Screening Urban Road Network for Corridors with Promise

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Civil Engineering

by

Yue Zhao

Dr. Zong Tian/Dissertation Advisor

August, 2015
We recommend that the dissertation prepared under our supervision by

YUE ZHAO

entitled

Screening Urban Road Network for Corridors with Promise

be accepted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Zong Tian, Ph.D., Advisor

Hao Xu, Ph.D., Committee Member

Sami Fadali, Ph.D., Committee Member

Elie Hajj, Ph.D., Committee Member

Shunfeng Song, Ph.D., Graduate School Representative

David Zeh, Ph.D., Dean, Graduate School

August, 2015
Abstract

Both federal and state policy makers increasingly emphasize the need to reduce traffic fatalities and serious injuries. Finding improved methods to enhance roadway safety has become a top priority. In an attempt to reduce traffic crashes, crash-prone locations should be identified for increased law enforcement activities, education programs, and engineering improvements. This dissertation addresses the critical issue in traffic safety research, methods for corridor-based screening for safety improvement. The role of corridor level screening is to periodically examine the entire urban roadway network in order to generate a list of corridors ranked in order of priority by which detailed engineering studies should be conducted. Ongoing debates in regards to corridor level network screening include what should constitute a corridor for the purpose of network screening, and how a local agency should perform a corridor screening. This research provides answers to these questions.

Firstly, a comprehensive literature review was conducted to summarize current practices across the nation pertaining to corridor level network screening. No consensus was found in terms of corridor definitions or screening methodologies. Observed traffic crashes are generally used in evaluating the safety of urban facilities in state departments of transportation although model based evaluation is highly recommended because of its exceptional merits. Secondly, this research proposes a spatial clustering based approach to define urban corridor boundaries. The idea is to group signalized intersections along urban arterial roads based on their spatial auto-correlations. Local
spatial measurements (local Moran’s I and Getis-Ord index) are adopted to cluster multivariate intersection data. The analyses of arterials in the Reno-Sparks area indicate that the proposed approach provides reasonable corridor definitions.

The next section of the dissertation proposes a model-based scheme for screening urban corridors. Significant efforts were made to collect crash, traffic and road inventory data at intersection level, segment level and corridor level. Data assembling and processing were conducted in ArcGIS. Statistical models, such as negative binomial regression models, have been widely used in developing crash prediction models over the past decades. This research investigated other models including the Poisson-Inverse Gaussian and Poisson-Lognormal models. Analysis results imply that for a certain data set different model assumptions will generate quite different results. Overall, the Poisson-Inverse Gaussian model and the Poisson-Lognormal model perform better than the negative binomial model in terms of goodness-of-fit statistics. Due to the high flexibility of Inverse Gaussian and Lognormal distributions, such models can be adopted as alternatives to the negative binomial models in developing crash prediction models.

Furthermore, this research explores the effect of spatial correlations in crash prediction modeling. Significant spatial correlations were found not only within different intersection data sets and segment data set, but also in the model residuals and fitted values. The spatial eigenvectors are introduced into the developed models to supplement the spatial effects. Different neighboring proximity structures are tested for assembled data sets to establish the configuration that results in the optimal performance of the prediction models. The comparisons of model goodness-of-fit statistics indicate that the spatial correlations contribute significantly to model heterogeneity. Ignoring spatial impacts may result in biased estimates of model parameters and incorrect inferences.
Combining safety performance of intersections and segments, a Corridor Safety Measurement (CSM) is proposed as the performance measure for corridor screening. The measurement is used to identify corridors in the Reno-Sparks area that have promise as locations where improvements will result in substantial crash reduction. The findings from this research will assist engineers to proactively identify and analyze high crash locations from a corridor perspective and detect potential problematic locations not identified through the traditional hot spot analysis.
Acknowledgment

Over the last five years, I have received support and encouragement from a great number of individuals. Dr. Zong Tian has been a mentor, colleague and friend. His unwavering support and encouragement have made my study and research at the University of Nevada, Reno a thoughtful and rewarding journey. I would like to thank my dissertation committee of Dr. Hao Xu, Dr. Shunfeng Song, Dr. Sami Fadali, Dr. Elie Hajj and Dr. Reed Gibby for their constructive comments and suggestions. In addition, Mr. Chuck Reider from CWR Solutions and Mr. Mike Colety from Kimley-Horn and Associates both provided valuable advice as this research moved from an idea to a completed research study.

During the dissertation writing, Anabel Hernandez spent hours proofreading and commenting on my research. Arafat Hossain Khan provided needed insights and support. I thank other members in the Center for Advanced Transportation Education and Research for being such good companions through my school years and for helping me survive all the stress. I would also like to thank Dr. Walt Johnson, who was associate dean of the College of Engineering, and his lovely wife Mrs. Marilyn Johnson for being there to talk and to offer guidance during difficult times of my life.

Finally, thanks to my parents for their unconditional love. They have taught me how to be persistent and independent and have given me confidence and motivated me in so many ways. I am so grateful for them both and thank them with all my heart and soul.
## Table of Contents

1 Introduction

1.1 Motivation ............................................. 1
1.2 Objectives and Scope ................................. 2
1.3 Organization of Dissertation ......................... 3

2 Literature Review

2.1 Corridor Screening Criteria .......................... 5
2.2 Current Practices ...................................... 10
   2.2.1 Nevada ............................................. 10
   2.2.2 New Hampshire .................................... 14
   2.2.3 Ohio ............................................... 15
   2.2.4 Other States ....................................... 15
2.3 Available Tools for Corridor Screening ............... 16
   2.3.1 SafetyAnalyst ..................................... 16
   2.3.2 ESRI ArcGIS ....................................... 17
2.4 Chapter Summary ...................................... 19

3 Define Urban Corridors ................................. 20

3.1 Components of Urban Corridors ........................ 20
3.2 Spatial Autocorrelation of Signalized Intersections .... 23
   3.2.1 Global Measures of Spatial Autocorrelation ...... 24
4 Techniques on Crash Prediction Models

4.1 Crash Prediction Models
   4.1.1 Generalized Poisson Regression Model
   4.1.2 Generalized Poisson Mixture Regression Models

4.2 Spatially Adjusted Crash Prediction Models
   4.2.1 Eigenvector Spatial Filtering (ESF)
   4.2.2 Spatially Adjusted Crash Prediction Models using ESF
   4.2.3 Spatially Adjusted Crash Prediction Models using CAR
   4.2.4 Section Summary

4.3 Model Development and Over Fitting Issues

4.4 Bayesian Paradigm

4.5 Goodness-of-Fit Measurements

4.6 Chapter Summary

5 Crash Prediction Models for Intersections along Corridors

5.1 Intersection Data Preparation
   5.1.1 Site Selection
   5.1.2 Data Collection and Processing

5.2 Model Development
   5.2.1 Four-Way Signalized Intersections (4SG)
   5.2.2 Four-Way Stop Controlled Intersections (4ST)
   5.2.3 Signalized T-Intersections (3SG)
   5.2.4 Stop Controlled T-Intersections (3ST)
5.3 Chapter Summary ............................................. 105

6 Crash Prediction Models for Segments along Corridors 107
  6.1 Data Set for Analysis ........................................ 107
  6.2 Model Development ........................................ 109
    6.2.1 Model Assumptions and Forms ......................... 109
    6.2.2 Spatially Adjusted Models ............................. 109
  6.3 Chapter Summary ........................................... 113

7 Corridor Safety Evaluation 114
  7.1 Corridor Safety Measurement (CSM) ....................... 114
  7.2 Case Demonstration ....................................... 115
  7.3 Chapter Summary .......................................... 119

8 Summary and Conclusions 120
  8.1 Findings .................................................. 120
  8.2 Future Extensions ........................................ 122

Bibliography 123

Appendix A Poisson Inverse-Gaussian Distribution 136

Appendix B R Codes 140
List of Tables

3.1 Global Measures of Spatial Auto-correlation for Selected Multivariates on Kietzke Ln in Reno, Nevada .................................................. 26
3.2 Standardized Local G-O Index $z(G_i)$ for Selected Multivariates on Kietzke Ln in Reno, Nevada .................................................. 30
3.3 Summary Statistics of Designated Corridors in the Study Region ... 37
4.1 General Coding Schemes for Weight Matrix Standardization ........ 61
5.1 Breakdown of Intersection Identification ................................. 75
5.2 Intersection Data Requirements in the HSM ............................. 76
5.3 Selected Intersections in the Study Region ............................... 76
5.4 Summary Statistics of Intersection Data ................................. 78
5.5 Generalized Poisson Models for 4SG ........................................ 82
5.6 Over-Dispersion Test for Generalized Poisson Models ............... 83
5.7 Generalized Poisson-Gamma Models for 4SG ............................ 84
5.8 Generalized Poisson-Lognormal Models for 4SG ...................... 84
5.9 Generalized Poisson-Inverse Gaussian Models for 4SG .............. 84
5.10 GOF Measurements of Unadjusted Models for 4SG ................. 85
5.11 Parameter Estimates of Spatial Adjusted Model SA-I for 4SG ...... 88
5.12 Parameter Estimates of Spatial Adjusted Model SA-II for 4SG ...... 91
5.13 Parameter Estimates of Spatial Adjusted Models SA-III and SA-IV for 4SG .................................................. 96
5.14 Parameter Estimates of Spatial Adjusted Model SA-V for 4SG ........ 98
5.15 Comparison of Model GOF Statistics for 4SG ................................. 99
5.16 GOF Measurements of Unadjusted Models for 4ST ......................... 100
5.17 GOF Measurements of Spatially Adjusted Models for 4ST ................. 101
5.18 Parameter Estimates of Spatial Adjusted Model for 4ST ..................... 101
5.19 GOF Measurements of Unadjusted Models for 3SG .......................... 102
5.20 GOF Measurements of Spatially Adjusted Models for 3SG .................. 102
5.21 Parameter Estimates of Spatial Adjusted Model for 3SG ..................... 102
5.22 GOF Measurements of Unadjusted Models for 3ST .......................... 104
5.23 GOF Measurements of Spatially Adjusted Models for 3ST .................. 104
5.24 Parameter Estimates of Spatial Adjusted Model for 3ST ..................... 104

6.1 Summary Statistics of Roadway Segment Data .................................. 108
6.2 GOF Measurements of Unadjusted Models for Roadway Segments .. 109
6.3 Parameter Estimates of Spatially Adjusted Model I for Roadway Segments ................................................................. 111
6.4 Parameter Estimates of Spatially Adjusted Model II for Segments within Corridors ............................................................... 112

7.1 Top Ten High Crash Corridors in Reno-Sparks based on Observed CSMs118
7.2 Top Ten High Crash Corridors in Reno-Sparks based on Expected CSMs118
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Demonstration of Sliding Mile Approach in Nevada</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Identified High Crash Corridors in Las Vegas using Geomedia</td>
<td>13</td>
</tr>
<tr>
<td>3.1</td>
<td>Determining the Number of Corridors using Elbow Method</td>
<td>33</td>
</tr>
<tr>
<td>3.2</td>
<td>Determining the Number of Corridors based on Gap Statistic</td>
<td>34</td>
</tr>
<tr>
<td>4.1</td>
<td>Poisson Distribution</td>
<td>41</td>
</tr>
<tr>
<td>4.2</td>
<td>Poisson-Gamma Distribution</td>
<td>45</td>
</tr>
<tr>
<td>4.3</td>
<td>Poisson-Lognormal Distribution</td>
<td>47</td>
</tr>
<tr>
<td>4.4</td>
<td>Poisson-Inverse Gaussian Distribution</td>
<td>50</td>
</tr>
<tr>
<td>5.1</td>
<td>Percentages of Traffic Crashes by Intersection Control Type</td>
<td>74</td>
</tr>
<tr>
<td>5.2</td>
<td>Conflicts between Major and Minor Approaches at a Typical Intersection</td>
<td>81</td>
</tr>
<tr>
<td>5.3</td>
<td>Plots of Fitted Values and Residuals of Model IV for 4SG</td>
<td>83</td>
</tr>
<tr>
<td>5.4</td>
<td>Coordinates of 4SG Intersections in the Study Region</td>
<td>86</td>
</tr>
<tr>
<td>5.5</td>
<td>Comparison of Spatially Adjusted Model SA-I and Unadjusted Model</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UA-VIII for 4SG</td>
<td>89</td>
</tr>
<tr>
<td>5.6</td>
<td>Comparison of Spatially Adjusted Model SA-II and Unadjusted Model</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UA-VIII for 4SG</td>
<td>92</td>
</tr>
<tr>
<td>5.7</td>
<td>4SG Intersections along Corridors</td>
<td>93</td>
</tr>
<tr>
<td>5.8</td>
<td>Comparison of Spatially Adjusted Models SA-III and SA-IV for 4SG</td>
<td>95</td>
</tr>
</tbody>
</table>
5.9 Comparison of Spatially Adjusted Model SA-V and Unadjusted Model UA-VIII for 4SG .............................. 97
5.10 Plots of Model Residuals for 3SG ............................................. 103
7.1 Corridor Screening Results based on CSMs ................................. 117
Chapter 1 Introduction

1.1 Motivation

The goal of transportation systems is to promote the safe and efficient movement of people and goods. The Federal Highway Administration (FHWA) held a national workshop in June 1990 to create a list of the five most promising traffic crash countermeasures. The Safety Corridor concept was proposed since then as one of five focus areas. In the United States (U.S.), each state was required by the FHWA as part of the Highway Safety Improvement Program (HSIP) to submit an annual report describing no less than 5% of their highway locations (intersections, segments or corridors) exhibiting the most severe safety needs, as well as develop a plan to remedy those hazardous locations, all in an effort to provide a safe and efficient transportation network.

State Departments of Transportation (DOTs) continue to improve the safety of roadway segments by placing safety as their forefront priority. There has been an increased interest in applying criteria to identify sites on a roadway network that present high potential in crash risks for future engineering investigation and safety improvement. Engineers and researchers have been developing approaches to evaluate safety on roadways and mitigate unsafe locations. Extensive research has been undertaken on identifying hot spot locations either at intersections or segments [18, 46, 48, 67, 82, 85, 89, 96]. Another solution, Safety Corridor Analysis, was proposed by several states to provide an opportunity to further improve safety especially
when highway funds are programmed. Screening sites from a corridor perspective is important and even more complicated because of mixed traffic conditions involving transit, bicycle and pedestrian activities along roadway segments as well as at intersections. It is also valuable to develop corridor-based safety improvement strategies to proactively identify sites that have the highest potential for safety improvement.

At present, there is no standard procedure to screen urban corridors and to prioritize hazardous locations along corridors for safety enhancement. In some states, roadway segments with observed crash and/or fatality rates above the statewide average may have some of their lengths designated as high crash corridors. Traffic fines are usually doubled in these areas. Several states have initiated *Safety Corridor Programs* dedicated to improving safety along the state’s most crash-prone corridors in rural areas. Nevertheless, in accord with a recent nationwide survey, relatively few states have developed measures to screen urban corridors. A safety index calculated using observational crashes is normally applied in preliminary corridor screening. The selected corridors require further examination to identify the factors that increase crash risk. A more proactive corridor-level safety management approach is required to identify high risk locations and to effectively implement a safe corridor program and select pilot corridors for consideration in future projects.

### 1.2 Objectives and Scope

Several state highway agencies have automated procedures for network screening to identify potential locations and rural corridors for improvements. Typically, the procedures include threshold values of observed crash frequency or crash rate, at times combined with crash severity index. Most state agencies selected corridors based solely on the observed number of crashes. It has been recognized that a more comprehensive procedure of identifying high crash corridors should be developed using
criteria in addition to crash frequency. Thus, the overall objective of this research is
to develop a scheme to screen high crash-prone corridors in urban areas. The primary
objectives are: 1) to propose approaches to define urban corridor boundaries; and 2)
to establish corridor performance measures to screen urban road network for corridors
with promise.

Traffic crashes may involve the interactions of vehicle, driver, pedestrian, roadway
and environmental factors. This research focuses on the impact of roadway factors
(such as roadway geometry and characteristics, traffic exposures and signals) on traffic
crashes. The main roadway types considered include principal arterial roads (exclud-
ing interstate highways, other freeways and expressways) and minor arterial roads.
Interchanges and ramps are beyond the scope of this research.

1.3 Organization of Dissertation

The dissertation documents all the findings and conclusions pertaining to urban corri-
dor definition, high crash corridor screening approaches, crash prediction models for
intersections and segments along corridors, and the development of corridor safety
measurements. Chapter 1 introduces the background, objectives and scope of this
research. Chapter 2 is a comprehensive literature review which documents current
practices across the U.S. and lessons learnt. Chapter 3 presents several options to
define an urban corridor and the ongoing debates regarding corridor level screen-
ing. Next, a new way of defining urban corridors is proposed and explained in detail.
Chapter 4 presents the theoretical details about the modeling methodologies adopted.
Chapter 5 develops the crash prediction models for intersections along urban corridors
and explores the spatial correlations among the sites of different facility types. Sim-
ilarly, Chapter 6 focuses on roadway segments along selected corridors and presents
the corresponding crash prediction models. Spatial effects among segments are also
briefly discussed. In Chapter 7, a corridor safety measurement is presented for promoting urban corridors with high potential crash risks. A case study of urban arterial roads in Reno-Sparks, Nevada is illustrated to demonstrate the application of developed models and approaches. Finally, Chapter 8 summarizes the major findings and contributions of this research. Future research extensions are suggested.
Chapter 2  Literature Review

This chapter provides a comprehensive review of previous research pertaining to urban corridor screening. At present, relatively few states across the U.S. have developed high crash corridor criteria and therefore there are challenges finding the appropriate information. The review focuses on limited publications from various sources including state DOTs’ documents and annual reports, public agency reports, journal papers, and conference proceedings. The author also contacted other DOT personnel through list server and personal contact to gather materials that are not available from the above sources.

