which is reflected on the regression coefficients on the number of lanes and segment length, it was found that more lanes on a given segment and long segments reduced the level of crash frequency during the analysis period. Longitudinal space which is reflected in the segment length exhibited differential impact across long freeway segments. This differential impact can be explained by the existence of differential transverse space reflected in the number of travel lanes for long segments. This implies that interacting influencing factors helps to reduce the impact of unobserved modifiers leading to biasing results. It is imperative that safety modeling include terms which explain any general forms of unobserved heterogeneity. This helps to come up with actual impacts of the influencing factors retained in the model.

8.1.2 Recommendation for application

The developed model can be applied as a discriminant model. This is based on the fact that spatial effects terms are included in the modeling process. Theories on the estimation process require integrating out these effects and summarize them in terms of variance, a method which leaves out the actual influences of the remaining factors. Based on figure 19 a researcher can point out locations on the freeway network from which its factors exhibited more impacts on the crash frequency. For instance, with these results segments with four and five lanes increased crash frequency compared to those with fewer number of lanes. These lanes require further investigation to be able to understand causes of crash frequency occurrence.

Another important application is based on the natural interpretation of most of the regression coefficients. Negative coefficients in most of the cases means the corresponding factors had a negative impact and therefore by increasing those factors help to reduce more crash frequency on freeways. For instance, the results for long segments with few numbers of lanes indicated that by increasing the longitudinal space we are able to reduce the number of crash frequency experienced. This is counter-intuiting with segments of the same length but have more lanes. This means we cannot adopt an alternative of increasing segment length on these locations. This helps to narrow down countermeasures alternative and focus on Intelligent Transportation Systems designed for safety purposes.

Furthermore, as detailed out on chapter 6 concerning issues arising from regression to the mean, it is possible under these results to account for the effect of regression to the mean by applying empirical Bayes approach to the results of spatial GLM Poisson model obtained in order to characterize the actual safety of a freeway given the actual crash counts observed at a site. Based on these results, decision making can be made on which locations can be prioritized for proper treatment.

8.1.3 Conclusion

The main focus of chapter 6 was to analyze safety by accounting for the effects of spatial autocorrelation for freeway segments which share arbitrary border. This was motivated by the fact that there are underlying processes from one area which are likely to affect another area there by generating crashes. An example to clarify this fact is when vehicles from one area spill over to another area in close proximity. This may happen when a road network is highly congested which in turn can generate secondary crashes.

Furthermore, there exists unobserved factors which are likely to propagate in space or time to abutting freeway segments leading to approximately equal pattern of crashes to the next area. An example is the pavement condition in close proximity of the abutting segments. If such defects occupy in both areas and it has known to cause crashes, this means the two areas will have the same factor generating crashes although the rate of cause may be different. The existing of spatial autocorrelation limits the applicability of statistical models which are based on the assumption of independent across crash counts. Solution to this challenging problem is to incorporate random effects terms which in turn help to capture spatial heterogeneity reflected in the spatial autocorrelation terms. Conditional Autoregressive (CAR) models have been design to be used when spatial autocorrelation is detected in crash frequency.

8.2 Model transferability

Model transferability refers to the adaptability of a developed model(s) from one region to be applied to another region (jurisdiction). This implies that safety performance functions can be imported between jurisdictions or between time period if the analysis differs and it can be done through model calibration process. For the results obtained in this study to be adapted to other regions with different features compared to the local roadway and traffic features used, a calibration process known as Bayesian model averaging (Chen, Persaud, & Sacchi, 2012) is proposed because it addresses model uncertainity in which no reasonable safety performance function would be discarded. Although this approach was applied on signalized intersections, the main focus is on calibration methodology principles used rather than the specific area used.

8.3 Future research need

This research has identified a number of issues surrounding the analysis of safety in road networks. The main focus of the whole research was to conduct safety study while accounting for factors which bias the end results. The main two issues discussed in this study was the unobserved heterogeneity reflected in the absence or unknown factors which are ultimately not included in modeling process and spatial heterogeneity issues which are reflected when values of crash frequency from areas sharing an arbitrary border exhibits spatial autocorrelation.