2.1 Corridor Screening Criteria

There are many different opinions about what constitutes a corridor among the states. Some defined corridors may only be a thousand feet in length and contain a few intersections while others may consist of a segment of roadway over 20 miles (normally in rural areas) in length. Presently, the corridor length used for network screening differs from one jurisdiction to another. A corridor is typically defined outside of the database for the purpose of a specific subject or program. This is probably why a corridor definition from a physical perspective varies widely.

Network screening is the first step of the Roadway Safety Management Process outlined in the American Association of State Highway and Transportation Officials’ (AASHTO) Highway Safety Manual (HSM) [66]. It is a process of reviewing a focused
network to rank and identify sites from most likely to least likely to reduce crash frequencies with implementation of countermeasures. The result of network screening is a list of sites that have the highest potential for safety improvement. Basic steps involved in network screening include: 1) establish focus populations; 2) select or develop performance measures; 3) select or develop screening method (e.g., sliding mile method, peak searching method or simple ranking approach); and 4) screen the focused network and report on the results. The simple ranking approach based on historical crash statistics is recommended for facilities composed of connected segments and intersections in the HSM. The FHWA SafetyAnalyst manual concluded that because corridor analyses are based upon extended arterial roads and hence multiple sites, there is less variability or randomness in the crash data. Thus, the biases that occur in analyzing observed crashes at individual sites are less. Subsequently, simple procedures based on observed crash frequencies and rates can be used and provide more reliable results than similar procedures when used to analyze individual sites [43].

Roadway sites of interest can be ranked based upon one or combinations of the following criteria:

- Observed number of crashes or crash rate over a certain period of time;
- Estimated mean crash frequency using the Bayes method;
- The amount by which the estimated mean crash frequency exceeds a standard value;
- The probability that the mean frequency exceeds a standard value.

Safety performance measures usually involve crash frequency, crash rate, crash density and severity. Combinations of those indexes are also available for screening segments or intersections. The key considerations when selecting criteria are the data
availability and the subsequent screening approach. Corridors can be ranked by one or both of two basic measures:

\[
\text{Average crash/mi/yr} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{Y} K_{i,j}/Y_{i}}{\sum_{i=1}^{N} SL_{i}} \tag{2.1}
\]

\[
\text{Average crash/mvmt/yr} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{Y} K_{i,j}}{\sum_{i=1}^{N} \sum_{j=1}^{Y} (SL_{i,j} \times ADT_{i,j} \times 365)/10^6} \tag{2.2}
\]

where,

- \(N\): total number of sites in the corridors;
- \(Y_{i}\): total number of years of available data at site \(i\);
- \(SL_{i}\): roadway segment length under consideration at site \(i\). The value equals zero for all intersections;
- \(K_{i,j}\): observed total number of crashes at site \(i\) during year \(j\);
- \(ADT_{i,j}\): average daily traffic at site \(i\) during year \(j\);
- \(mvmt\): Million vehicle miles traveled.

The calculations in formula (2.1) and (2.2) are based on observed crashes. Formula (2.1) is essentially a crash frequency and does not consider traffic exposure when analyzing the safety potential of a corridor. Comparing crash frequencies between two corridors may not be meaningful unless the two corridors have the same level of exposure. Formula (2.2) is essentially a crash rate and takes traffic exposure into account. The fundamental problem with crash rate based criterion lies in the fact that the linear assumption between the number of crashes and vehicles miles traveled (VMT) must hold. However, many studies have argued that this rigid assumption is false especially when the volume range among the focused sites is wide. Also, using the
raw crash frequency or crash rate as an estimate of the true safety measurement is very crude and inefficient. The performance measures based on historical crash records tend to have simple data requirements yet do not require restrictive assumptions. They can be used in a preliminary screening to identify locations in need of more detailed analysis.

The site ranking or screening approaches include: 1) raw/naïve ranking; 2) scan based ranking; and 3) model based ranking. Naïve ranking refers to using historical crashes as the sole criterion to measure the safety of sites. As discussed in previous paragraphs, this criterion is rather coarse and is likely to prioritize sites with low traffic volumes and low recorded number of crashes. The scan based approach is also known as the sliding window- or peak searching-based approach as expressed in the HSM. This approach also suffers from the “low volume and low crashes” problem. In practice, it is also difficult to select the best window size and moving step size and the analysis results can be very sensitive to these selections. In short, the naïve and scan based ranking methods purely based on observational crashes may suffer potential issues including but not limited to:

- Very sensitive to random variations because of the regression-to-the-mean nature of crash data;

- Tend to prioritize low-volume and low-crash locations;

- Generally does not consider data from other locations; and

- Assume linear relationship between crash and traffic exposure.

More in-depth research recommended the model based screening approach in which crash prediction models are developed to identify locations that have a significant higher number of crashes than would be expected based on site conditions
[41, 51, 59, 88, 106]. Most of the preceding research was tailored for intersections or segments screening while a few studies approached corridor perspective.

El-Basyouny and Sayed developed crash prediction models for urban corridors with consideration of random parameters [28]. They found that even though covariates representing AADT, segment length, crosswalks density, etc., were related to crash frequencies, their effects vary significantly across corridors. The proposed Poisson-Lognormal model with random parameters for each corridor provided the best fit. The conclusions from this study also proved the benefit of clustering road segments into rather homogeneous groups (i.e., corridors) and incorporating random corridor parameters in the crash prediction models. In addition, Abdel-Aty and Wang pointed out that signalized intersections especially closer ones along a certain corridor, are spatially correlated and will influence each other in many respects [1]. They developed crash prediction models using Generalized Estimating Equations (GEE) for analyzing the safety of signalized intersections along corridors. Later, Xie and Wang applied the Bayesian Hierarchical models to analyze the intersection safety with consideration of potential correlations [107]. Qin et al. developed a truck corridor crash severity index to evaluate selected corridors [88]. They introduced statistical models of crash frequencies and injury severity into the truck crash severity index. Negative Binomial models were used to estimate truck crash frequencies while Multinomial Logit (MNL) models were generated to estimate the proportion of specific crash severity. The truck safety index was calculated as a function of the number of crashes expected to occur on selected corridors multiplied by the proportion of severity as well as the associated crash unit costs.
2.2 Current Practices

A few states across the U.S. have developed safety corridor programs that provide procedures to identify rural and urban corridors with safety issues, including but not limited to high crash frequencies and rates. This section presents the state of the art of high crash corridor identification and a summary of the criteria applied.

2.2.1 Nevada

In Nevada, the selection of high crash corridors in the past has been based solely on crash frequency. Nevada DOT (NDOT) has recognized that a more detailed method of selecting the corridors should be developed using additional criteria. Recently, NDOT researchers have developed a corridor analysis procedure based on the Sliding Mile approach in Geomedia. The idea of the method is demonstrated in Figure 2.1.

![Figure 2.1: Demonstration of Sliding Mile Approach in Nevada](image)

In the sliding window approach, a window with a user-specified length would move incrementally along contiguous roadway segments such that the current window location would overlap the previous one [66]. In Nevada’s practice, the corridor length is designated as 0.5 mile whereas the sliding window is selected as 1 mile. The corridor length is determined using engineering judgment. Utilizing Geomedia software, three
major databases including the NDOT TRINA database, UrbanRural RDWYSYS, and CrashData database are first connected and prepared for future query. The second step is to generate the corridors for analysis. This was achieved by using the Linear Referencing System (LRS) Event Generation and Dynamic Segmentation functions in Geomedia. The LRS function allows to define corridor length and to locate features by specifying a distance (mile point) along a route. The dynamic segmentation function then takes two or more different route event tables or layers (e.g., traffic data from TRINA database and crash data from CrashData database) and merges them into one file with mile point locations adjusted accordingly. The attributes of each layer will both be presented in a new file.

For each 0.5 mile corridor, a total ranking score is computed using weighted crash rate and weighted severity index. The crash rate and weighted severity are normalized first by dividing by the maximum rate and maximum weighted severity in the county or study area, and then the measures are added as the final ranking score. The total score is given by (2.3).

\[
Total \ Score = \frac{Weighted \ Severity}{(Weighted \ Severity)_{\text{max}}} \times 0.6 + \frac{Crash \ Rate}{(Crash \ Rate)_{\text{max}}} \times 0.4 \quad (2.3)
\]

where, the numerical factors assigned to severity index are: fatal crash=542; A injury crash=29; B injury crash=11; C injury crash=6.

The final ranking scores are divided into four categories with 0.4-0.5 threshold indicates corridors have the highest priority for consideration of improvements and 0.1-0.2 indicates the lowest priority. In NDOT’s current application, intersection-related crashes are not specified or distinguished from segment-related crashes. Therefore, the high crash intersections actually dominate the identified corridors. Figure 2.2 demonstrates the identified high crash corridors in the Las Vegas area. NDOT researchers find that current practice using the Sliding Mile approach in Geomedia
provides quick and gross rankings of corridors. The main shortcomings reported by NDOT safety engineers include: 1) intersection related crashes are not specified in this procedure; 2) the screening process should include PDO crashes. Locations with significant amount of PDO crashes may have potential issues in regards to geometry design, traffic signs or marks placement, etc. Serious injury crashes may occur at these locations if the PDO crash records are ignored; and 3) short sections with red (i.e., corridors with high potential for improvements) may be surrounded by several purple and green sections. Significant amount of engineering judgment needs to be applied.
Figure 2.2: Identified High Crash Corridors in Las Vegas using Geomedia
2.2.2 New Hampshire

The State of New Hampshire uses the Peak Searching approach in SafetyAnalyst to screen for excess fatal and all injury crash frequency for the crash types relevant to the critical emphasis area along corridors, such as lane departure, impaired driving, seat belts, etc. For instance, when screening for lane departure crashes, an agency-defined lane departure indicator for roadway segments in SafetyAnalyst was developed using head-on, sideswipe opposite, fixed object, off roadway in shoulder/median, and off roadway beyond the shoulder. This method is applicable for any combination of peak searching and sliding window, expected and excess frequency, and any severity level.

The peak searching approach divides roadway segments into windows of similar lengths (e.g., 0.1 mile, 0.2 mile, etc.), potentially growing incrementally in length until the length of the window equals the length of the entire roadway segment [43, 66]. Unlike the sliding window approach, the windows do not overlap in peak searching approach. For each window, the average crash frequency is calculated first and the results are subjected to precision testing. Based upon the statistical precision of the performance measure, the window with the maximum value of the performance measure within a roadway segment is used to rank the potential for reduction in crashes of that site relative to the other sites being screened.

Initially, a default coefficient of variation (CV) threshold of 0.50 was set up in SafetyAnalyst and no frequency limiting values were used. The engineers then take the frequency results and create a frequency distribution (frequency on Y-axis, percentile of sites ranked on X-axis) to determine where a crash frequency threshold should be set based on a steep break in the curve that separates the highest crash locations from the majority of sites. The threshold was applied as the CV limiting value to develop the final list of sites. This process was completed separately for urban and rural sites within each region in New Hampshire. Using the output reports
in CSV format, engineers can easily join the screening results to the roads feature class in ArcGIS software and identify high crash corridors based on the excess crash frequencies.

2.2.3 Ohio

The Ohio Department of Transportation (ODOT) considers a location as a candidate high crash location when the total number of crashes equal to or greater than 10 in a 3-year period, and the crash rate equal to or greater than 2 crashes per million vehicles [72]. For urban corridors, crashes occurring within a roadway section are measured by crashes per mile instead of raw numbers of crashes. This is done to take into account the length of roadway section. Crashes per mile are calculated by dividing the total crashes by the section length. The crash rate for an arterial section also takes into account the section length. The crash rate is the crash total divided by the section traffic volume multiplied by the section length. It is expressed in crashes per million vehicle miles traveled (MVMT). A minimum criteria of 10 crashes per mile and a crash rate of 2 crashes per million vehicle miles traveled is used in identifying high crash corridors.

2.2.4 Other States

Other investigated states including California, Virginia, Florida, Wyoming, Arizona, Oregon, Iowa, Missouri, New Jersey and New Mexico, conducted hot spot studies in both urban and rural areas. The studies were not necessarily focused on corridor level screenings. Their applications utilized traditional network screening approaches in the HSM and are based on historical observations.
2.3 Available Tools for Corridor Screening

Several analytical tool and techniques are available to analyze traffic crash data. However, questions like “where are most of the crashes occurring and why” were hard to answer. Nowadays, these questions can be answered using software tools. This section has been prepared as a brief summary of available tools for analyzing crash data and identifying high crash corridors. The main focus is given to AASHTOware Safety Analyst and ESRI’s ArcGIS.

2.3.1 Safety Analyst

Safety Analyst is a software package developed by FHWA and participating state and local agencies [43]. The package supports implementation of many of the analytic methods presented in Part B of the AASHTO Highway Safety Manual (HSM). The corridor screening procedure in Safety Analyst enables multiple sites to be aggregated together and be analyzed as a single entity such as corridor. Safety Analyst specifies whether corridors will be prioritized based upon two criteria as presented in formula (2.1) and (2.2).

There are a number of limitations with Safety Analyst’s corridor screening approach. The software typically requires comprehensive data sets that can provide sufficient information in a specific format to capture intricate interactions in the traffic system. In Safety Analyst, sites are aggregated to investigate the crash history of roadway segments and intersections. Each corridor is considered a single entity, the actual computations are performed on a site-by-site basis, i.e., one intersection or one segment at a time. This indicates that corridor boundaries need to be predefined. Additionally, the corridor screening approach in Safety Analyst is essentially a crash rate based method. Therefore, it presents the same problems as using any crash rate
network screening method.

2.3.2 ESRI ArcGIS

ESRI’s ArcGIS is a GIS software for working with maps and geographic information. The ArcGIS Desktop software package has been used as the primary platform to develop the framework for crash data analysis in several studies and research projects [14, 94].

GIS has improved the ability of analysts to perform advanced statistical analysis of crash data. The crash data analysis in GIS involves processing data to enable interpretation and performing statistical analysis, similar to how they would be done outside of a GIS environment. Nevertheless, by adding spatial context, the analyses give results that are relatively easy to connect to actual locations. GIS aids statistical analysis in several aspects. For instance, GIS allows more accurate data selection, screening and spatial analysis of the results.

ESRI has developed three standard tools for crash analysis in ArcMap including the Spot, Strip and Sliding Scale analyses. Firstly, the spot analysis helps analyze crashes within a specified distance of selected points. The points identify locations of interest for analysis (typically intersections), and the distance represents a buffer around those points (typically 250 feet). The tool will output a data set containing all crashes occurring within the specified buffers. The spot analysis tool allows the user to quickly place points at an intersection and analyze the safety at selected intersection by running the spot analysis query.

Secondly, the strip analysis tool helps analyze roadway segments. The tool breaks the input roadway system into segments that contain a user-specified number of crashes. A user defined window length is selected along with a minimum crash threshold. The tool works by laying the window over each roadway segment and counting
the number of crashes for each window. Any window that has at least the specified number of crashes will be output. When utilizing this approach, each intersection will typically be split in the center, creating four different approaches/segments that include the intersection crashes and diluting the effects of the intersection. This analysis has a greater chance of including both intersection approaches in the same segment. This tool also serves as a method to analyze only segments that have a meaningful number of crashes. Pre-defined segments often lead to segments with very high crash rates but a low number of occurrences, making analysis difficult and countermeasures not financially viable. The tool allows the analyst to ignore the segments that do not meet the minimum crash frequency threshold and focus on areas where treatments may have a significant impact. It can be used for identifying high crash corridors with approved critical crash frequency criterion threshold.

Likewise, the sliding scale analysis is suitable for corridor searching and crash data analysis. The sliding scale is very similar to the strip analysis except that it moves the window along the network routes increments rather than segments of predetermined lengths. This approach may be even more beneficial for corridor screening since there is more flexibility in where the segment begins and ends and a hazard corridor can be identified automatically. This tool aggregates all adjacent candidate segments into one long segment and form a corridor. This raises another concern when the designated segment is too long to manage especially in rural areas.

A study conducted by South Dakota State University [89] adopted the sliding scale method and accounted for this problem. This study does not aggregate the segments to one linear feature; instead, only segments of a specified length limit were output for the analysis. However, the segments with high number of crashes might overlap in between and the researchers have to break the state into small regions to reduce the processing time. In short, the sliding scale approach is similar to the sliding window
and peak searching approaches documented in the HSM but can be achieved directly in ArcGIS environment. This approach is suitable for accurately identifying relatively short corridors that need safety improvements.