The results of the former have indicated the existence of unobserved heterogeneity on freeway segments. This implies that many factors are involved in the crash generating mechanisms but are unknown or cannot be collected although they might be known. An example is information processing capability which requires a road user to quickly be able to screen information while moving on the network to be able to avoid collisions. This factor is not easy to measure at the time a crash has occurred but also even if it was the main cause, chances are that it may not properly be recorded as the main cause of a crash. Under this situation, more research is required to account for these factors in order to clearly understand crash generating mechanisms and device proper method of improving our networks.

Furthermore, the results from the spatial GLM Poisson study have indicated the existence of spatial autocorrelation across contiguous freeway segments. This implies that segments with spatial proximity constitute traffic and geometric characteristics which influence crash occurrence in a similar trend. This phenomenon violates the distribution assumption of Poisson process under which crash events occur. My future research will involve simultaneity treatment of freeway segments in analyzing safety effects of factors which are believed to influence crash occurrence. Simultaneity behavior of observation unit can be incorporated in safety analysis by using simultaneous equation models which are the special case of the general structural equation models. In addition to solving the aforementioned problem of distributional assumption, general unobserved heterogeneity terms can further be incorporated to account for random effects.

An alternative approach to extend the aforementioned techniques which are based on maximum likelihood estimation approach is modeling crash frequency based on a model which does not assume any probability distribution. Generalized method of moments (GMM) (Hansen, 1982) derives estimators which use assumptions about the moments of the random variables to derive an objective function. The assumed moments of the random variables provide population moment conditions from which parameters estimates can be obtained by finding the parameters that make the sample moment conditions as true as possible. When compared to maximum likelihood estimation approach, generalized method of moments (GMM) is characterized by the fact that many estimators can be viewed as special cases of GMM. Furthermore, GMM techniques are appropriate where a likelihood analysis is difficult.

APPENDIX

Table 9 Model set #1: Untransformed Traffic volume and segment base length

(Dep. var = crash frequency)	PO		ME PO		NB		ME NB	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Traffic volume	0.000	10.43	0.000	2.18	0.000	5.57	0.000	2.95
Right shoulder	-0.193	-7.13	-0.085	-1.93	-0.176	-3.50	-0.146	-2.70
Seg. base length	-0.000	-2.62	-0.000	-1.88	-0.000	-1.58	-0.000	-1.97
Weaving segments (enex seg.)	-0.190	-2.65	-0.341	-2.47	-0.179	-1.29	-0.289	-1.83
Intercept	-1.998	-14.76	1.604	6.37	1.592	5.35	1.754	5.76
Covariance parameters								
Variance - Intercept (Std. err.)				0.309 (0.172)				0.206 (0.105)
Variance - enex seg. (Std. err.)				0.412 (0.099)				0.193 (0.187)
Model fit criteria								
lnL		-831.014		-655.203		-620.319		-614.043
AIC		1672.028		1324.407		1252.64		1244.09
BIC		1689.831		1349.331		1274.00		1272.57
Alpha (α)						0.667		0.370
LR test for α = 0						421.39 (0.000)		-4.06 (0.000)
LR test Vs PO & NB (p-value)				351.62 (0.000)				15.55 (0.0002)

Table 10 Model set #2: Untransformed Traffic volume and log (segment base length)

(Dep. var = crash frequency)	PO		ME PO		NB		ME NB	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Traffic volume	0.000	10.64	0.000	2.18	0.000	5.63	0.000	2.93
Right shoulder	-0.194	-7.15	-0.088	-1.98	-0.180	-3.54	-0.151	-2.76
Log (base length)	-0.410	-2.38	-0.546	-1.86	-0.524	-1.49	-0.765	-2.06
Weaving segments (enex seg.)	-0.207	-2.94	-0.366	-2.73	-0.195	-1.41	-0.308	-1.99
Intercept	3.018	5.68	2.973	3.24	2.929	2.65	3.720	3.19
Covariance parameters								
Variance - Intercept (Std. err.)				0.288 (0.166)				0.214 (0.106)
Variance - enex seg. (Std. err.)				0.421 (0.101)				0.185 (0.187)
Model fit criteria								
lnL		-831.748		-655.31		-620.443		-613.968
AIC		1673.496		1324.61		1252.89		1243.94
BIC		1691.300		1349.54		1274.25		1272.42
Alpha (α)						0.669		0.368
LR test for α = 0						422.61 (0.000)		-4.10 (0.000)
LR test Vs PO & NB (p-value)				352.88 (0.000)				12.95 (0.005)

Table 11 Model set #3: log (Traffic volume) and untransformed segment base length

(Dep. var = crash frequency)	PO		ME PO		NB		ME NB	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Log (Traffic volume)	1.723	12.76	0.727	2.90	1.623	6.30	1.291	4.30
Right shoulder	-0.172	-6.43	-0.091	-2.11	-0.180	-3.61	-0.147	-2.80
Base length	-0.000	-1.85	-0.546	-1.86	-0.000	-1.48	-0.000	-1.78
Weaving segments (enex seg.)	-0.365	-3.30	-0.361	-2.63	-0.183	-1.31	-0.308	-1.97
Intercept	-10.410	-10.11	-3.472	-1.90	-9.531	-4.92	-7.252	-3.27
Covariance parameters								
Variance - Intercept (Std. err.)				0.318 (0.166)				0.174 (0.087)
Variance - enex seg. (Std. err.)				0.357 (0.090)				0.197 (0.174)
Model fit criteria								
lnL		-792.000		-653.542		-618.500		-610.90
AIC		1594.000		1321.08		1248.99		1237.80
BIC		1611.801		1346.01		1270.36		1266.29
Alpha (α)						0.636		0.364
LR test for α = 0						347.01 (0.000)		-4.39 (0.000)
LR test Vs PO & NB (p-value)				276.91 (0.000)				15.19 (0.000)

Table 12 Model set #4: log (Traffic volume) and log (segment base length)

(Dep. var = crash frequency)	**PO**		**ME PO**		**NB**		**ME NB**	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Log (Traffic volume)	1.737	12.91	0.732	2.92	1.634	6.37	1.291	4.30
Right shoulder	-1.172	-6.41	-0.094	-2.16	-0.183	-3.65	-0.152	-2.86
Log (Base length)	-0.155	-0.87	-0.534	-1.85	-0.498	-1.44	-0.690	-1.88
Weaving segments (enex seg.)	-0.243	-3.48	-0.385	-2.89	-0.200	-1.43	-0.324	-2.11
Intercept	-10.111	-7.91	-2.167	-1.03	-8.334	-3.55	-5.48	-2.07
Covariance parameters								
Variance - Intercept (Std. err.)				0.363 (0.166)				0.179 (0.088)
Variance - enex seg. (Std. err.)				0.301 (0.162)				0.191 (0.173)
Model fit criteria								
lnL		-792.182		-653.604		-618.555		-610.76
AIC		1594.363		1321.21		1249.11		1237.53
BIC		1612.167		1346.13		1270.48		1266.01
Alpha (α)						0.639		0.362
LR test for α = 0						347.25 (0.000)		-4.42 (0.000)
LR test Vs PO & NB (p-value)				277.15 (0.000)				15.58 (0.000)

Table 13 Model set #5: Traffic volume in 10 mill vehicles and untransformed segment length

(Dep. var = crash frequency)	**PO**		**ME PO**		**NB**		**ME NB**	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Traffic volume in 10millth	0.150	10.43	0.080	2.18	0.316	5.57	0.181	2.95
Right shoulder	-0.193	-7.13	-0.085	-1.93	-0.176	-3.50	-0.145	-2.70
Base length	-0.001	-2.62	-0.000	-1.88	-0.000	-1.58	-0.000	-1.97
Weaving segments (enex seg.)	-0.190	-2.65	-0.341	-2.47	-0.179	-1.27	-0.289	-1.83
Intercept	2.000	14.76	1.604	6.37	1.592	5.35	-1.754	5.76
Covariance parameters								
Variance - Intercept (Std. err.)				0.412 (0.100)				0.193 (0.186)
Variance - enex seg. (Std. err.)				0.309 (0.172)				0.206 (0.105)
Model fit criteria								
lnL		-831.014		-655.203		-620.319		-614.04
AIC		1672.03		1324.41		1252.64		1244.06
BIC		1689.831		1349.33		1274.00		1272.57
Alpha (α)						0.667		0.370
LR test for α = 0						421.39 (0.000)		-4.06 (0.000)
LR test Vs PO & NB (p-value)				351.62 (0.000)				12.55 (0.002)

Table 14 Model set #6: Traffic volume in 10 mill. Vehicles and log (segment base length)