2.4 Chapter Summary

Some states across the U.S. have initiated *Safety Corridor Programs* that are dedicated to improving safety along the State’s most crash-prone corridors. The Safety Engineering Division at the Nevada Department of Transportation (NDOT) and the Center for Advanced Transportation Education and Research (CATER) at the University of Nevada, Reno (UNR) conducted a nationwide survey through list server and personal contact to gather related information. It was found that relatively few states have developed procedures to promote urban high crash corridors. The survey revealed major hurdles for corridor level screening. Some of the deterrents include non-availability of corridor level safety performance functions (SPFs) and absence of guidance in regards to the combination of intersection-related crashes into corridor level analysis. The HSM *Safety Performance of Urban and Suburban Arterials* chapter and previous research emphasized the need to fully consider intersection characteristics along a studied corridor, but all intersections contained should be modeled separately. Model-based network screening approaches are highly recommended because of their exceptional merits, such as accounting for regression-to-the-mean bias. Previous research and states’ applications provide valuable information and reference for this research.
Chapter 3  Define Urban Corridors

The lengths of urban corridors may vary from approximately 1 mile in highly developed central business districts to 10 miles or more in sparsely developed urban fringes. The features of urban roadways vary tremendously from place to place. The lengths of arterial, collectors and local streets may also vary in the same city, let alone the safety conditions. Consequently, a universal definition of urban corridors is hard to find. Limits of urban corridors will depend on the nature of the study. Corridors can be determined beforehand or integrated based on hot spot analyses. Some states define an urban corridor in the planning stage as a road stretch between two important intersections along which it is assumed that the traffic conditions and some roadway features are somewhat homogeneous. The purpose of this chapter is to shed light on the question of what should be defined as a corridor by introducing several options.

3.1 Components of Urban Corridors

An urban corridor may be comprised of several intersections and multiple segments with different lengths. Usually sites are assigned to one specific corridor based on engineering judgment. Corridor screening aims at comparing the safety performance of extended corridors, rather than comparing the safety performance of individual sites. Part of the rationale for corridor level screening is that many agencies are taking a proactive approach towards implementing countermeasures with proven effectiveness
over extended corridors to reduce or eliminate traffic crashes before they occur [43]. Therefore, from a corridor perspective, sites are aggregated to investigate the safety performance of a group of intersections and roadway segments. In other words, sites within the same corridor are analyzed as a single entity. This is what makes the corridor screening process unique simply because all other screening analyses (i.e., intersection or segment screening) are performed on a site-by-site basis.

The most intuitive way is to consider the entire focused arterial section as one corridor for the purposes of screening. A roadway section normally comprises a few segments and intersections. The traffic conditions and roadway geometry along the route will be diverse. The safety of the corridor is considered as the average over the entire arterial. This definition is the simplest but has several disadvantages. Relatively more crashes may be expected to occur on longer corridors. A limited budget may not be able to handle the entire corridor at once. According to the advanced techniques documented in the HSM, average conditions should not be used to represent the safety of the entire corridor. It is more reasonable to focus on the locations with elevated crash frequency and severity.

Some states divide roadways into segments of fixed length and consider each segment as a corridor. For example, in the State of Nevada, 0.25 mi long corridors are considered for screening; in New York, 0.3 mi segments are considered. This option is also straightforward and is more suitable for jurisdictions that do not possess predefined corridor limits from the planning perspective. This approach also evaluates and compares the average safety of the defined corridors and is relatively easy to implement. However, the screening results are usually dominated by the high crash intersections along the predefined short sections. Also, this arbitrary decision lacks scientific support.

Another option is to aggregate adjacent hot spot locations (i.e., segments and
intersections) identified based on traditional screening approaches (e.g., sliding mile or peak searching approaches for segments, simple ranking approach for intersections). The highest value of calculated crash frequency of the sites encompassed along the corridor represents the safety of the corridor and should be used to compare with other entities. Different engineers may come up with different definitions for corridors. In real world applications, this option cannot be applied without the exercise of sound engineering judgment.

Furthermore, the HSM defines a roadway segment as a portion of a facility that has a consistent roadway cross-section and has two defined endpoints [66]. The endpoints can be two intersections, a change in roadway cross-section, mile markers or mile posts, or a change in any of the roadway characteristic, such as number of lanes per direction, access density, traffic volume ranges, median type, etc. Homogeneous segments are generated using this guide and similar concept can be adopted to define urban corridors. The cutting points of adjacent corridors mainly depend upon the changes of roadway geometry and traffic conditions. In most collected data scenarios, roadway segments are split at intersections where key attribute values, such as road name, AADT, number of lanes, speed limit, etc. change. The limits of urban corridors can therefore be generated based on the changes of these data items. The author applied this concept to the urban road network in the Reno-Sparks area in Nevada and found that creating homogeneous corridors using this concept would result in many tiny corridor sections (e.g., as short as 0.01 mile) especially when many additional and detailed attributes are involved. In order to conduct a meaningful corridor analysis, a number of sites with short lengths should be treated as a group. Therefore, several short segments were aggregated upward in size thereby increasing the average corridor length of the entire database and removing a large portion of short segments. It can be seen that this way of defining corridors requires massive engineering judgment.
This research proposes a clustering based approach to determine corridor boundaries. Previous research has found that the safety conditions at adjacent intersections and/or segments are correlated because of close distances, similar traffic patterns and geometry. This correlation or association indicates a non random pattern in the data values over a study region (e.g., a corridor, arterial or traffic analysis zone). Nevertheless, corridors in previous research are mostly predetermined and accordingly, the intersections and segments had already been integrated into groups that represent different corridors. This study analyzes corridors from a reverse perspective in this regard by clustering adjacent intersections and/or segments with similar spatial features, traffic patterns and geometry into corridors. It is proposed to cluster signalized intersections along urban arterial roads based on their spatial auto-correlations. The proposal is for the following reasons. First and foremost, signalized intersections are a major source of traffic crashes and urban congestion. In the U.S., signalized intersections comprise less than 10% of the total number of intersections but about 25% of total crashes and 35% of intersection fatal crashes occurred at signalized intersections. Additionally, a couple of adjacent signalized intersections are usually coordinated during peak hours in order to maximize corridor bandwidth so that similar traffic patterns are formed in-between intersections in the same corridor. The details pertaining to the proposed approach are presented in the next sections.

3.2 Spatial Autocorrelation of Signalized Intersections

According to Tobler’s First Law of Geography,

“Everything is related to everything else, but near things are more related than distant things.” [100]

The spatial autocorrelation concept is used to describe the coincidence of variables similarity with location proximity [8, 19, 101]. Assume that there are \( n \) sites in the
study region, \( x_i \ (i = 1, ..., n) \) are the values of an uni-variate \( X \) on \( n \) sites. The concept of spatial auto-correlation indicates that pairs of subjects (e.g., \( i \) and \( j \)) that are close to each other are more likely to have more similar variable values (i.e., \( x_i \) and \( x_j \)). Spatial auto-correlation indexes are often used to measure the spatial association in the data considering both locational and attribute information. The location information refers to the latitude and longitude information whereas the attribute information represents the features (e.g., AADT, traffic control type, number of through lanes, number of left-turn lanes, etc.) of a roadway facility. The existence of spatial auto-correlation can be measured using two indexes: 1) global measures, that summarize the averaged spatial association with respect to the whole region, and 2) local measures, that refer to the association of a single site in relation to its neighborhoods. Both types are briefly discussed below.

### 3.2.1 Global Measures of Spatial Autocorrelation

There are two indexes that are often used to describe the global spatial auto-correlation. The first one is known as the Moran’s I which was proposed by Moran in 1950 [71]. An alternative is the Getis-Ord statistic which was proposed by Getis and Ord in 1992 [35]. In short, global indexes of spatial auto-correlation provide summaries over the entire study area for the similarity of variable values among neighboring observations.

#### 3.2.1.1 Moran’s I

Moran’s I is one of the oldest and most popular indicators of spatial auto-correlation. It can be applied to either zones or points with continuous variables associated. Moran’s I takes the form of a classic correlation coefficient in which the mean of a variable is subtracted from each sample value in the numerator as formula (3.1)
where, $x_i$ are the data values of a random variable $X$ on $n$ sites in the study region, $\bar{x}$ is the mean of $X$, $w_{ij}$ is a measure of spatial contiguity between data sites $i$ and $j$ and $w_{ij}$ forms the spatial weight matrix $W$.

Moran’s I index has a range from -1 to 1 where values between -1 and 0 indicate a negative association, values between 0 and 1 indicate a positive association between variables, and 0 indicates that there is no correlation between variables. To calculate spatial auto-correlation measurement, a weight matrix is required. In this study, the matrix is generated based on the inverse distance weights in which the distances are calculated using intersections’ latitude and longitude information. The weights for pairs of intersections that are close together are higher than for pairs of intersections that are far apart.

### 3.2.1.2 Getis-Ord Index

Another widely used index to describe global spatial auto-correlation is the Getis-Ord (G-O) index and is calculated as:

$$ G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} x_i x_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j} $$

Compared to the Moran’s I measurement, G-O’s G index takes values on $[0,1]$. G-O’s G value close to 1 indicates clustering of high values while G-O value close to 0 indicates clustering of low values. G-O’s G is able to distinguish the clustering of high and low values but cannot capture the presence of negative spatial auto-correlations.
Moran’s I, on the other hand, is able to detect both positive and negative spatial auto-correlations but cannot cluster high or low values. Therefore, they can be used as complementary tools to detect spatial correlation and also cluster units. The inference of either Moran’s I or Getis-Ord index is based on normal distribution so standardized $z$-values can be computed and significance tests can be performed for the statistics.

Next, the Kietzke Ln arterial in Reno, Nevada is used as a demonstration example. Currently there are a total of 25 intersections and 11 signalized intersections along Kietzke Ln. Table 3.1 presents the recommended features (or variables) from the HSM that may impact the safety of signalized intersections. They can be broadly divided into three categories: 1) traffic patterns, 2) intersection geometry, and 3) signalization.

Table 3.1: Global Measures of Spatial Auto-correlation for Selected Multivariates on Kietzke Ln in Reno, Nevada

<table>
<thead>
<tr>
<th>Selected Features</th>
<th>Abbreviation</th>
<th>Moran’s I</th>
<th>p-value</th>
<th>Getis-Ord(G)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT of 2007</td>
<td>AADT07</td>
<td>0.2009</td>
<td>0.0009</td>
<td>0.6520</td>
<td>0.0003</td>
</tr>
<tr>
<td>AADT of 2008</td>
<td>AADT08</td>
<td>0.3205</td>
<td>0.0025</td>
<td>0.7250</td>
<td>0.0001</td>
</tr>
<tr>
<td>AADT of 2009</td>
<td>AADT09</td>
<td>0.1523</td>
<td>&lt;10e-5</td>
<td>0.6255</td>
<td>0.0000</td>
</tr>
<tr>
<td>AADT of 2010</td>
<td>AADT10</td>
<td>0.1822</td>
<td>0.0000</td>
<td>0.6588</td>
<td>0.0001</td>
</tr>
<tr>
<td>AADT of 2011</td>
<td>AADT11</td>
<td>0.2250</td>
<td>0.0025</td>
<td>0.7004</td>
<td>&lt;10e-6</td>
</tr>
<tr>
<td>Number of through lanes</td>
<td>NTL</td>
<td>0.8802</td>
<td>&lt;10e-7</td>
<td>0.9252</td>
<td>&lt;10e-5</td>
</tr>
<tr>
<td>Number of approaches with left turn lanes</td>
<td>NAWLTL</td>
<td>0.1225</td>
<td>0.0480</td>
<td>0.1990</td>
<td>0.0500</td>
</tr>
<tr>
<td>Number of approaches with right turn lanes</td>
<td>NAWRTL</td>
<td>-0.1572</td>
<td>0.0462</td>
<td>0.1600</td>
<td>0.0362</td>
</tr>
<tr>
<td>Cycle lengths</td>
<td>CL</td>
<td>0.6529</td>
<td>&lt;10e-8</td>
<td>0.8943</td>
<td>&lt;10e-5</td>
</tr>
<tr>
<td>Presence of left turn phasing</td>
<td>PLTP</td>
<td>-0.0025</td>
<td>0.1025</td>
<td>0.0002</td>
<td>0.1590</td>
</tr>
<tr>
<td>Type of left turn phasing</td>
<td>TLTP</td>
<td>-0.0171</td>
<td>0.2856</td>
<td>0.0017</td>
<td>0.3177</td>
</tr>
</tbody>
</table>
To investigate the spatial auto-correlation of 2007 AADT data element along Kietzke Ln, for instance, the Moran’s I and G-O index are calculated using formula (3.1) and (3.2), respectively. Latitude and longitude coordinates of signalized intersections are used to generate the matrix of inverse distance weights \((w_{ij})\). The calculated Moran’s I value is equal to 0.2009 with a p-value of 0.00088, and G-O index equals 0.652. Based on this result, the null hypothesis that there is zero spatial auto-correlation present in the variable AADT is rejected at a significance level of 0.05.

Both Moran’s I and G-O index proved the existence of spatial auto-correlation and provide an average of spatial dependency over the study route. Nevertheless, global measures cannot be directly used to define corridors. Instead, local measurements were proposed to identify individual clusters within the study routes.

### 3.2.2 Local Measures of Spatial Autocorrelation

There are two indexes to describe local spatial auto-correlation, well known as the local Moran’s I and local Getis-Ord indices. Local measures aim at identifying patterns of spatial dependency within a certain arterial road.

**Local Moran’s I**

The local Moran’s I index, proposed by Anselin in 1995 [6], is defined as:

\[
I_i = n(x_i - \bar{x}) \frac{\sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2}
\]  

(3.3)

Formula (3.3) is similar to (3.1) and the relationship between Moran’s I and local
Moran’s I is:

\[ I = \sum_i \frac{I_i}{\sum_i \sum_j w_{ij}} \]  

(3.4)

Also similar to Moran’s I, large positive values of this index indicate local clustering of data values around site \(i\). The expected value of \(I_i\) under the complete randomization assumption is given by:

\[ E(I_i) = \frac{-w_i}{n-1} \]  

(3.5)

The variance is given by [6]:

\[ Var(I_i) = \frac{-w_{i(2)}(n - b_2)}{n - 1} + \frac{2w_{i(kh)}(2b_2 - n)}{(n - 1)(n - 2)} - \frac{w_i^2}{(n - 1)^2} \]  

(3.6)

\[ b_2 = \frac{m_4}{m_2^2} \]  

(3.7)

\[ m_r = \frac{\sum_i (x_i - \bar{x})^r}{n} \]  

(3.8)

\[ w_i = \sum_j w_{ij} \]  

(3.9)

\[ w_{i(2)} = \sum_{i \neq j} w_{ij}^2 \]  

(3.10)

\[ w_{i(kh)} = \sum_{k \neq i} \sum_{h \neq i} w_{ik}w_{ih} \]  

(3.11)

The local Moran’s I is a member of the class of so-called “local indicators of spatial association” (LISA). A LISA statistic must satisfy two requirements: (a) indicate the extent of significant spatial clustering for each location, and (b) the sum of local statistics is proportional to a global indicator of spatial association [6], which was demonstrated in (3.4).
3.2.2.1 Local G-O Index

Getis and Ord also proposed a local G-O index in the form of (3.12) [35].

\[ G_i = \frac{\sum_{j=1}^{n} w_{ij}x_j}{\sum_{j=1}^{n} x_j} \]  \hspace{1cm} (3.12)

Similar to the local Moran’s I, the expected value and variance are given by:

\[ E(G_i) = \frac{w_i}{n} \]  \hspace{1cm} (3.13)

\[ Var(G_i) = \frac{w_i(n - w_i)}{n^2(n - 1)} \frac{s^2}{\bar{x}^2} \]  \hspace{1cm} (3.14)

\[ \bar{x} = \frac{\sum_{i} x_i}{n} \]  \hspace{1cm} (3.15)

\[ s^2 = \frac{\sum_{i}(x_i - \bar{x})^2}{n} \]  \hspace{1cm} (3.16)

The mostly used local G-O index is the standardized statistic given by:

\[ z(G_i) = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{x}w_i}{\sqrt{\frac{s^2}{n-1} \left( n \sum_{j=1}^{n} w_{ij}^2 - w_i^2 \right)}} \]  \hspace{1cm} (3.17)