(Dep. var = crash frequency)	**PO**		**ME PO**		**NB**		**ME NB**	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Traffic volume in 10millth	0.152	10.64	0.080	2.18	0.318	5.63	0.180	2.93
Right shoulder	-0.194	-7.15	-0.088	-1.98	-0.180	-3.54	-0.151	-2.04
Log (Base length)	-0.410	-2.38	-0.546	-1.86	-0.524	-1.49	-0.765	-2.04
Weaving segments (enex seg.)	-0.207	-2.94	-0.366	-2.73	-0.195	-1.41	-0.308	-1.99
Intercept	2.018	5.68	2.973	3.24	2.929	2.65	-3.720	3.19
Covariance parameters								
Variance - Intercept (Std. err.)				0.421 (0.101)				0.214 (0.106)
Variance - enex seg. (Std. err.)				0.288 (0.166)				0.185 (0.185)
Model fit criteria								
lnL		-831.748		-655.306		-620.442		-613.97
AIC		1673.50		1324.61		1252.88		1243.94
BIC		1691.30		1349.54		1274.25		1272.42
Alpha (α)						0.669		0.368
LR test for α = 0						422.61 (0.000)		-4.10 (0.000)
LR test Vs PO & NB (p-value)				352.88 (0.000)				12.95 (0.002)

Table 15: Model set #7: log (Traffic volume) and short segment base length (< 0.5 mile)

	PO		ME PO		NB		ME NB	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Log of traffic volume	1.679	13.05	0.756	3.03	1.642	6.54	1.32	4.43
Right shoulder	-0.180	-6.68	-0.101	-2.32	-0.193	-3.87	-0.16	-3.03
Short base length	0.232	3.31	0.282	2.55	0.321	2.26	0.38	2.56
Weaving segments (enex seg.)	-0.239	-3.64	-0.427	-3.37	-0.241	-1.93	-0.37	-2.61
Intercept	-10.191	-10.71	-3.952	-2.20	-9.900	-5.36	-7.74	-3.60
Covariance parameters								
Variance - Intercept (Std. err.)				0.259 (0.154)				0.139 (0.165)
Variance - enex seg. (Std. err.)				0.367 (0.093)				0.196 (0.092)
Model fit criteria								
lnL		-787.234		-652.119		-617.019		-609.246
AIC				1318.239				1234.493
BIC				1343.163				1262.978
Alpha (α)						0.626		0.353
LR test for α = 0						340.43 (0.000)		-4.51 (0.000)
LR test Vs PO & NB (p-value)				270.23 (0.000)				15.55 (0.0002)

Table 16: Model set #8: Traffic volume and short segment base length (< 0.5 mile)

	PO		ME PO		NB		ME NB	
Fixed effects parameters	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**	**Coef.**	**Stat.**
Traffic volume	0.000	10.80	0.000	2.20	0.000	5.74	0.00	2.96
Right shoulder	-0.199	-7.31	-0.093	-2.10	-0.189	-3.73	-0.16	-2.91
Short base length	0.279	3.97	0.276	2.44	0.313	2.16	0.39	2.56
Weaving segments (enex seg.)	-0.234	-3.52	-0.408	-3.20	-0.247	-1.95	-0.36	-2.51
Intercept	-1.762	-16.37	1.323	6.76	1.342	5.61	1.41	5.76
Covariance parameters								
Variance - Intercept (Std. err.)				0.433 (0.102)				0.243 (0.115)
Variance - enex seg. (Std. err.)				0.243 (0.157)				0.118 (0.177)
Model fit criteria								
lnL		-826.921		-654.088		-619.206		-612.871
AIC				1322.176				1241.562
BIC				1347.101				1270.047
Alpha (α)						0.659		0.357
LR test for α = 0						415.43 (0.000)		12.85 (0.005)
LR test Vs PO & NB (p-value)				345.67 (0.000)				15.55 (0.0002)

REFERENCES

AASHTO. (2010). *Highway Safety Manual, First Edition, Volume 2*. Washington, DC: American Association of State Highway and Transportation Officials (AASHTO).

AASHTO. (2011). *A Policy on Geometric Design of Highways and Streets, 6th Edition.* Washington, DC: American Association of State Highways and Transportation Officials.

Abdel-Aty, M., & Huang, H. (2010). Multilevel data and Bayesian analysis in traffic safety. *Accident Analysis and Prevention*, Vol. 42, pp. 1556–1565.

Abdel-Aty, M., & Huang, H. (2013). Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis and Prevention*, 371-376.

Agresti, A. (2002). *Categorical Data Analysis, 2nd Edition.* John Willey & Sons, Inc.

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

Aguero-Valverde, J. (2014). Direct Spatial Correlation in Crash Frequency Models: Estimation of the Effective Range. *Journal of Transportation Safety & Security*, 6:1, 21-33, DOI: 10.1080/19439962.2013.799108.

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

Anders, S., & Sophia, R. (2012). *Multilevel and Longitudinal modeling using Stata, 3rd Edition.* Texas, College station: Stata Press.

Arthur, R. M. (2015, 05 20). *Spatial and Temporal Analysis of Traffic collisions (Order No. NQ87017).* Retrieved from ProQuest Dissertations & Theses. (305346803).: http://ezproxy.library.unlv.edu/login?url=http://search.proquest.com/docview/3

Black, R. W., & Thomas, I. (1998). Accidents on Belgium's motorways: a network autocorrelation analysis. *Journal of Transport Geography*, Vol. 6, No.1, pp. 23-31.

Bolstad, P. (2005). *GIS Fundamentals: A First Text on Geographic Information Systems, Second Edition.* Eider Press.

Bonneson, A. J., Geedipaly, S., & Pratt, P. J. (2014, November 03). *Safety Prediction Methodology and Analysis Tool for Freeways and Interchanges.* Retrieved from Google sites: https://sites.google.com/site/jbreportsandtools2/home/reports/1745

Bristol, U. o. (2014, July 01). *Centre for multilevel modeling*. Retrieved from www.bristol.ac.uk

Cameron, C. A., & Trivedi, K. P. (2013). *Regression analysis of count data, 2nd Edition.* Cambridge University Press.

Centre for Multilevel modeling. (2014, July 01). Retrieved from University of Bristol: www.bristol.ac.uk

Chen, H., Liu, P., Lu, J. J., & Behzad, B. (2009). Evaluating the Safety impacts of the number and arrangement of lanes on freeway exit ramps. *Accident Analysis and Prevention*, Vol. 41, pp. 543-551.

Chen, H., Zhou, H., & Liu, P. (2014). Freeway deceleration lane lengths effects on traffic safety and operation. *Journal of Safety Science*, Vol. 64, 00. 39-49.

Chen, H., Zhou, H., & Liu, P. (2014). Freeway deceleration lane lengths effects on traffic safety and operation. *Journal of Safety Science*, Vol. 64, 00. 39-49.

Chen, H., Zhou, H., Zhao, J., & Hsu, P. (2011). Safety performance evaluation of left-side off-ramps at freeway diverge areas. *Accident Analysis and Prevention,*, Vol. 43, pp. 605-612.

Chen, Y., Persaud, B., & Sacchi, E. (2012). Improving Transferability of safety performance functions by Bayesian Model Averaging. *Transportation Research Record*, Volume 2280, 162-172.

Cirillo, J. A. (1970). *The relationship of accidents to length of Speed-Change Lanes and Weaving Areas on Interstate Highways, Report HRR 312.* Washington, DC: Highway Research Record.

Diggle, J. P., Heagerty, P., Liang, K., & Zeg. (2002). *Analysis of Longitudinal Data, 2nd Edition.* Oxford University Press.

Dinu, R. R., & Veeraragavan, A. (2011). Random parameter models for accident prediction on two-lane undivided highways in India. *Journal of Safety Research*, 39-42.

Dobson, A. (2002). *An Introduction to Generalized Linear Models, 2nd Edition.* Chapman & Hall/CRC.

El-Basyouny , K., & Sayed, T. (2009). Urban Arterial Accident Prediction Models with Spatial Effects. *TRR: Journal of the Transportation Research Board*, No. 2102, pp. 27-33.

El-Basyouny, K., & Sayed, T. (2009). Urban Arterial Accident Prediction Models with Spatial Effects. *TRR: Journal of the Transportation Research Board*, No. 2102, pp. 27-33.

ESRI. (2013). *ArcMap 10.2.* Redlands, CA: Environmental Systems Resource Institute (ESRI).

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Boca Laton: Taylor & Francis Group, LLC.

FAST. (2014, June 01). *Regional Transportation Commission of Southern Nevada (RTC).* Retrieved from Freeway and Arterial System of Transportation: http://bugatti.nvfast.org/

Flahaut, B., Mouchart, M., Martin, E. S., & Thoma. (2003). The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. *Accident Analysis and Prevention*, Vol. 35, pp. 991-1004.