For the Kietzke Ln example, the calculated local spatial measures for each variable are presented in Table 3.2.
<table>
<thead>
<tr>
<th>Intersect Arterial</th>
<th>AADT07</th>
<th>AADT08</th>
<th>AADT09</th>
<th>AADT10</th>
<th>AADT11</th>
<th>NTL</th>
<th>NAWLTL</th>
<th>NAWRTL</th>
<th>CL</th>
<th>PLTP</th>
<th>TLTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victorian Ave</td>
<td>1.1020</td>
<td>0.8253</td>
<td>0.9523</td>
<td>0.8120</td>
<td>1.5203</td>
<td>0.6255</td>
<td>-1.6252</td>
<td>0.3252</td>
<td>-1.8223</td>
<td>0.8563</td>
<td>-1.2033</td>
</tr>
<tr>
<td>Galletti Way</td>
<td>0.7211</td>
<td>1.3746</td>
<td>0.4955</td>
<td>1.5288</td>
<td>1.6233</td>
<td>0.6522</td>
<td>1.8955</td>
<td>0.2417</td>
<td>-1.6577</td>
<td>0.7419</td>
<td>-1.5005</td>
</tr>
<tr>
<td>E 2nd St</td>
<td>0.8760</td>
<td>1.2314</td>
<td>0.5068</td>
<td>1.0123</td>
<td>2.2158</td>
<td>0.5687</td>
<td>-1.7716</td>
<td>0.1635</td>
<td>-1.6947</td>
<td>0.7511</td>
<td>-1.3557</td>
</tr>
<tr>
<td>Mill St</td>
<td>-0.8415</td>
<td>-0.4328</td>
<td>-0.5483</td>
<td>-0.4823</td>
<td>-0.3702</td>
<td>0.6112</td>
<td>-1.9061</td>
<td>0.2227</td>
<td>-1.8468</td>
<td>0.7696</td>
<td>-1.4774</td>
</tr>
<tr>
<td>Vassar St</td>
<td>0.6414</td>
<td>0.7718</td>
<td>0.3138</td>
<td>0.5701</td>
<td>0.8447</td>
<td>0.3492</td>
<td>-1.2157</td>
<td>0.6742</td>
<td>-1.4439</td>
<td>-0.0525</td>
<td>-1.7327</td>
</tr>
<tr>
<td>E Plumb Ln</td>
<td>1.5658</td>
<td>1.7168</td>
<td>1.4238</td>
<td>1.4662</td>
<td>0.4792</td>
<td>-0.0443</td>
<td>0.3424</td>
<td>1.5402</td>
<td>-0.4584</td>
<td>-1.4400</td>
<td>-0.5500</td>
</tr>
<tr>
<td>E Grove St</td>
<td>1.0766</td>
<td>1.0716</td>
<td>0.5899</td>
<td>0.7961</td>
<td>0.9087</td>
<td>-1.0026</td>
<td>-0.4902</td>
<td>0.9111</td>
<td>-0.9019</td>
<td>-0.5055</td>
<td>-1.0822</td>
</tr>
<tr>
<td>Gentry Way</td>
<td>1.5630</td>
<td>1.4607</td>
<td>1.2946</td>
<td>1.1755</td>
<td>-0.0716</td>
<td>0.6041</td>
<td>1.1742</td>
<td>1.6797</td>
<td>0.2292</td>
<td>-1.9708</td>
<td>0.5125</td>
</tr>
<tr>
<td>E Moana Ln</td>
<td>1.3069</td>
<td>1.1911</td>
<td>1.0510</td>
<td>0.8625</td>
<td>-0.2270</td>
<td>-3.0621</td>
<td>1.3154</td>
<td>1.5914</td>
<td>0.4019</td>
<td>-1.9686</td>
<td>1.2139</td>
</tr>
<tr>
<td>E Packham Ln</td>
<td>-0.0147</td>
<td>-0.2312</td>
<td>-0.2850</td>
<td>-0.2327</td>
<td>-1.4953</td>
<td>0.3644</td>
<td>1.6041</td>
<td>-0.3267</td>
<td>1.6238</td>
<td>-0.5330</td>
<td>0.3837</td>
</tr>
<tr>
<td>S Virginia St</td>
<td>0.5857</td>
<td>0.1485</td>
<td>1.0377</td>
<td>0.4949</td>
<td>0.3106</td>
<td>-1.3793</td>
<td>1.7155</td>
<td>0.2508</td>
<td>1.4365</td>
<td>-1.0643</td>
<td>3.2123</td>
</tr>
<tr>
<td>S Mccarran Blvd</td>
<td>-1.2010</td>
<td>-1.5687</td>
<td>-1.0248</td>
<td>-1.2415</td>
<td>-1.3014</td>
<td>0.0849</td>
<td>1.6007</td>
<td>-1.5482</td>
<td>2.2315</td>
<td>0.4749</td>
<td>4.9901</td>
</tr>
<tr>
<td>Sierra Rose Dr</td>
<td>-0.9109</td>
<td>-1.3109</td>
<td>-0.5816</td>
<td>-0.8146</td>
<td>-0.8904</td>
<td>0.1550</td>
<td>1.4782</td>
<td>-1.5250</td>
<td>2.1084</td>
<td>0.5171</td>
<td>2.2524</td>
</tr>
</tbody>
</table>
3.3 Cluster Signalized Intersections

The next step is to cluster adjacent signalized intersections into corridors. Previous research had shown that neighboring sites typically have similar environmental, geographic characteristics, and traffic patterns, and thereby form a cluster that has similar crash occurrence [29]. Hence, the goal of clustering signalized intersections is to find groups of adjacent intersections and segments such that the sites within a cluster are as similar as possible, whereas sites from different clusters are as dissimilar as possible. The clustering considers both multivariate profiles of each intersection and their spatial distributions. The local spatial auto-correlation measures are used as input of clustering algorithm.

In statistical literature, clustering has been extensively applied in many areas such as machine learning and data mining. Rare applications were found in the transportation field. Most clustering algorithms can be broadly divided into two categories: 1) partitioning methods, which seeks to divide study sites into several classes; and 2) hierarchical methods, which seeks to produce nested sequence of clusters [95]. As for the nature of this study, the popular K-means partitioning algorithm is selected for the corridor analysis.

Given \( n \) signalized intersections along an urban arterial spanned by a multivariate set of variables \( x_{ij} \) (\( i = 1, 2, ..., n; j = 1, 2, ..., p \)), the \( K \)-means algorithm seeks to assign each intersection to one of \( K \) corridors. It seeks to partition \( n \) intersections into \( K \) sets \( S_k \) (\( k = 1, 2, ..., K \)) so that the within-cluster sum of squares (\( WCSS \)) (formula (3.18)) is minimized.

\[
WCSS = \sum_{k=1}^{K} \sum_{i=1}^{n} \|x_{i}^{k} - c_{k}\|^2 \quad (3.18)
\]

where, \( x_{i}^{k} \) indicates that site \( i \) belongs to cluster \( k \), \( c_{k}(k = 1, 2, ..., K) \) are selected
centroid for cluster $k$, $\|x_k^i - c_k\|^2$ represents the distance between an intersection and the cluster’s centroid, and $\|x_k^i - c_k\| = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$ is the Euclidean distance to represent the within-cluster dissimilarity. The identification of corridors can be achieved based on the following steps.

**Step 1: Create local spatial auto-correlation measures matrix $Z$**

Given the spatial weight matrix obtained using latitude and longitude information, standardized local G-O index $z(G_j(x_i))$ is computed for the $j^{th}$ ($j = 1, 2, ..., p$) variable at each site $i$ ($i = 1, 2, ..., n$), and $z(G_j(x_i))$ forms the matrix $Z$ of dimension $n \times p$. Each column of matrix $Z$ expresses the local G-O index $z(G_j(x))$ for variable $j$, while each row ($z(G(x_i))$) of $Z$ provides the clustering profile around each intersection $i$. As for the Kietzke Ln example, Table 3.2 represents the matrix $Z$.

**Step 2: Choose the optimal number of clusters $K$**

Determining the number of clusters is the critical step in the $K$-means algorithm. Three approaches are widely used to choose the optimal number including: 1) simple rule of thumb: $K \approx \sqrt{n/2}$; 2) the elbow method [95]; and 3) the gap statistic [97]. Similar to the rule of thumb, the elbow method is also an empirical method. It looks for the bend or “elbow” point in the plot of the error measure, sum of squared errors ($WCSS$) versus the number of clusters $k$ employed. Figure 3.1 is the plot of $WCSS$ versus $k$ for the Kietzke Lane example. It can be seen that the measure $WCSS$ decreases monotonically as $k$ increases but from $k = 3$ or $k = 4$ the decrease trend flattens markedly. The location of the so-called “elbow” indicates the appropriate number of clusters. However, the “elbow” point may not always be identified unambiguously. A relatively more sophisticated approach is the Gap Statistic.

The gap statistic was proposed by Tibshirani et al. in 2000. The idea is to standardize the comparison of $logW_k$ ($W_k = WCSS$) with the null reference distribution
of the data. The optimal number of clusters is the value of $k$ for which $\log W_k$ falls the farthest below this reference curve. The gap statistic is given by,

$$\text{Gap}_n(k) = E_n^*[\log W_k] - \log W_k$$  \hspace{1cm} (3.19)

To obtain the values of $E_n^*[\log W_k]$, the average of $B=100$ copies of $\log W_k$ generated using a Monte Carlo sample from the reference distribution is calculated. For each replicate of Monte Carlo samples, a standard deviation $sd(k)$ is obtained and,

$$sd(k) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (\log W_{kb}^* - \frac{1}{B}(\log W_{kb}^*))^2}$$  \hspace{1cm} (3.20)

and

$$s_k = \sqrt{1 + \frac{1}{B}sd(k)}$$  \hspace{1cm} (3.21)
The optimal number of clusters $K$ is the smallest value of $k$ that satisfies $Gap(k) \geq Gap(k + 1) - s_{k+1}$. Other details can be seen in the original paper. Figure 3.2 shows the gap statistics of the Kietzke Ln example.

![Figure 3.2: Determining the Number of Corridors based on Gap Statistic](image)

**Step 3: Apply K-means algorithm to cluster signalized intersections into corridors**

$K$ data points are randomly selected as cluster centers and the multivariate values as presented in Table 3.2 are assigned to clusters. This process is repeated for several iterations to minimum the within-cluster sum of squares.

The clustering results for Kietzke Ln are presented below. Overall, this 4.5 mile long arterial is divided into three corridors: 1) Corridor 1: from $N$ Kietzke Ln & Galletti Way to Kietzke Ln & E Plumb Ln, 2) Corridor 2: from Kietzke Ln & E Plumb Ln to $S$ Virginia St, and 3) Corridor 3: from $S$ Virginia St to Del Monte Ln roundabout.
3.4 Chapter Summary

Corridors are most likely defined based upon project needs using subjective sectioning. This chapter proposes an objective way to define urban corridor boundaries based on the spatial auto-correlations of signalized intersections. As a key component of urban corridors, signalized intersections play an important role in vehicle mobility and safety. Most fatal and serious injury crashes occurred at signalized intersections especially in urban areas. The underlying reasons include but not limited to more conflicting points, traffic volumes, pedestrian presence, etc. Previous research proposed that signalized intersections close by may be correlated because of similar traffic pattern, land use and intersection geometries. In these cases, corridors were predefined. A reverse action was taken in this regard to define corridors from the perspective of the spatial correlations of signalized intersections. Global and local measures of spatial auto-correlation for signalized intersections indicate significant correlations between signalized intersections within arterial roads. This finding is consistent with findings from previous research. In short, urban corridor boundaries are defined based on the following integrated steps:

1. Select and define features to describe signalized intersections based on data availability;
2. Calculate the local spatial measures as the input of the clustering algorithm; and
3. Utilize the partitioning algorithm to cluster signalized intersections and adjacent intersections that fall into the same cluster form a corridor.

The same procedure was repeated for 32 principal arterial roads and 46 minor arterial roads in Reno-Sparks, Nevada. For each arterial road, the method was repeated for 30 times until the results stop changing. A total of 152 urban corridors
were generated with minimum length of 0.25 mile and maximum length of 4.62 miles. A summary of the basic statistics of defined urban corridors is presented in Table 3.3.
Table 3.3: Summary Statistics of Designated Corridors in the Study Region

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>STDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>Corridor lengths (miles)</td>
<td>0.2558</td>
<td>4.6218</td>
<td>1.9666</td>
<td>0.7002</td>
</tr>
<tr>
<td>$K$</td>
<td>Number of segments</td>
<td>1</td>
<td>15</td>
<td>5.1815</td>
<td>3.2351</td>
</tr>
<tr>
<td>$J$</td>
<td>Number of intersections</td>
<td>0</td>
<td>15</td>
<td>3.9020</td>
<td>3.7854</td>
</tr>
<tr>
<td>IntDen</td>
<td>Number of intersections per mile</td>
<td>0</td>
<td>12.8103</td>
<td>4.8095</td>
<td>5.2386</td>
</tr>
<tr>
<td>$l$</td>
<td>Segment lengths (miles)</td>
<td>0.1558</td>
<td>1.0562</td>
<td>0.3526</td>
<td>0.3205</td>
</tr>
<tr>
<td>$AADT_{seg}$</td>
<td>AADT on segments</td>
<td>3800</td>
<td>43500</td>
<td>19805</td>
<td>9902.25</td>
</tr>
<tr>
<td>$P_{two-lane}$</td>
<td>Proportion of two-lane arterial along corridors</td>
<td>0</td>
<td>1</td>
<td>0.2811</td>
<td>0.2850</td>
</tr>
<tr>
<td>$P_{multilane-undi}$</td>
<td>Proportion of multilane undivided arterial along designated corridors</td>
<td>0.36</td>
<td>1</td>
<td>0.5107</td>
<td>0.4216</td>
</tr>
<tr>
<td>$P_{multilane-di}$</td>
<td>Proportion of multilane divided arterial along designated corridors</td>
<td>0</td>
<td>0.81</td>
<td>0.3199</td>
<td>0.3327</td>
</tr>
<tr>
<td>$P_{3leg}$</td>
<td>Proportion of 3-legged intersections</td>
<td>0</td>
<td>0.61</td>
<td>0.2219</td>
<td>0.1025</td>
</tr>
<tr>
<td>$P_{4leg}$</td>
<td>Proportion of 4-legged intersections</td>
<td>0</td>
<td>1</td>
<td>0.5217</td>
<td>0.3155</td>
</tr>
<tr>
<td>$P_{sig}$</td>
<td>Proportion of signalized intersections</td>
<td>0</td>
<td>1</td>
<td>0.3255</td>
<td>0.1216</td>
</tr>
<tr>
<td>$P_{unsig}$</td>
<td>Proportion of unsignalized intersections</td>
<td>0</td>
<td>0.68</td>
<td>0.2100</td>
<td>0.0821</td>
</tr>
<tr>
<td>$AADT_{maj}$</td>
<td>AADT for major approach</td>
<td>5000</td>
<td>80000</td>
<td>15523</td>
<td>6852.03</td>
</tr>
<tr>
<td>$AADT_{min}$</td>
<td>AADT for minor approach</td>
<td>3000</td>
<td>55000</td>
<td>12030</td>
<td>6205.88</td>
</tr>
<tr>
<td>$P_{3leg-mistop}$</td>
<td>Proportion of 3-legged intersections with minor road stop control</td>
<td>0</td>
<td>0.56</td>
<td>0.1270</td>
<td>0.0521</td>
</tr>
<tr>
<td>$P_{3leg-AWSC}$</td>
<td>Proportion of 3-legged intersections with all-way stop control</td>
<td>0</td>
<td>0.37</td>
<td>0.1924</td>
<td>0.0745</td>
</tr>
<tr>
<td>$P_{3leg-sig}$</td>
<td>Proportion of 3-legged intersections with signal</td>
<td>0</td>
<td>0.80</td>
<td>0.3763</td>
<td>0.2015</td>
</tr>
<tr>
<td>$P_{4leg-TWSC}$</td>
<td>Proportion of 4-legged intersections with minor road stop control</td>
<td>0</td>
<td>0.68</td>
<td>0.1052</td>
<td>0.0575</td>
</tr>
<tr>
<td>$P_{4leg-AWSC}$</td>
<td>Proportion of 4-legged intersections with all-way stop control</td>
<td>0</td>
<td>0.50</td>
<td>0.1179</td>
<td>0.0026</td>
</tr>
<tr>
<td>$P_{4leg-sig}$</td>
<td>Proportion of 4-legged intersections with signal</td>
<td>0</td>
<td>1</td>
<td>0.4215</td>
<td>0.1005</td>
</tr>
</tbody>
</table>

Chapter 4  Techniques on Crash Prediction Models

A sound corridor screening analysis requires good estimates of crash risks at individual intersections and segments. Traffic crashes are random, rare, discrete, non-negative and sporadic events. This nature is closely related to two well established statistical problems, the “small area estimation” problem [91] and the “disease mapping” problem. The small area problem arises when sample survey data for a small geographical area is used. The disease mapping problem stems from estimating rare disease or incidents (i.e., traffic crash) for a small geographical area. These two issues cause high random variability in crash data. The main statistical problems associated are that the reliable estimates of model parameters are hard to obtain due to small sample size and the rareness of crashes. Over the past few decades, the model-based approach has been a prevailing choice for analyzing traffic crashes. Since the late 19th century, the Poisson process has been used to model traffic crashes. With enormous progress in mathematics and computing capabilities, more complex modeling schemes have been developed to improve model fitting.

To discuss crash prediction models, it is useful to classify the likely effects on the distribution of crashes into three categories:

• Poisson variation;

• Heterogeneity which indicates the extra variation due to within-site effects reflecting its individual characteristics; and
• Spatial effects that represent the similarities of neighboring sites.

Most of the literature related to the development of crash prediction models accounts for the first two types of variation. Recently, researchers often add the third effect, spatial correlation, into the model development procedures. For example, recent publications covered intersection crash prediction models considering correlations along corridors, direct spatial correlations between segments, and random parameters for corridors. It has been demonstrated that spatial dependence can be a surrogate for unknown and relevant co-variate, thereby improving model estimation. Previous studies showed that ignoring spatial dependence leads to underestimation of variability [10, 20, 29].

This chapter discusses the statistical techniques pertaining to traffic crash prediction models. The HSM recommended Poisson-Gamma regression model is extended to other Poisson mixtures. Further steps were made towards spatially adjusted models for predicting traffic crashes.

4.1 Crash Prediction Models

Over the last few decades, statistical models have been extensively applied in traffic safety analyses to identify contributing factors to crashes, establish relationships between crash and explanatory variables and predict crash frequencies. The accurate characterization of the distribution of traffic crash data is important for safety risk assessment. Statistical distributions that adequately model traffic crash are often related to the Poisson mixture distributions, especially when sampling data is integer-valued.

Crash counts are highly skewed and include solely positive integers. The common assumption is that crash counts are approximately Poisson distributed. This serviceable approximation has been known for a long time [31]. The first application was
initiated by L. von Bortkiewitz in 1898 who had data of the number of deaths by horse-kick in ten Prussian army corps over 20 years. When comparing the number of years with number of deaths to the number predicted by the Poisson distribution the fit was remarkably good. Quine and Seneta rechecked Bortkiewitz’s data and claimed that the main reason for the good fit was that the numbers were small and the variations from year to year was small too [90]. Attempts to examine if the reported crash counts actually obey the Poisson distribution found that “there is little (if any) justification for rejecting it [74, 75, 76].”

4.1.1 Generalized Poisson Regression Model

The Poisson regression model is a standard probability model for traffic crash data. Nonetheless, the Poisson model is somewhat limited in that the conditional mean and variance of crashes are modeled by a single parameter as shown below.

\[
Y_i | \theta_i \sim \text{Poisson}(\theta_i), \ i = 1, 2, ..., n \tag{4.1}
\]

\[
P(Y_i = y_i | \theta_i) = \frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!} \tag{4.2}
\]

\[
\theta_i = E(Y_i | \theta_i) = Var(Y_i | \theta_i) = e^{(X^T \beta)} \tag{4.3}
\]

where, \( Y_i \) which represents crash count at site \( i \) is considered as a random variable following a Poisson distribution with parameter \( \theta_i > 0 \); \( y_i \) is the observed number of crashes at site \( i \); \( X \) is a vector of covariates; \( \beta \) represents a vector of regression parameters associated with corresponding covariates to be estimated by maximum likelihood using the iterative weighted least squares (IWLS) algorithm. Assume the crash counts for sites \( i = 1, 2, ..., n \) are independent and Poisson distributed with means \( \theta_1, \theta_2, ..., \theta_n \), the likelihood function is obtained by multiplying those independent events.
Figure 4.1: Poisson Distribution
\[ L(y_1, y_2, \ldots, y_n | \theta_i) = \prod_{i=1}^{n} \frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!} \]  

(4.4)

The log-likelihood is:

\[ \ln[L(y_1, y_2, \ldots, y_n | \theta_i)] = \ln \prod_{i=1}^{n} \frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!} = \sum_{i=1}^{n} \ln \frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!} = \sum_{i=1}^{n} \left( -\theta_i + y_i \ln \theta_i - \ln(y_i!) \right) \]  

(4.5)

The term \( \ln(y_i!) \) is a constant that does not depend on any parameter so the abridged log-likelihood will be:

\[ \ln[L^*(y_1, y_2, \ldots, y_n | \theta_i)] = \sum_{i=1}^{n} \left( -\theta_i + y_i \ln \theta_i \right) \]  

(4.6)

As presented earlier, \( \theta_i = e^{(X^T \beta)} \), replacing \( \theta_i \) by model parameters, the likelihood function became a function of parameters. Fitting the model function to data amounts to finding those values of parameters which maximize the \( \ln[L^*(\beta_i)] \).

\[ \frac{\partial \ln[L^*(\cdot)]}{\partial \beta} = \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \left( -\theta_i + y_i \ln \theta_i \right) = \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \left( -e^{(X^T \beta)} + y_i (X^T \beta) \right) = 0 \]  

(4.7)

The Poisson likelihood function is rather simple and the results can be manually calculated. In reality, the conditional variance will exceeds the conditional mean, which is widely known as over-dispersion. This may result from unobserved heterogeneity. Thus, most applications introduce an error term \( \varepsilon_i \) to \( \theta_i \) to create a more flexible Poisson formulation to accommodate the extra variability.
4.1.2 Generalized Poisson Mixture Regression Models

By introducing the error term to Poisson distribution,

\[ Y_i | \mu_i \sim \text{Poisson}(\mu_i), \ i = 1, 2, ..., n \]  

\[ \mu_i = e^{(X^T \beta + \epsilon_i)} = \theta_i e^{\epsilon_i} = \theta_i \epsilon_i \]  

\[ P(Y_i = y_i | \theta_i, \epsilon_i) = \frac{(\theta_i \epsilon_i)^y e^{-\theta_i \epsilon_i}}{y_i!} \]  

The formulas above imply that: 1) each intersection has its own expected safety condition, 2) the safety condition differs for different intersections even if their characteristics are similar in some relevant way, and 3) the safety conditions of similar intersections (aka intersections with a certain set of characteristics) are assumed to have a common overall mean. The last statement indicates that the mean number of crashes at similar intersections is assumed to be draws from a common population distribution with its own mean (\( \mu_{\epsilon_i} \)) and variance (\( \sigma_{\epsilon_i}^2 \)).

There is a big list in the Poisson mixture family. However, the mostly widely used mixture distributions are Poisson-Gamma, Poisson-Lognormal, and zero-inflated models. Other mixtures are not gaining attention because of complex derivations and interpretation. The Zero-Inflated models were derived because naturally there are many zeros in the crash database. However, previous research argued the plausibility of using such models and concluded that the values of zeros do not necessarily indicate zero-inflated [62, 63]. The assumption of Poisson mixtures is crucial when developing crash prediction models. This section shed light on some of the most popular mixtures and proposed another mixture by introducing the Poisson-Inverse Gaussian regression models.
4.1.2.1 Poisson-Gamma Model

The mixed Poisson distribution depends on the specific distribution of \( \epsilon_i \). The most frequently used model for \( \epsilon_i \) is the Gamma distribution due to its convenience and flexibility. Assume that \( g(\epsilon_i) \) is the Gamma probability density function of \( \epsilon_i \) with mean equals to 1 and variance of \( 1/\phi \) (in other words, the parameters of the Gamma distribution \( \alpha = \beta = \phi \)). Then the resulting marginal density of crash counts \( Y_i \) is the Negative Binomial Distribution.

\[
g(\epsilon_i; \phi) = \frac{\phi^\phi \epsilon_i^{\phi-1} e^{-\epsilon_i \phi}}{\Gamma(\phi)} \tag{4.11}
\]

The marginal distribution for \( Y_i \) can be obtained by integrating \( \epsilon_i \) as:

\[
P(Y_i = y_i | \theta_i, \phi) = \int f(y_i | \theta_i, \epsilon_i) g(\epsilon_i) d\epsilon_i = \int \frac{(\theta_i \epsilon_i)^{y_i} e^{-\theta_i \epsilon_i} \phi^\phi \epsilon_i^{\phi-1} e^{-\epsilon_i \phi}}{y_i!} \frac{\Gamma(\phi)}{\Gamma(\phi)} d\epsilon_i
\]

\[
= \frac{\Gamma(y_i + \phi)}{\Gamma(\phi) \Gamma(y_i + 1)} \left( \frac{\phi}{\theta_i + \phi} \right)^\phi \left( \frac{\theta_i}{\theta_i + \phi} \right)^{y_i} \tag{4.12}
\]

The Poisson-Gamma application is based on two main assumptions: 1) crash counts on every site are Poisson distributed, and 2) the means of the sites, \( \mu_i \) (of which a population is comprised) come from a Gamma distribution. The plausibility of the first assumption resides in the similarity between the realities of crash occurrence and the conditions from which the Poisson distribution is derived by logic. The second assumption has no similar source of plausibility and there is no obvious reason why the mean \( \mu_i \) should resemble a Gamma distribution. The motivations using a Gamma distribution are convenience and flexibility. Convenience because assumptions 1) and 2) together imply another closed-form probability distribution: the Negative Binomial distribution, and flexibility because the Gamma PDF can take on many shapes.

The inverse dispersion parameter \( (\phi) \) is one of the outputs of negative binomial
Figure 4.2: Poisson-Gamma Distribution
regression model. The parameter is essentially a value that measures how widely dispersed crash data are around the estimated mean. Models that generate a lower over-dispersion parameter are generally considered to be more accurate.

The likelihood functions below can be used to estimate the regression parameters of the Poisson-Gamma model.

\[ L(y_1, y_2, \ldots, y_n | \theta, \phi) = \prod_{i=1}^{n} \frac{\Gamma(y_i + \phi)}{\Gamma(\phi) \Gamma(y_i + 1)} \left( \frac{\phi}{\theta_i + \phi} \right)^\phi \left( \frac{\theta_i}{\theta_i + \phi} \right)^{y_i} \]  \hspace{1cm} (4.13)

The log-likelihood is:

\[ \ln[L(y_1, y_2, \ldots, y_n | \theta, \phi)] = \ln \prod_{i=1}^{n} \frac{\Gamma(y_i + \phi)}{\Gamma(\phi) \Gamma(y_i + 1)} \left( \frac{\phi}{\theta_i + \phi} \right)^\phi \left( \frac{\theta_i}{\theta_i + \phi} \right)^{y_i} \]

\[ = \sum_{i=1}^{n} \left[ \ln \Gamma(y_i + \phi) - \ln \Gamma(\phi) + \phi \ln(\phi) + y_i \ln(\theta_i) - (\phi + y_i) \ln(\theta_i + \phi) \right] \] \hspace{1cm} (4.14)

Substituting \( \theta_i \) by model parameters, the likelihood function becomes a function of parameters. The estimates of the parameters can be obtained by maximizing the log-likelihood \( \ln[L(y_1, \ldots, y_n | \theta, \phi)] \).

4.1.2.2 Poisson-Lognormal Model

The Poisson-Lognormal (P-LN) distribution is generally defined as a mixture of Poisson distributions with parameters \( \theta_i \) where \( \theta_i \) follows a lognormal distribution with parameters \( \mu \) and \( \sigma^2 \), \( \theta_i \sim \Lambda(\mu, \sigma^2) \) [21]. The probability mass function for P-LN is given by:

\[ P(Y_i = y_i | \theta_i; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi y_i!}} \int_{0}^{\infty} e^{-\theta_i} \theta_i^{y_i-1} \exp \left( -\frac{(ln\theta_i - \mu)^2}{2\sigma^2} \right) d\theta_i \]  \hspace{1cm} (4.15)

The marginal distribution of this model does not have a closed form as the Poisson-
Figure 4.3: Poisson-Lognormal Distribution
Gamma model and no explicit expression for the generating function is available. The Poisson component accommodates integer inputs to describe the actual number of traffic crashes at a site. The log-normal component of the distribution describes the over-dispersion in the Poisson rate parameter due to random effects and/or heterogeneity and describes how the average number of crashes varies across the population. To analyze traffic crashes, most applications introduce the error term $\varepsilon_i$ to $\theta_i$, $\varepsilon_i \sim N[0, \sigma^2]$ and $\epsilon_i = e^{\varepsilon_i} \sim \text{Lognormal}[0, \sigma^2]$.

$$g(\varepsilon_i; \sigma) = \frac{1}{\varepsilon_i \sigma \sqrt{2\pi}} e^{-\frac{(\ln \varepsilon_i)^2}{2\sigma^2}}$$ (4.16)

$$\mu_i = e^{(X^T \beta + \sigma \varepsilon_i)} = \theta_i e^{(\sigma \varepsilon_i)}$$ (4.17)

and,

$$P(Y_i = y_i | \theta_i, \varepsilon_i) = \frac{(\theta_i e^{\sigma \varepsilon_i})^y_i e^{-\theta_i e^{\sigma \varepsilon_i}}}{y_i!}$$ (4.18)

The unconditional density will be:

$$P(Y_i | x_i) = \int_{-\infty}^{\infty} \frac{(\theta_i e^{\sigma \varepsilon_i})^y_i e^{-\theta_i e^{\sigma \varepsilon_i}}}{y_i!} \phi(\varepsilon_i) d\varepsilon_i$$ (4.19)

where, $\phi(\varepsilon_i)$ denotes the standard normal density. Correspondingly, the unconditional log-likelihood function is:

$$Ln[L(y_1, y_2, ..., y_n | \theta_i, \phi)] = \sum_{i=1}^{n} \left[ \int_{-\infty}^{\infty} \frac{(\theta_i e^{(\sigma \varepsilon_i)})^y_i e^{-\theta_i e^{\sigma \varepsilon_i}}}{y_i!} \phi(\varepsilon_i) d\varepsilon_i \right]$$ (4.20)

The integrals of the log likelihood function do not exist in closed form. The Markov Chain Monte Carlo (MCMC) algorithm can be applied to estimate model
parameters. On the other hand, the quadrature based approach suggested by Butler and Moffitt [15] is a convenient method of approximating them. Related details are briefly presented below.

Given, \( v_i = \varepsilon_i / \sqrt{2} \) and \( \omega = \sigma \sqrt{2} \),

\[
P(Y_i | x_i) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} EXP(-v_i^2)P(Y_i | x_i, v_i)dv_i
\]  (4.21)

\[
E[Y_i | x_i, v_i] = EXP(X^T \beta + \omega v_i)
\]  (4.22)

The parameters can be obtained by maximizing the re-parameterized log-likelihood function:

\[
Ln[L_Q] = \sum_{i=1}^{n} Ln \left[ \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} W_h P(Y_i | x_i, V_h) \right]
\]  (4.23)

where, \( Ln[L_Q] \) is the Gauss-Hermite quadrature approximation, \( V_h \) are the nodes for the quadrature, and \( W_h \) are the weights for the quadrature.

4.1.2.3 Poisson-Inverse Gaussian (P-IG) Model

It is expected that Poisson-Inverse Gaussian (P-IG) can be used in cases where P-LN is applicable. This is conjectured because of the similarities between the log-normal distribution and the inverse Gaussian distribution. It has been proved that the P-IG distribution is more tractable than the P-LN distribution. Similar conclusion can be drawn based on the derivatives below. If \( \varepsilon_i \) is assumed to have an Inverse-Gaussian distribution, the following derivations can be obtained.

Given \( \varepsilon_i \) follows the Inverse-Gaussian distribution with mean \( \gamma = 1 \) and shape parameter \( \delta = 1/\tau \),

\[
g(\varepsilon_i; \gamma, \delta) = \left[ \frac{\delta}{2\pi\varepsilon_i^3} \right]^{1/2} e^{-\frac{\delta(\varepsilon_i - \gamma)^2}{2\gamma^2\varepsilon_i}}
\]  (4.24)
Figure 4.4: Poisson-Inverse Gaussian Distribution
Therefore, 
\[
g(\epsilon_i; 1, 1/\tau) = \left[ \frac{1}{2\pi \tau \epsilon_i^3} \right]^{1/2} \exp \left( -\frac{(\epsilon_i - 1)^2}{2\tau \epsilon_i} \right) \tag{4.25}
\]

The marginal distribution of \( Y_i \) is derived as:
\[
P(Y_i = y_i|\theta_i, \tau) = \int_0^\infty f(y_i|\theta_i, \epsilon_i)g(\epsilon_i)d\epsilon_i
\]
\[
= \int_0^\infty \frac{(\theta_i \epsilon_i)^{y_i} e^{-\theta_i \epsilon_i}}{y_i!} \left[ \frac{1}{2\pi \tau \epsilon_i^3} \right]^{1/2} \exp \left( -\frac{(\epsilon_i - 1)^2}{2\tau \epsilon_i} \right) d\epsilon_i
\]
\[
= \frac{\theta_i^{y_i} \epsilon_i^{1/\tau}}{y_i!(2\pi \tau)^{1/2}} \int_0^\infty \epsilon_i^{y_i-3/2} e^{-\theta_i \epsilon_i - \frac{\epsilon_i}{2\tau} - \frac{1}{2\tau \epsilon_i}} d\epsilon_i \tag{4.26}
\]

Detailed derivations is attached in Appendix A. Eventually we can get,
\[
P(0|\theta_i, \tau) = e^{1/\tau(1-\sqrt{1+2\tau \theta})} \tag{4.27}
\]
\[
P(1|\theta_i, \tau) = \theta_i (1 + 2\tau \theta)^{-1/2} P(0|\theta_i, \tau) \tag{4.28}
\]
\[
P(y_i|\theta_i, \tau) = \frac{2\tau \theta}{1 + 2\tau \theta} \left( 1 + \frac{3}{2y} \right) P(y_i - 1|\theta_i, \tau) + \frac{\theta^2}{1 + 2\tau \theta \ y(y - 1)} P(y_i - 2|\theta_i, \tau) \tag{4.29}
\]

The marginal probabilities can be derived recursively. The maximum likelihood functions for the P-IG distribution are derived as:
\[
L(y_1, y_2, ..., y_n|\theta_i, \tau) = \prod_{i=1}^{n} P(y_i|\theta_i, \tau) \tag{4.30}
\]

Both Inverse-Gaussian and Log-normal models provide versatile alternative specifications that are more flexible and more natural than the gamma form.
4.2 Spatially Adjusted Crash Prediction Models

In classical statistical modeling, the observed data is often assumed to be spatially independent. Consequently, the joint probability distributions are the products of marginal probability distributions, as demonstrated when deriving the likelihood functions of regression models. However, when checking the model residuals and model fitting, this assumption is occasionally violated. In reality, observations such as traffic crashes may be related to each other in some way. When spatial correlation is not explicitly modeled, serious consequences may arise because model assumptions are violated [50]. Along with the inflation of degrees of freedom, standard errors and estimated coefficients may be biased as well as inconsistent, risking erroneous conclusions on the basis of a mis-specified regression model [8, 50].