Garber, J. N., & Hoel, A. L. (2009). *Traffic and Highway Engineering, 4th Edition.* Toronto: Cengage Learning.

Garnowski, M., & Manner, H. (2011). On factors related to car accidents on German Autobahn connectors. *Accident Analysis and Prevention*, 1864-1871.

Garson, G. D. (2013). *Hierarchical Linear Models: Guide and applications.* Thousand Oaks, CA: SAGE Publications, Inc.

Gelman, A., Carlin, J. B., Stern, H. S., & Dunson, D. (2014). *Bayesian Data Analysis, 3rd Edition.* FL: Taylor & Francis Group, LLC.

Golob, T. F., Recker, W. W., & Alvarez, V. M. (2004). Safety Aspects of Freeway Weaving Sections. *Transportation Research Part A*, Vol. 38, pp. 35-51.

Guo, F., Wang, X., & Abdel-Aty, M. (2010). Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis and Prevention*, Vol. 42, pp. 84-92.

Hansen, L. P. (1982). Large sample properties of Generalized Method of Moments Estimators. *Econometric society*, Vol. 50, No.4, pp. 1029-1054.

Hauer, E. (1997). *Observational Before-After Studies in Road Safety: Estimating the effect of highway and traffic engineering measures on road safety.* Elsevier Science Ltd.

Hauer, E. (2001). Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. *Accident Analysis and Prevention*, Vol. 33,pp. 799–808.

Hepple, L. (2009). *Spatial Autocorrelation*. Retrieved from In The dictionary of human geography: http://search.credoreference.com .ezproxy.library.unlv.edu/content/entry/bkhumgeo/spatial_autocorrelation/0

Hilbe, J. M. (2011). *Negative Binomial Regression, 2nd Edition.* United Kingdom: Cambridge University Press.

Huang, H., & Abdel-Aty, M. (2010). Multilevel data and Bayesian analysis in traffic safety. *Accident Analysis and Prevention*, 1556-1565.

Joe, B., Greg, G., & Davey, W. (1999). Safety evaluation of acceleration and deceleration lane lengths. *Institute of Transportation Engineers,* Vol. 69, No. 5, pp. 50.

Karlaftis, G. M., & Tarko, P. A. (1998). Heterogeneity consideration in accident modeling. *Accident Analysis and Prevention,* Vol. 30, pp. 425–433.

Kery, M. (2010). *Introduction to WINBUGS for Ecologists: A Bayesian approach to regression, ANOVA, Mixed models and related analyses.* Elsevier Inc.

Lee, D. (2014). CARBayes: Spatial Areal unit Modeling. *R package version 3.1.2*.

Lee, D. A. (2011). comparison of Conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology,* Vol. 2, pp. 79-89.

LeSage, J. P. (1999). *The Theory and Practice of Spatial Econometrics.*

Li, L., Zhu , L., & Sui, Z. D. (2007). A GIS-based Bayesian approach for analyzing spatial-temporal patterns of intra-city motor vehicle crashes. *Journal of Transport Geography*, Vol. 15, pp. 274-285.

Li, L., Zhu, L., & Sui, Z. D. (2007). A GIS-based Bayesian approach for analyzing spatial-temporal patterns of intra-city motor vehicle crashes. *Journal of Transport Geography*, Vol. 15, pp. 274-285.

Liu, P., Chen, H., Lu, J. J., & Cao, B. (2010). How Lane Arrangements on Freeway Mainlines and Ramps Affect Safety of Freeways with closely spaced Entrance and Exit Ramps. *Journal of Transportation Engineering*, Vol. 136, No. 7, pp. 614-622.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A*, Vol. 44, pp. 291-305.

Lord, D., & Park, Y. P. (2008). Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. *Accident Analysis and Prevention*, Vol. 40, pp. 1441–1457.

Lord, D., Washington, P. S., & Ivan, N. J. (2005). Poisson, Poisson-Gamma and Zero-inflated regression models of motor vehicle crashes: balancing statistical fir and theory. *Accident Analysis and Prevention*, Vol. 37, pp. 35–46.

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1-22.

Miaou, S., & Song, J. J. (2005). Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention*, Vol. 37, pp. 699-720.

Mitra, S., & Washington, S. (2007). On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis and Prevention*, Vol. 39, pp. 459–468.

Mulokozi, E., & Teng, H. (2015). Safety analysis of freeway segments with random parameters. *Transporatation Research Board, 2015.* Washington, DC: TRB.