A few studies have confirmed the existence of spatial auto-correlations in crash data [1, 2, 4, 28, 29, 51, 55, 60, 61]. The spatial dependency can be a surrogate for mis-specified model forms and unknown but relevant covariates. Potential unknown factors that may significantly impact traffic crashes include land use, environmental effects (e.g., temperature, sunlight), economic factors, driver population, law enforcement, etc.

The first analyses to consider spatial component were conducted by Levine et al. [60, 61]. The authors developed a spatial lag model using time series auto-regressive lag-1 model in which the previous time was replaced by the weighted average of the neighbors. Nevertheless, the authors assumed that the crashes follow a normal distribution rather than Poisson or Negative Binomial distributions. Afterwards, Jones et al. [55] and Nicholson [73] both conducted studies to look for patterns in crash data. They both applied the K-function analysis and confirmed the existence of spatial correlations in crash data. Nicholson assessed statistical analysis techniques
for spatial data including nearest-neighbor methods, K-function, and quadrant. The nearest-neighbor approach was found to be the most powerful and robust technique in detecting the crash patterns by then.

Black and Thomas [11], and Aguero-Valverde and Jovanis [4] analyzed spatial correlations at the segment level. Black and Thomas applied the classic Moran’s I index and concluded that there was a significant level of positive spatial correlation in the data even though the results were descriptive. Aguero-Valverde and Jovanis tested different segment neighboring structures and found that high proportion of variability is explained by the spatial correlation term. The authors also emphasized that the importance of spatial correlation is to reduce the bias associated with model mis-specification as shown by the changes of the estimated coefficients. The authors proposed three levels of neighboring structures which are summarized as follows: 1) the first-order neighbors share the edge with each other, 2) the second-order neighbors connect directly to first-order neighbors, and 3) the third-order neighbors are the ones that connect to the second-order neighbors. Two strong assumptions are associated with the definitions of the neighboring structures. Firstly, the distances between roadway segments are ignored. Secondly, the spatial weights assigned to each order of neighbors are equal to the inverse of the order (i.e., 1, 1/2, and 1/3). Nonetheless, only the first-order neighboring structure was found to display significant contribution to the spatial models.

Similarly, El-Basyouny and Sayed [29] investigated the spatial effects in crash prediction models on urban arterial roads. The spatial effects were separately modeled using the Gaussian conditional auto-regressive (CAR) model, the multiple membership (MM) model and an extended MM (EMM) model. The EMM model was found to fit the data the best, followed by the CAR model. The authors found significant correlation between the spatial effects and heterogeneity and spatial effects were
considerably alleviated by clustering segments within the same corridor. Following Aguero-Valverde and Jovanis’s work discussed in the paragraph above, El-Basyouny and Sayed only considered the first-order neighboring structure in their study. The adjacency-based proximity was used to create the spatial weight matrix.

Miaou et al.[70], MacNab [65], Aguero-Valverde and Jovanis [3, 4], El-Basyouny and Sayed [29], Miaou and Song [69] all developed spatial Bayesian models in which the spatial correlation was modeled by the CAR model. Significant spatial correlation was found in those studies. The main concern about the spatial Bayesian models in these studies is that adjacency-based spatial weights are often used, and as pointed out by Miaou and Song [69], the spatial effect indicator ($\rho$) needs to be plugged in beforehand into the CAR models and is usually set up as a high value. Thereby, the interpretation of the parameter obtained from model output cannot represent the actual spatial effects of the model. Nonetheless, by accounting for the spatial random effects, the spatial models were found to perform relatively better than traditional model techniques.

Wang et al. [1, 103, 104, 107] performed spatial and temporal analysis using the generalized estimating equations (GEE) approach. The authors grouped intersections into clusters based on spatial location, distances and corridor locations. When analyzing the spatial correlations, intersections within the same corridor were considered correlated whereas intersections from different corridors were independent. Apparently the correlations between intersections within the corridors that intersect each other was ignored. Three correlation structures for intersections within corridors were analyzed and compared including: 1) independent correlation, 2) exchangeable correlation where a constant was assigned to the pairs of intersections within the corridor, and 3) AR-1 correlation, where the correlation decreases as the distance between the pair of intersections increase. The authors also tried the unstructured correlation
where different correlations were estimated for each pair of intersections within a corridor but the spatial model failed to converge. Significant spatial correlations were found between intersections along corridors.

Shortly, previous research work revealed the practical consequences of not accounting for the spatial correlations within the data. The main deficiency would be misleading safety conditions of the sites of interest. Griffith and Haining [40] argued that spatial independence of count data does not hold true and anticipated that there will be inter-dependencies in the model residuals. Current practice focuses on the applications of CAR or SAR models [4]. A relatively new concept that has never been used to model spatial auto-correlation in traffic safety field is the eigenvector based spatial filtering. Previous research had found that the filtered model performed moderately better than a SAR model with the same spatial weight matrix because the Moran’s I for residuals in the filtered model was just half the size of that for the SAR model [38, 39]. The presence of spatial correlations is violating the elemental assumption of independently and identically distributed errors of most statistical procedures [7]. To receive unbiased estimates and inferences, spatial correlations must be explicitly modeled [99]. Next sections of this chapter will introduce the eigenvector spatial filtering algorithm that accounts for spatial effects within count data.

4.2.1 Eigenvector Spatial Filtering (ESF)

The spatial eigenvector filtering approach is based on the idea that the spatial arrangement of traffic crashes can be translated into explanatory variables that capture the spatial effects at different spatial resolutions. The spatial filtering concept was first implemented by Griffith in 1978 [37]. Depending on the methods that are used to generate these synthetic variables or filters, the spatial filtering technique can be divided into three different categories: 1) Dray et al.’s distance-based eigenvector-
tor procedure [27], 2) Getis’ G-statistics-based approach [32, 33], and 3) Griffith’s eigenfunction-based procedure [38]. Some parts of these three approaches are very similar and the basic idea is the same, which is to extract eigenvectors from the transformed weight matrix. These procedures are derived based on the finding discovered by de Jong et al. in 1984 [56], that there exists a linkage between eigenvalues and Moran’s I [5] as shown in (4.31) where C is a constant. Roughly speaking, the larger a positive eigenvalue is, the stronger the positive spatial auto-correlation the associated eigenvector represents.

\[
\text{Extreme Moran’s I values} = C \times \text{Extreme eigenvalues of } \Omega \quad (4.31)
\]

Dray et al. proposed a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). This approach is used to create spatial predictors that can be easily incorporated into regression models. Dray et al. proposed more formal mathematics and the linkage between the PCNM method and spatial auto-correlation structure. The method basically consists of diagonalizing a spatial weighting matrix, then extracting the eigenvectors that maximize the Moran’s index of auto-correlation. These eigenvectors can then be used directly as explanatory variables in regression models. This paper extended the original PCNM approach by introducing both positive and negative eigenvalues into the analysis.

Getis’s work used the differences between observed and expected local spatial statistics to separate spatial effects from non-spatial effects. Each variable is therefore separated into two parts with only one part related to the spatial effect. Griffith used orthogonal and uncorrelated map patterns represented as eigenvectors obtained from a doubly centered contiguity matrix to filter out meaningful spatial forces. Getis and Griffith compared their work in 2002 [34]. It was found that two filtering methods yielded similar goodness-of-fit statistics. The results were also similar to those from
the SAR model. They found that Griffith’s model is preferable due to its flexibility for application in non-linear model specifications. Getis’ approach requires that analysts have variables with a natural origin and a linear model specification, thereby limiting its use [78].

The ESF method is a relatively new technique in analyzing spatial data, and previous research proved that it appears to offer much promise. Currently there are applications in fields like ecology [26, 30, 54], economic analysis [98], optical imaging [108] and spatiotemporal crime mapping and analysis [50]. These studies confirmed that generalized linear models linked with ESF can effectively address spatial auto-correlation.

A recent application of spatial filtering in analyzing land use data sets is presented by Wang et al. in 2011 [105]. They pointed out that the application of the spatial filtering idea in transportation is rare but is promising. They compared the applications of spatial filtering with spatial auto-regressive models with distance decay parameters estimated using Bayesian techniques. They found that the ESF approach demonstrates great potential as a worthy competitor to more conventional Bayesian spatial-auto-regressive models, offering high fit statistics, somewhat shorter computing times, and more straightforward computations. The authors concluded that the spatial auto-correlation arises due to imperfect information on observational units and measurement errors [105]. The missing information results in correlations across error terms of nearby sites. Aggregated spatial data values such as AADT can introduce forms of spatial auto-correlation.

A traditional treatment for the spatial effect is to directly specify a spatial structure, such as a spatial auto-regressive (SAR), conditional auto-regressive (CAR) or spatial moving average (SMA) models [5, 9, 58]. Most applications to date rely on specific functional forms, such as SAR, CAR, and SMA, and arbitrarily pre-determined
weight structures to anticipate spatial structures in the data. When incorporating SAR or CAR models in regression analysis, the likelihood functions do not have a closed form. MCMC approach must be used to the model estimation. This is one of the reasons why the spatial analyses are mostly achieved using Bayesian sampling inferences.

Traditional SAR and CAR models assume that the responses are more likely to be influenced by nearby neighbors. An $N \times N$ weight matrix is used to describe the distance-decay spatial pattern. ESF addresses the spatial auto-correlation from a quasi semi-parametric point of view [38]. In other words, the spatial arrangement of data points can be translated into explanatory variables, which can be used to capture spatial effects at different spatial resolutions (e.g., global, regional and local) [30]. Apart from the observed explanatory variables, ESF generates synthetic explanatory variables to represent the spatial structures within the data set. Spatially adjusted regression models are consequently generated. By including synthetic explanatory variables, the spatially adjusted models display more flexibility compared to the original regression models.

4.2.1.1 Introduction of ESF

Griffith found that with the increase of sample size $n$, greater agreement was observed between the eigenvalues of the weight matrix $W$ and the eigenvalues of its transformation matrix $\Omega$.

$$\Omega = \left( I - \frac{11^T}{n} \right) W \left( I - \frac{11^T}{n} \right)$$

(4.32)

where, $I$ is a $N \times N$ identity matrix, $1$ is a $N \times 1$ vector of $1$s and $T$ denotes the matrix transpose.

The eigen function decomposition of matrix $\Omega$ generates $n$ eigenvectors, denoted as
\[ E = (E_1, E_2, \ldots, E_n) \text{ and } n \text{ eigenvalues } \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n). \] Matrix \( \Omega \) is guaranteed to have an eigenvector of purely \( 1/\sqrt{n} \) values, corresponding to an eigenvalue of unity. The advantage of using \( \Omega \) instead of \( W \) is the fact that the eigenvectors of \( \Omega \) are orthogonal and thus uncorrelated. The Moran’s I coefficient in (3.1) can be rewritten as (4.33).

\[ I(X) = \frac{n}{1^T W 1} \frac{x^T \left( I - \frac{11^T}{n} \right) W \left( I - \frac{11^T}{n} \right) x}{x^T \left( I - \frac{11^T}{n} \right)} \] (4.33)

Griffith had proved that the extreme values of Moran’s I for a specific spatial configuration can be expressed as a function of matrix \( \Omega \)’s eigenvalues. The corresponding Moran Coefficient for the corresponding eigenvectors are,

\[ M_j = \frac{n}{1^T W 1} \lambda_j \] (4.34)

Extreme Moran’s I values = \[ \frac{n}{1^T W 1} \times \text{ Extreme eigenvalues of } \Omega \] (4.35)

The eigenvalues \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n) \) are related to Moran’s I which represents the spatial auto-correlation. Similarly, they can be either positive or negative. The first eigenvector of matrix \( \Omega \) is the vector of values yielding the strongest spatial auto-correlation, thus having the largest Moran’s I value. It is thus the most important principal component of the spatial structure. The second eigenvector offers the second largest eigenvalue and is orthogonal to the first eigenvector. Eigenvectors associated with high positive eigenvalues have high positive auto-correlation and describe global structures. Similarly, eigenvectors associated with high negative eigenvalues have high negative auto-correlation and describe local structures. Eigenvectors associated with eigenvalues with extremely small absolute values correspond to spatial auto-correlation with low intensity and are not suitable for defining spatial structures.
4.2.1.2 Define Spatial Weight Matrix $W$

The choice of the spatial weight matrix $W$ is the critical step in ESF analysis. Although the definition of the neighborhood matrix is exogenous, it requires that the actual spatial process and relations are mimicked [50]. The determination of the matrix $W$ and the subsequent coding scheme are fundamental for ESF and will influence the actual filtering as well as model goodness-of-fit.

Two types of spatial proximity are commonly used: distance-based measures and adjacency-based (aka neighboring-based) measures. Distance-based weight is calculated based on either inverse distance weighting function or exponential distance-decay weighting function. Typical adjacency-based weighting has binary codes $w_{ij} = 0$ or $1$ depending on whether sites $i$ and $j$ are neighbors. The definition of “neighbors” also depends on the nature of study and can be customized for different data sets. Frequently applied approach is the first-order queen contiguity which assigns the weight $1$ to adjacent pair of sites that share an edge and/or node. Other sites have spatial weight of $0$ [57].

In most cases, the choice of a particular matrix may be rather difficult and a data-driven specification could then be applied. The objective is to select a configuration of $W$ that best represents the spatial structures of focused areas and objectives. The generation of the matrix $W$, which led to the matrix $\Omega$, is the key to the optimal performance of the spatially adjusted models. In Chapter 5 and 6, specific weight matrices for intersections and segments along urban corridors are developed. Moreover, the standardization of weight matrices can be achieved using different coding schemes as summarized in Table 4.1 [79].
Table 4.1: General Coding Schemes for Weight Matrix Standardization

<table>
<thead>
<tr>
<th>Coding</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>B: binary</td>
<td>$w_{ij} = 1$ if location $i$ and $j$ are neighbors; $w_{ij} = 0$ otherwise.</td>
</tr>
<tr>
<td>C: global</td>
<td>Weights of all links sum to $n$.</td>
</tr>
<tr>
<td>W: row</td>
<td>Weights of rows sum to 1.</td>
</tr>
<tr>
<td>S: variance</td>
<td>Stabilize the variance by compensating the level of variation within weights.</td>
</tr>
</tbody>
</table>

4.2.1.3 Extraction of Eigenvectors

The other crucial component in implementing ESF in crash prediction modeling is the extraction of eigenvectors. The extraction will result in $N$ eigenvectors whose elements are attached to each site. The methodology requires judicious selection of eigenvectors. Using the complete set of eigenvectors is not feasible due to missing degrees of freedoms [78]. This requires a pre-selection to uncover potential eigenvector candidates. The selected eigenvectors are the ones that best reduce spatial auto-correlation in the residuals. The selected eigenvectors are included in the crash prediction models as the spatial predictors. Tiefelsdorf and Griffith proposed a criterion of 0.25 to pre-select candidate eigenvectors [98].

$$\frac{\lambda_i}{\lambda_{max}} > 0.25$$ (4.36)

where, $\lambda_i$ represents the extracted eigenvalues and $\lambda_{max}$ is the largest positive eigenvalue.

Griffith recommended using the step-wise regression to select the eigenvectors. It is essential to select the top performing set of eigenvectors in order to optimally
reduce the residual sum of squares calculated as,

\[ \text{RSS} = \hat{\varepsilon}^* T \hat{\varepsilon}^* \] (4.37)

or increase the proportion of explained variation represented by \( R^2 \),

\[ R^2 = 1 - \frac{\hat{\varepsilon}^* T \hat{\varepsilon}^*}{y^T My} \] (4.38)

where, \( M = \left( I - \frac{1}{n} X^T X \right) \) is an idempotent matrix, and \( \hat{\varepsilon}^* \) are the estimates of the error terms.

To ensure the spatial pattern underlying each selected eigenvector merits inclusion in the regression equation, Tiefelsdorf and Griffith (2007) proposed minimizing the following Z-score objective function to remove spatial auto-correlations in the error terms [98]:

\[
\min \ z[I(\hat{\varepsilon}^*)] \equiv \sqrt{\frac{y^T M_{(X|E)} W M_{(X|E)} y}{y^T M_{(X|E)} y} - E(I(\hat{\varepsilon}^*))} \frac{V ar (I(\hat{\varepsilon}^*))}{n - k} \] (4.39)

where,

\[ E(I(\hat{\varepsilon}^*)) = \frac{tr(M_{(X|E)} W M_{(X|E)})}{n - k} \] (4.40)

\[ V ar (I(\hat{\varepsilon}^*)) = \frac{2 \left[ (n - p) tr\left(M_{(X|E)} W M_{(X|E)}\right)^2 - \left(tr\left(M_{(X|E)} W M_{(X|E)}\right)\right)^2\right]}{(n - k)^2 (n - k + 2)} \] (4.41)

\[ M_{(X|E)} \equiv I - X(X^T X)^{-1} X^T \] (4.42)

Eigenvectors are added to the unadjusted model until the Moran’s I index in model residuals is no longer significant. Each eigenvector represents a particular spatial pattern. The spatial auto-correlation is effectively allowed to vary in space which relaxes the assumption of both spatial isotropic and stationary [30]. Either adjacency- or distance-based connectivity matrices can be decomposed, offering a
great deal of flexibility regarding topology and transformations.

4.2.2 Spatially Adjusted Crash Prediction Models using ESF

Incorporating spatial variation in the crash prediction models requires tools to explicitly describe spatial relationships as predictors or covariables. The consideration of spatial filters increases the explanatory power of the regression. Spatially adjusted crash prediction models will support the understanding of underlying spatial processes affecting the presence or absence of crashes. For the typical Poisson-Gamma model, the following expressions are derived.

\[ Y_i | \mu_i \sim \text{Poisson}(\mu_i), \ i = 1, 2, ..., n \]  

\[ \mu_i = e^{(X^T \beta X + \epsilon_i + \phi_i)} = \theta_i \epsilon_i e^{\phi_i} \]  

\[ \phi_i = E_k \beta_E \]  

\[ P(Y_i = y_i | \theta_i, \epsilon_i, \phi_i) = \frac{(\theta_i \epsilon_i e^{\phi_i})^y e^{-\theta_i \epsilon_i e^{\phi_i}}}{y_i!} \]

where, \( X \) is a 1 by \( p + 1 \) vector of \( p \) covariates and a 1 for the intercept term, \( \beta_E \) is the \( p + 1 \) by 1 vector of covariate regression coefficients including the intercept term, \( \beta_E \) is a \( k \) by 1 vector of selected eigenvector regression coefficients with \( E_k \beta_E \) denotes the spatially structured random effects term. \( \phi_i \) is the spatial random effect, one for each observation. The maximum likelihood method can be used to obtain an estimate of model parameters.
4.2.3 Spatially Adjusted Crash Prediction Models using CAR

Besides ESF, other spatial autocorrelation term can be added to the crash prediction models. Therefore,

\[ \mu_i = e^{(X^T \beta + \varepsilon_i + \psi_i)} \]  

(4.47)

The assumptions on the uncorrelated term \( \varepsilon_i \) remain the same as the Poisson mixture model. The term \( \psi_i \) denotes the spatial random effect, one for each observation. Two common ways to express the spatial component proposed in several publications are the CAR and the SAR model. In practice, the CAR and SAR models were found to produce similar results.

**CAR Model**:

\[ E(y_i|y_{j \neq i}) = \mu_i + \rho \sum_{ij} [w_{ij}(y_i - \mu_j)] \]

**SAR Model**:

\[ E(y_i|y_{j \neq i}) = \mu_i + \rho \sum_{ij} [w_{ij}y_j] \]

where, \( \mu_i \) is the expected value for observation \( i \), \( w_{ij} \) is the spatial weight between observation \( i \) and \( j \), and \( \rho \) is the spatial auto-correlation parameter.

The CAR adjusted Poisson-Gamma model cannot be estimated using the maximum likelihood method. They must be estimated using the MCMC technique. The spatial effect is assumed to have the distribution as:

\[ P(\psi_i|\psi_{-i}) \propto e^{\left( -\frac{w_{ii}}{2\sigma^2} \left( \psi_i - \rho \sum_{i \neq j} \frac{w_{ij}}{w_{ii}} \psi_j \right)^2 \right)} \]  

(4.48)

\[ w_{ii} = \sum_{i \neq j} w_{ij} \]  

(4.49)

Using the Poisson-Gamma model as an example, the adjusted crash prediction model with CAR spatial terms can be written as:
\[ Y_i|\theta_i \sim \text{Poisson}(\theta_i), \ i = 1, 2, \ldots, n \] (4.50)

\[ \theta_i = e^{(X^T \beta X)} \] (4.51)

\[ \mu_i = e^{(X^T \beta X + \epsilon_i + \psi_i)} \] (4.52)

\[ \epsilon_i = e^{\epsilon_i} \sim \text{Gamma}(\phi, \phi) \] (4.53)

\[ \beta_j \sim \text{Uniform}(-\infty, \infty) \] (4.54)

\[ \phi \sim \text{Gamma}(a_\phi, b_\phi) \] (4.55)

\[ \tau_\psi = \sigma_\psi^{-2} \sim \text{Gamma}(a_\psi, b_\psi) \] (4.56)

\[ \rho \sim \text{Uniform}(\lambda^{-1}_{\text{min}}, \lambda^{-1}_{\text{max}}) \] (4.57)

### 4.2.4 Section Summary

This section explores the spatially adjusted crash prediction models and introduces the eigenvector filtering technique to account for such randomness. The most popular interpretation for having the spatial correlations in traffic crash data is the omitted covariables which vary in space. Ignoring such effects may cause underestimation of model variability, result in biases parameter estimates and decrease model’s prediction performance as demonstrated in Chapter 5 and 6. After obtaining the estimates of parameters of a model, the dependency of model residuals and fitted values should be checked using a spatial index such as the Moran coefficient. If noticeable spatial auto-correlation is present in the residuals of an unfiltered model, there may be unknown variables or mis-specified model forms that have impacts on the results. The eigenvector spatial filtering can be used to account for this correlation and filter out the correlations in the model to achieve better fitted models.


4.3 Model Development and Over Fitting Issues

Reasonable model forms should be identified for different facility data sets. Traffic volume is often found to be significantly related to crashes. However, there is no fixed formula for measuring the impact of traffic exposures. At intersections, more conflicts between traffic from major and minor approaches imply higher opportunities for crash to occur. It is necessary to identify major and minor approaches at each intersection. Various model forms are tested in Chapter 5 and 6 to find the best model forms for intersection data sets and segment data.

When too many variables are included in crash prediction models, over fitting may occur especially when large sample sizes are used. Even though the variables are found to be statistically significant, the inclusion of such variables may not be of practical importance. This may be because there are underlying correlations between the variables. One has simply modeled the “noise” in the model once the most relevant explanatory variables are included in the model. One way to deal with over fitting is using cross-validation. That is, the data set used to develop the regression models is divided into two parts with one part used for estimating the model parameters and the other part for model validation. For instance, the data set is divided into two groups with 70% of the data used for model development and the other 30% for model validation.

The other effective way to check model over fitting is to use relative goodness-of-fit measures such as $R^2$, Pearson chi-square, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), etc. AIC and BIC measures deal with the trade-off between the model goodness-of-fit and the complexity of the model. They are statistics typically used as model selection criteria rather than model goodness-of-fit assessment. Model details in regards to model goodness-of-fit measures are
discussed in Section 4.5.

4.4 Bayesian Paradigm

As discussed previously, it is crucial to obtain a reliable estimate of expected safety condition at a site. Because of the randomness and scarcity of crashes, the naive statistical analysis relying on the information from small number of sites fail to capture the true long term behavior at that site. Besides, the true expected number of crashes at a particular site is unobservable. Using observed crashes or crash rate to estimate the true expected number of crashes will suffer the regression-to-the-mean bias which is referred as the phenomenon that when a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement, and paradoxically, if it is extreme on its second measurement, it will tend to have been closer to the average on its first [12].

In road safety analysis, crashes are random and rare events that naturally fluctuate over time at any given site. The fluctuation is measured by variance and the variations are usually due to the normal randomness of crash occurrence. Because of this random fluctuation, an extremely high number of crashes chosen in one period is very likely to experience a lower number in the next period and vice versa. The specific concern in road safety is that entities are commonly selected for improvements or treatments because of their high crash records in only one year or a short period of time. The counts will naturally regress back toward the mean in subsequent years.

There comes another question. Which one should we trust: the average number of crashes in the long run, or the temporary extreme value? The answer should be clear. Even if high crashes were temporarily observed, the best guess about the magnitude of the crash frequency in the next period should still be the expected average number. This is the essence and meaning of expected value. When the first observed number of
crashes is a down-fluctuation or an up-fluctuation, a return towards the mean should be expected for the next observation.

The first and most influential step towards the application of the Empirical Bayes method in analyzing traffic crash data was the work by Hauer and colleagues [42, 44, 45, 46, 47, 49, 52, 53, 81, 83, 84, 86]. Afterwards, several researchers proposed the application of the full Bayesian (FB) method to estimate the expected number of crashes [22, 23, 24, 25, 29, 51, 64, 80, 87, 92, 93, 102]. As a special case of FB, the EB method assumes that the effect of explanatory variables such as traffic volumes and segment lengths on the site’s safety is represented via crash prediction models and the associated crash modification factors (CMFs). The EB method is recommended by the HSM and is now being implemented by most states in the U.S. Often times, the EB results are very different to FB results.

According to the HSM, application of the Empirical Bayes (EB) method provides a way to combine the estimate using a crash predictive model and observed crash frequencies to obtain a more reliable estimate of expected average crash frequency. The EB method is also widely applied in identifying crash-prone locations. The equation to obtain the EB estimate is defined using equation (4.44).

\[
\hat{\mu}_i = w \times E(Y_i) + (1 - w) \times y_i \tag{4.58}
\]

where, \( \hat{\mu}_i \) is the EB estimate of the expected number of crashes obtained based on the Best Linear Predictor (BLP); \( E(Y_i) \) is the predicted number of crashes estimated by the crash prediction models; \( w \) is a weight factor which is a function of mean and variance of \( Y_i \).

\[
w = \frac{1}{1 + \frac{\text{Var}(Y_i)}{E(Y_i)}} \tag{4.59}
\]

Equation (4.44) indicates that the estimated number of crashes at site \( i \) is calcu-
lated as the weighted average of the observed number of crashes collected at that site and the mean number of crashes at similar sites. For example, if the Poisson-Gamma model is used to develop the crash prediction models, the weight $w$ is calculated as $w = 1/(1+1/\phi)$ where $\phi$ is the inverse dispersion parameter from the Poisson-Gamma model. Thereby, when the dispersion of the model decreases, the inverse dispersion parameter increases and $w$ increases. This is intuitively appealing, as the developed model is more reliable, more weight is put into the model prediction.

4.5 Goodness-of-Fit Measurements

Three categories of goodness-of-fit measures can be used in combination to check how well the developed crash prediction model fits the focused data, including: 1) likelihood statistics; 2) model error estimates; and 3) over-dispersion tests.

Likelihood statistics

Likelihood statistics include the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Deviance Information Criterion (DIC). They are normally used in selecting models. These measures penalize models with more estimated parameters than needed, and help reduce the possibility of over fitting. Smaller values indicate better model fit.

The AIC statistic has the form of $AIC = -2ln(L) + 2p$, where $p$ is the number of unknown parameters, $ln(L)$ is the log-likelihood function. BIC statistic is also known as the Schwarz Information Criterion. It has the form of $BIC = -2ln(L) + pln(n)$, where $n$ is the sample size. The DIC statistic is often used as the goodness-of-fit measure when Bayesian estimation method is used. AIC and BIC are often used for the maximum likelihood method. DIC has the form of $DIC = \bar{D} + p_D$, where, $p_D = \bar{D} - D(\theta)$, $\bar{D} = E^\theta[D(\theta)]$ is the expectation to measure how well the model fits
the data, and \( D(\theta) = -2ln(p(y|\theta)) + C \) represents the deviance.

The Pearson Chi-Square is another likelihood statistic for measuring model goodness-of-fit. When the mean and variance are properly specified, \( E[\text{Pearson} - \chi^2] = n \) [17]. Hence, values closer to \( n \) show a better fit, where,

\[
\text{Pearson} - \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\theta}_i)^2}{\text{var}(y_i)}
\]  

(4.60)

**Model Error Estimates**

Two statistics can be used to estimate errors in the model:

1) **Mean Absolute Deviation (MAD)**

MAD was proposed by Oh et al. in 2003 [77]. It is the sum of the absolute differences between the estimated (or fitted) and observed values. Higher values indicate the estimated values are more spread out from the observed values. A value close to 0 indicates that the model predicts the observed data well.

\[
MAD = \frac{1}{n} \sum_{i=1}^{n} |\hat{\theta}_i - y_i|
\]  

(4.61)

where, \( n \) is the validation data sample size.

2) **Mean Squared Prediction Error (MSPE)**

Similarly, the MSPE is the sum of the squared differences between the estimated and observed crashes divided by validation sample size. A value closer to 1 indicates better fit.

\[
\text{MSPE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - y_i)^2
\]  

(4.62)
Over-dispersion Tests

Adjusted deviance or adjusted Pearson Chi-Square can be used to test the over-dispersion in the fitted models. The adjusted deviance is defined as the deviance divided by the degrees of freedom (N-K-1). A value closer to 1 indicates a good fit. Usually, values greater than 1 indicate over-dispersion, while values below 1 indicate under-dispersion. The adjusted Pearson Chi-Square is defined as Pearson Chi-Square divided by the degrees of freedom. A value closer to 1 indicates a satisfactory fit. The inverse dispersion parameter can also be used to check model fit. The larger the inverse dispersion parameter value, the better the model performance and the smaller the unexplained and isolated site-specific heterogeneity.

4.6 Chapter Summary

This chapter discussed the methodologies for modeling traffic crash data. The nature of crash causes high variability and the empirical reality of over-dispersion. Statistical models and the underlying assumptions must be used properly in order to model the variation and withstand logical scrutiny since different assumptions will draw different results and may fit different types of data sets.

Spatial effect is another possible source of variability and may have significant impact on the overall model goodness-of-fit. Spatially adjusted regression models can be built within a Poisson mixture model. Chapter 5 and 6 will present the development of crash prediction models using real world data and detect the existence of spatial correlations in model residuals.
Chapter 5  Crash Prediction Models for Intersections along Corridors

This chapter explores the crash prediction models for intersections along corridors. Major data collection involving site selection and data acquisition was conducted towards further analysis. Model development involves checking model assumptions, selecting model forms and analyzing the spatial effects in the model. All the models presented in this research were fitted using R software and the codes are attached in Appendix B.

5.1  Intersection Data Preparation

Intersection data used in this chapter was collected in Northern Nevada’s urban and suburban areas covering the City of Reno, City of Sparks, and Carson City. According to Nevada Traffic Crash Report, the majority of crashes were recorded at four-way and T-intersections. Therefore, typical urban intersections are considered, including 3-way signalized and stop controlled intersections and 4-way signalized and stop controlled intersections. “Y” intersections, multi-leg intersections and roundabout are beyond the scope of this study. Sites to be selected must have the following characteristics: 1) at least 5-year crash records, and 2) availability of AADT from either past studies or agency counting stations.

Basic roadway geometry data was collected from Nevada’s Highway Performance
Monitoring System (HPMS) database and Linear Referencing System (LRS) of road network. Five years (2007-2011) of crash data was collected from the Nevada Accident and Citation Tracking System. In the five year period, a total of 32,745 traffic crashes (with 144 fatal crashes, 1087 A injury crashes and 3389 B level injury crashes) were recorded in Reno-Sparks-Carson City area.

AADT data was extracted from Nevada’s Traffic Records Information Access (TRINA). Most principal and minor arterials have AADT measured. AADT estimate models are readily available in Nevada for some of the minor arterials where AADT data is absent. The major and minor AADTs are required as independent variables in developing the crash prediction models for intersections. The major entering volume is identified as the sum of the two highest approach volumes and the minor entering volume should be the sum of the two lowest approach volumes for a typical four-way intersection. For T-intersections, the major entering volume is calculated as the sum of the two highest approach volumes. A buffer with radius of 200 feet was created for each intersection so that crashes occurring within this threshold of an intersection was deemed to be intersection-related crashes. Original data was stored and processed in ArcGIS software.

5.1.1 Site Selection

The primary purpose of site selection is to provide sufficient study objects for the safety performance analysis and developing crash prediction models. The selection process is constitutive of two phases, each of which has specific screening criterion. The selection of study intersections starts from predefined urban corridors, which are associated with concentrated representative locations for the safety performance assessment.
PHASE I–Identify Intersection Crashes in Northern Nevada

Crash data from 2007 to 2011 in the Reno-Sparks-Carson City areas in Nevada were extracted from NDOT database. After the first screening, 9846 crash data entries (with 70% recorded at signalized intersections) were recorded in the 5-year period. Figure 5.1 summarizes the percentages of crashes for intersections with different control type and geometry.

Data assessment was conducted to ensure that all the crash data were assigned to the correct intersections. When the crash location could not be identified in the original shape file, further research was conducted to confirm the specific location in order to obtain all valuable crash information. Later, a circular buffer with radius of 200 feet was created for each intersection in ArcGIS. Any crash occurring within the buffer was considered as intersection-related crashes and was counted automatically by spatial join. At the end of this phase, a total of 487 intersections with detailed crash records were selected to establish the candidate study sites database.
Table 5.1: Breakdown of Intersection Identification

<table>
<thead>
<tr>
<th>Intersections Type</th>
<th>Number of Selected Intersections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop controlled T-intersection (3ST)</td>
<td>152</td>
</tr>
<tr>
<td>Signalized T-intersection (3SG)</td>
<td>55</td>
</tr>
<tr>
<td>Stop controlled four-way intersection (4ST)</td>
<td>120</td>
</tr>
<tr>
<td>Signalized four-way intersection (4SG)</td>
<td>160</td>
</tr>
<tr>
<td>Total</td>
<td>487</td>
</tr>
</tbody>
</table>

PHASE II–Select Intersections based on Data Availability

The recommended data elements for urban and suburban intersections are presented in Table 5.2. Most of the required data elements were provided by NDOT. The key data elements such as the AADT at minor road was estimated based on similar locations. Other missing data items such as presence of lighting were collected using Google Maps and Google Earth. One of the key data elements not presented in the HSM is the spatial location (i.e., latitude and longitude) of each intersection. The intersection coordinates were extracted from GIS files. 487 candidate sites were screened again to select sites with the most available data. Eventually 262 intersections were selected as the final study sites as shown in Table 5.3.