Navidi, W. (2011). *Statistics for Engineers and Scientists. 3rd Edition.* New York: McGraw-Hill Companies, Inc.

Noland, R. B., & Quddus, M. A. (2004). A spatially disaggregate analysis of road causalities in England. *Accident Analysis and Prevention*, Vol. 36, pp. 973-984.

Nshankar, N. V., & Sittikariya, S. (2014, July 01). *Bayesian Formulations for Heterogeneity in Crash Count Models.* Retrieved from DKS – Transportation solutions: http://www.dksassociates.com/wp-content/files_mf/13348756001BayesianFormulations1.pdf

Ogden, K. W. (1996). *Safer Roads: A guide to Road Safety Engineering.* England: Avebury Technical.

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

Quddus, M. A. (2008). Modeling area-wide count outcomes with spatial autocorrelation and heterogeneity: An analysis of London crash data. *Accident Analysis and Prevention*, Vol. 40, pp.1486-1497.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models, 2nd Edition.* Thousands Oaks: Sage Publications, Inc.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edition.* Thousand Oaks, CA: Sage Publications, Inc.

Roess, P. R., Prassas, S. E., & McShane, R. W. (2011). *Traffic Engineering.* NJ: Pearson Higher Education, Inc.,.

Ross, M. S. (2010). *Introduction to Probability Models, 10th Edition.* Elsevier Inc.

Sarhan, M., Hassan, Y., & Adb El Halim, O. A. (2008). Safety of Freeway Sections and relation to speed-change lanes. *Canadian Journal of Civil Engineering*, Vol. 35, pp. 531-541.

Schnabel, K. U., Little, T. D., & Baumert, J. (2000). *Modeling longitudinal and Multilevel data: Practical Issues, applied approaches, and specific examples.* Mahwah, NJ: Lawrence Erlbaum.

Scott, M. A., Simonoff, J. S., & Marx, B. D. (2013). *The SAGE Handbook of Multilevel Modeling.* SAGE Puplications Inc.

Shankar, N. V., Albin, R. B., Milton, J. C., & Mannering, F. L. (1998). Evaluating Median Crossover Likelihoods with Clustered Accident Counts. *Transportation Research Record*, 98-0238.

Shankar, V., Mannering, F., & Barfield, W. (1995). Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention,*, Vol. 27, No. 3, pp. 371-389.

Sharma, S. (2006). *Investigation of regression to mean effect in Traffic safety evaluation methodologies.* ProQuest.

Simon, P. W., Matthew, K. G., & Fred L, M. (2011). *Statistical and Econometric Methods for Transportation Data Analysis, Second Edition.* Boca Raton, FL: Taylor & Francis Group.

Sittikariya, S. (2006). *Methodological approaches to Incorporate Heterogeneity in Traffic Accident Frequency Models.* Pennsylvania: Thesis – The Pennsylvania State University.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Modeling.* Chapman & Hall/CRC.

Snijders, T. A., & Bosker, R. J. (1999). *Multilevel Analysis: An intoduction to basic and advanced multilevel modeling.* SAGE Pupbications Inc.

Stata. (2011). *Data Analysis and Statistical Software.* Texas, College Station: StataCorp LP.

TRB. (2010). *Highway Capacity Manual, Volume 2: Uninterrupted flow.* Washington, DC: National Academy of Sciences.

University of Bristol. (2015, May 31). Retrieved from Center for Multilevel modeling: http://www.bristol.ac.uk/cmm/

Veeraragavan, A., & Dinu, R. R. (2011). Random parameter models for accident on two-lane undivided highways in India. *Journal of Safety Research*, Vol. 42, pp. 39-42.

Venkataraman, N. S., Ulfarsson, G. F., Shankar, V., Oh, J., & Park, M. (2011). Model of Relationship between Interstate Crash Occurrence and Geometrics. *Transportation Research Record*, 41-48.

Wang, Y., & Kockelman, K. M. (2013). A Poisson – lognormal Condition-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis and Prevention*, Vol. 60, pp. 71-84.

Washington, P. S., Karlaftis, G. M., & Mannering, L. F. (2011). *Statistical and Econometric Methods for Transportation Data Analysis, 2nd Edition.* Boca Raton: Taylor & Francis Group.

West, T. B., Welch, B. K., & Gatecki, T. A. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software.* Boca Laton: Taylor & Francis Group, LLC.

Wong, J., & Chung, Y. (2008). Analyzing heterogeneous acident data from the perspective of accident occurrence. *Accident Analysis and Prevention*, 357-367.

Yu, R., Abdel-Aty, M., & Ahmed, M. (2013). Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis and Prevention*, 371-376.

CURRICULUM VITAE

Eneliko Mulokozi
emulokozi@aol.com

EDUCATION

University of Nevada, Las Vegas, Nevada, 2011- 2013
M.S. in Engineering, Civil Engineering, 2013

University of Dar es salaam, Dar es Salaam, Tanzania, 1999 – 2003
B.S in Engineering**;** Civil & Structural Engineering, 2003

AWARDS AND HONORS

- U.S DOT - Dwight David Eisenhower Transportation Fellowship Program
- Member, Golden Key International Honor Society
- Member, The National Society of Leadership and Success (Sigma Alpha Pi)
- Outstanding Academic Achievement - Alliance of Professionals of African Heritage

RESEARCH AND INDUSTRY EXPERIENCES

Feasibility study of Public Bike Sharing Program in Las Vegas: July 2014 – May 2015

- Assessed the feasibility of establishing a public bike share transport mode on and around the main campus of the University of Nevada, Las Vegas.
- Recommended measures to minimize bike / pedestrian conflicts on campus.

Supervisor: Prof. Hualiang (Harry) Teng

Next Generation Performance Monitoring Data Needs for Nevada DOT: Jan. 2014 – June 2014

- Recommended an algorithm of a safety performance function for predicting the likelihood of a crash occurrence using real-time traffic characteristics while accounting for the effect of geometric elements as confounding factors.
- Tested the performance of an algorithm using real-time traffic data and compared different safety performance measures to the threshold safety measures for decision making on improving safety levels.

Supervisor: Prof. Hualiang (Harry) Teng, Prof. Alexander Paz, and Dr. Morris

Evaluation of Geometric Design Needs of Freeway Systems Based on Archived ITS and Safety Data: May 2012 – June 2013

- Evaluated safety performance levels of freeway geometric elements in the city of Las Vegas, Nevada.
- Recommended network screening technique to identify locations with potential for improvement

Supervisor: Prof. Hualiang (Harry) Teng

Manager – Dar es Salaam water and sewerage system (Dawasco) – Tanzania: March 2007 – Feb 2010.

- Designed and implemented min-projects of water distribution systems
- Conducted maintenance activities of water distribution systems
- Supervised revenue generation

Supervisor: Mr. Alex Kaaya (Managing Director)

Head of business and civil/building works department, ELCT - Tanzania: June 2004 – Dec. 2005

- Analyzed and designed water distribution systems and buildings
- Rehabilitated Diocesan buildings
- Prepared contract documents and civil works maintenance schedule
- Conducted project cost estimates and site supervision

Supervisor: Mr. George Chobya (Diocesan General Secretary)

PUBLICATIONS

3. **Mulokozi, E.,** and Teng, H (2015). Safety Analysis of Freeway Segments with Random Parameters. *Transportation Research Record:* Journal of the Transportation Research Board. **In press**
2. Kwigizile, V., **Mulokozi, E.,** Xu, X., Teng, H., and Ma, C (2014). Investigation of the Impact of Corner Clearance on Urban Intersection Crash Occurrence. *The Journal of Transportation Statistics* V10, N1 2014. pp. 35 - 48
1. Xu, X., Teng, H., Kwigizile, V., and **Mulokozi, E** (2014). “Modeling Signalized Intersection Safety with Corner Clearance”. *ASCE Journal of Transportation Engineering* V140, N6

RESEARCH INTERESTS

- Road and Traffic safety
- Traffic operations and simulations
- Intelligent Transportation Systems
- Transportation systems design and evaluation
- Data mining

TEACHING EXPERIENCE

Course - Introduction to Railway Transportation: Jan 2013 – May 2013

- Taught three lectures

TEACHING INTERESTS

- Transportation safety
- Geometric design of highways
- Transportation planning
- Traffic engineering
- Intelligent transportation systems
- Sustainable transportation

SERVICES AND PROFESSIONAL AFFILIATION

- President – American Railway Engineering and Maintenance – of - the Way Association Student Chapter at University of Nevada, Las Vegas, 2011 - 2012
- Member, Institute of Transportation Engineers (ITE), 2011 - Present