5.1.2 Data Collection and Processing

The purpose of data collection was to retrieve sufficient intersection information in order to develop crash prediction models. Over the past decades, many highway departments have been equipped with improved computing and data processing capabilities to link traffic crashes, roadway inventory and other related databases using GIS and related technologies. This made the data collection process relatively smooth and improved the data quality. Two phases were involved in the data collection process. The first phase involved collecting the basic characteristics of study sites. The
Table 5.2: Intersection Data Requirements in the HSM

<table>
<thead>
<tr>
<th>Roadway/Intersection Types</th>
<th>Data Element</th>
<th>Required</th>
<th>Desired</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Intersections</td>
<td>Number of intersection legs</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type of traffic control</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AADT for major road</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AADT for minor road</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of approaches with left-turn lanes</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of approaches with right-turn lanes</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presence of lighting</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Latitude and Longitude</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Signalized Intersections</td>
<td>Presence of left-turn phasing</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type of left-turn phasing</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Use of RTOR signal operation</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Use of red light cameras</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pedestrian volumes</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum number of lanes crossed by pedestrian</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presence of bus stops within 1000 ft</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presence of schools within 1000 ft</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presence of alcohol sales within 1000 ft</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: Selected Intersections in the Study Region

<table>
<thead>
<tr>
<th>Intersection Type</th>
<th>Number of Selected Intersections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop controlled T-intersection (3ST)</td>
<td>60</td>
</tr>
<tr>
<td>Signalized T-intersection (3SG)</td>
<td>20</td>
</tr>
<tr>
<td>Stop controlled four-way intersection (4ST)</td>
<td>50</td>
</tr>
<tr>
<td>Signalized four-way intersection (4SG)</td>
<td>132</td>
</tr>
<tr>
<td>Total</td>
<td>262</td>
</tr>
</tbody>
</table>
second phase consisted of vehicular volume counting and collection of other key data elements. Consequently, a database involving all available data was constructed to accommodate information for model development.

5.1.2.1 PHASE I–Intersection Information

Basic intersection information such as location and geometry layout was identified. To be specific, the location of each study site was identified using county, city, name of major streets and name of minor streets. The geometry layout of the intersections was identified through Google Maps. Other substantial intersection information includes:

- Intersection Control Devices, such as signalized, Two-way Stop-controlled (TWSC), or All-way Stop-controlled (AWSC);

- Number of Vehicle Travel Lanes on Major Street, such as 2-lane in each direction or 4-lane on both direction. The number of travel lanes on either two approaches or each approach needed to be specified. Also, the number of through, left-turn, and right-turn lanes in all directions were recorded;

- Land Use Pattern Around, which included residential or commercial land use, as well as whether there is a bus stop nearby etc.;

- Site Geometry Characteristics, such as angled T-intersection, lighting condition, etc.

- For some intersections without minor volume information, a proper interpolation between post miles was performed.

5.1.2.2 PHASE II–Data Assembling

A total of five data sets were assembled and employed in this study: four-way signalized intersections (4SG) data set, four-way stop controlled intersections (4ST) data
set, signalized T-intersections (3SG) data set, stop controlled T-intersections (3ST) data set, and urban and suburban roadway segment data set. The segment data set is introduced in Chapter 6. A summary of the key intersection related data is presented in Table 5.4. Take the four-way signalized intersection data as an example, the data set includes 130 signalized intersections, which have a total of 6,087 reported crashes for the 5-year period, with approximately 25% injury crashes and 75% non-injury crashes. From Table 5.4, it can be seen that individual intersections experienced crashes from 0 to 47 per year. The traffic volumes on major and minor approaches vary widely for different intersections.

Table 5.4: Summary Statistics of Intersection Data

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Description</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>STDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>4SG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{5,i}$</td>
<td>Annual crash counts</td>
<td>0</td>
<td>169</td>
<td>47</td>
<td>3.1732</td>
</tr>
<tr>
<td>$T_{1,i}$</td>
<td>AADT on major approaches</td>
<td>4620</td>
<td>43000</td>
<td>22564</td>
<td>701.4187</td>
</tr>
<tr>
<td>$T_{2,i}$</td>
<td>AADT on minor approaches</td>
<td>200</td>
<td>24400</td>
<td>7714</td>
<td>476.4139</td>
</tr>
<tr>
<td>$FR_i$</td>
<td>Flow ratio ($T_{2,i}/T_{1,i}$)</td>
<td>0.0096</td>
<td>1</td>
<td>0.4663</td>
<td>0.0224</td>
</tr>
<tr>
<td>4ST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{5,i}$</td>
<td>Annual crash counts</td>
<td>0</td>
<td>63</td>
<td>8</td>
<td>1.4408</td>
</tr>
<tr>
<td>$T_{1,i}$</td>
<td>AADT on major approaches</td>
<td>1170</td>
<td>63600</td>
<td>15561</td>
<td>1860.1009</td>
</tr>
<tr>
<td>$T_{2,i}$</td>
<td>AADT on minor approaches</td>
<td>96</td>
<td>13700</td>
<td>2678</td>
<td>429.1008</td>
</tr>
<tr>
<td>$FR_i$</td>
<td>Flow ratio ($T_{2,i}/T_{1,i}$)</td>
<td>0.0076</td>
<td>1</td>
<td>0.2041</td>
<td>0.0295</td>
</tr>
<tr>
<td>3SG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{5,i}$</td>
<td>Annual crash counts</td>
<td>6</td>
<td>78</td>
<td>28</td>
<td>4.8631</td>
</tr>
<tr>
<td>$T_{1,i}$</td>
<td>AADT on major approaches</td>
<td>7460</td>
<td>43000</td>
<td>18746</td>
<td>2208.3364</td>
</tr>
<tr>
<td>$T_{2,i}$</td>
<td>AADT on minor approaches</td>
<td>1660</td>
<td>20000</td>
<td>6436</td>
<td>1136.2907</td>
</tr>
<tr>
<td>$FR_i$</td>
<td>Flow ratio ($T_{2,i}/T_{1,i}$)</td>
<td>0.1186</td>
<td>1</td>
<td>0.3845</td>
<td>0.0793</td>
</tr>
<tr>
<td>3ST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{5,i}$</td>
<td>Annual crash counts</td>
<td>0</td>
<td>157</td>
<td>11</td>
<td>2.6902</td>
</tr>
<tr>
<td>$T_{1,i}$</td>
<td>AADT on major approaches</td>
<td>2360</td>
<td>38100</td>
<td>20284</td>
<td>870.0525</td>
</tr>
<tr>
<td>$T_{2,i}$</td>
<td>AADT on minor approaches</td>
<td>500</td>
<td>10200</td>
<td>2826</td>
<td>299.6628</td>
</tr>
<tr>
<td>$FR_i$</td>
<td>Flow ratio ($T_{2,i}/T_{1,i}$)</td>
<td>0.0256</td>
<td>1</td>
<td>0.3545</td>
<td>0.0178</td>
</tr>
</tbody>
</table>
5.2 Model Development

Crash prediction models are developed for four types of urban intersections including: 1) four-way signalized intersections (4SG), 2) four-way stop controlled intersections (4ST), 3) signalized T-intersections (3SG), and 4) stop controlled T-intersections (3ST). Unfortunately, not all the required data, such as use of RTOR signal operation, use of red light cameras and type of left-turn phasing listed in Table 5.2 was available for all selected sites. So they are excluded from the model development.

5.2.1 Four-Way Signalized Intersections (4SG)

As shown in Table 5.3, 132 four-way signalized intersections are selected as study samples, 60% (roughly 80 sites) of which is used for model development and the other 40% is used for model validation. Different model forms and assumptions are first explored and attempts are made to mimic the spatial process appropriately.

5.2.1.1 Model Assumptions and Forms

The general model form explored is

\[ \mu_i = f(T_{1,i}, T_{2,i}, x_i; \beta) \exp(\varepsilon_i + \phi_i) \]  (5.1)

where, \( T_{1,i} \) represents the AADT on major approach, \( T_{2,i} \) represents the AADT on minor approach, \( f(.) \) is a positive valued function with different forms for different facility type.

a. Generalized Poisson Models

Poisson generalized linear models are first evaluated. There is no fixed model forms for analyzing traffic crash data so that a variety of model specifications were tested before narrowing down to four representative ones:
Model Form I: \[ \mu_i = \exp(\beta_0 + \beta_1 T_{1,i} + \beta_2 T_{2,i} + \sum \beta_i x_i) \]

Model Form II: \[ \mu_i = \exp(\beta_0 + \beta_1 T_{1,i} + \beta_2 T_{2,i}) \]

Model Form III: \[ \mu_i = \exp(\beta_0 + \sum \beta_i x_i) T_{1,i}^{\beta_2} T_{2,i}^{\beta_3} \]

Model Form IV: \[ \mu_i = \exp(\beta_0 + \beta_1 FR_i + \sum \beta_i x_i) \left[ T_{1,i}^{\beta_2} \exp(\beta_3 T_{2,i}) \right] \left[ T_{2,i}^{\beta_4} \exp(\beta_5 T_{1,i}) \right] \]

Model form III is the general form used in the HSM. In model form IV, \( FR_i \) stands for intersection flow ratio, \( FR_i = T_{2,i}/T_{1,i} \). The part, \( T_{1,i}^{\beta_2} \exp(\beta_3 T_{2,i}) \) and \( T_{2,i}^{\beta_4} \exp(\beta_5 T_{1,i}) \) are based on the idea that vehicles entering from the major and minor approaches may have different risks. This concept of different risks was first introduced by Miaou and Lord in 2003 [68] but the general forms applied here are modified from the original form. The underlying concept can be elaborated as follows. Vehicles entering from the major approach are exposed to a certain level of safety risks involving the vehicles itself, vehicles in the same direction and vehicles from opposing direction such as left-turning vehicles. This risk is captured by parameter \( \beta_2 \) in the part \( T_{1,i}^{\beta_2} \). Meanwhile, these vehicles are exposed to risks due to vehicles entering from minor approach and this risk is captured by the term \( \exp(\beta_3 T_{2,i}) \). Similar logic can be applied to minor approach vehicles so as to generate the term \( T_{2,i}^{\beta_4} \exp(\beta_5 T_{1,i}) \) in model IV. The concept is demonstrated in Figure 5.2.

The estimated parameters and model performance measures are presented in Table 5.5. Comparing the AIC values, Model IV has better fit than the other three model forms. The residual deviance is used to perform a goodness-of-fit test for the overall model. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed. Therefore, if the residual difference is small enough, the goodness of fit test will not be significant, indicating that the model fits the data well. The conclusion can be drawn that the model fits reasonably well because the goodness-of-fit chi-squared test is not statistically significant. If the test is statistically
significant, it would indicate that the data do not fit the model well. In that situation, we may try to determine if there are omitted predictor variables, if our linearity assumption holds and/or if there is an issue of over-dispersion. From Table 5.5, none of the four models fit the data well.

If the residual deviance is greater than the degrees of freedom, that suggests that the conditional mean is not equal the conditional variance. Therefore, one rule of thumb to check for over-dispersion is to look at the ratio of residual deviance to degrees of freedom. If it is greater than 1, then the data are usually over-dispersed. Therefore, the Poisson assumption does not hold. Another way to test the over-dispersion in the Poisson model was proposed by A. Colin Cameron and Pravin K. Trivedi in 1990 [16]. The test simply tests this assumption as a null hypothesis against an alternative where \( Var(y) = \mu + cf(\mu) \) where constant \( c > 0 \) indicates over-dispersion and \( c < 0 \) indicates under-dispersion. The test is equivalent to testing the hypothesis that \( H_0 : c = 0; \ H_1 : c \neq 0 \). The test results indicate that the true constant \( c \) in all four models are significantly greater than 0. Therefore, over-dispersion exists in the data and other model forms must be investigated.
Table 5.5: Generalized Poisson Models for 4SG

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Model Form I</th>
<th>Model Form II</th>
<th>Model Form III</th>
<th>Model Form IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>3.286</td>
<td>3.294</td>
<td>-4.57035</td>
<td>-7.645e-01</td>
</tr>
<tr>
<td></td>
<td>(4.521e-02)</td>
<td>(4.470e-02)</td>
<td>(0.41349)</td>
<td>(9.497e-01)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>1.977e-05</td>
<td>1.962e-05</td>
<td>-0.21018</td>
<td>-2.099e-01</td>
</tr>
<tr>
<td></td>
<td>(1.715e-06)</td>
<td>(1.710e-06)</td>
<td>(0.01179)</td>
<td>(1.240e-02)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>7.344e-05</td>
<td>7.326e-05</td>
<td>0.31291</td>
<td>-3.689e-01</td>
</tr>
<tr>
<td></td>
<td>(2.256e-06)</td>
<td>(2.254e-06)</td>
<td>(0.03910)</td>
<td>(1.886e-01)</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>-1.806e-01</td>
<td>-1.924e-01</td>
<td>0.66643</td>
<td>-3.138e-01</td>
</tr>
<tr>
<td></td>
<td>(1.545e-02)</td>
<td>(1.203e-02)</td>
<td>(0.02004)</td>
<td>(1.123e-01)</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>-1.533e-02</td>
<td>-</td>
<td>-</td>
<td>-1.307e-05</td>
</tr>
<tr>
<td></td>
<td>(1.267e-02)</td>
<td>(         )</td>
<td>(         )</td>
<td>(8.752e-06)</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.105e-01</td>
</tr>
<tr>
<td></td>
<td>(         )</td>
<td>(         )</td>
<td>(         )</td>
<td>(5.030e-02)</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.392e-05</td>
</tr>
<tr>
<td></td>
<td>(         )</td>
<td>(         )</td>
<td>(         )</td>
<td>(5.394e-06)</td>
</tr>
<tr>
<td>( df )</td>
<td>126</td>
<td>127</td>
<td>127</td>
<td>124</td>
</tr>
<tr>
<td>( \text{Residual deviance} )</td>
<td>2409.8</td>
<td>2411.3</td>
<td>1921.1</td>
<td>1879.8</td>
</tr>
<tr>
<td>( \text{MAD} )</td>
<td>(4.9236e-318)</td>
<td>(2.2362e-200)</td>
<td>(7.6691e-318)</td>
<td>(3.0823e-311)</td>
</tr>
<tr>
<td>( \text{MSPE} )</td>
<td>22.3635</td>
<td>22.3836</td>
<td>19.9753</td>
<td>19.8267</td>
</tr>
<tr>
<td>( \text{AIC} )</td>
<td>937.5214</td>
<td>935.8763</td>
<td>796.1142</td>
<td>781.9318</td>
</tr>
</tbody>
</table>

\( \cdot \) standard errors of estimated parameters

b. Generalized Poisson Mixture Models

Four similar model forms are assumed initially but Negative Binomial generalized linear models are fitted instead. The estimated parameters are presented in Table 5.7. The chi-square test results indicate that Model V to VIII fit data well, and much better than the Poisson generalized linear models. Among the models, model VIII has the best fit so far.

Next, the Poisson-Lognormal and Poisson-Inverse Gaussian models are fitted and presented in Table 5.8 and 5.9. Variables such as the presence of left turn lanes and presence of right turn lanes are found to be non-significant in the Poisson-Lognormal Model. The estimated parameters changed slightly when different model assumptions
Table 5.6: Over-Dispersion Test for Generalized Poisson Models

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Model Form I</th>
<th>Model Form II</th>
<th>Model Form III</th>
<th>Model Form IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample estimates</td>
<td>19.16334</td>
<td>19.17601</td>
<td>15.67931</td>
<td>15.08733</td>
</tr>
<tr>
<td>z value</td>
<td>4.5323</td>
<td>4.5616</td>
<td>3.5236</td>
<td>3.4729</td>
</tr>
<tr>
<td>p-value</td>
<td>2.917e-06</td>
<td>2.539e-06</td>
<td>0.0002129</td>
<td>0.0002574</td>
</tr>
</tbody>
</table>

and forms are applied. Comparing the goodness-of-fit measures, it can be seen that NB model VIII fits this data set the best. PIG fits the data similar as other models, a little less well than the PLN model. Overall, the generalized Poisson-mixture models provide better fit than the Poisson models. Model forms considering the conflicts between major and minor approaches provide better fit than others.
Table 5.7: Generalized Poisson-Gamma Models for 4SG

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Model Form V</th>
<th>Model Form VI</th>
<th>Model Form VII</th>
<th>Model Form VIII</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>3.025e+00(2.046e-01)</td>
<td>3.045e+00(2.034e-01)</td>
<td>-4.35927(1.44336)</td>
<td>-2.437e+00(2.495e+00)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>2.173e-05(7.848e-06)</td>
<td>2.085e-05(7.828e-06)</td>
<td>-0.20057(0.04887)</td>
<td>-1.926e-01(5.901e-02)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>9.507e-05(1.256e-05)</td>
<td>9.449e-05(1.248e-05)</td>
<td>0.23274(0.14444)</td>
<td>-1.170e-01(7.576e-01)</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>-1.571e-01(6.963e-02)</td>
<td>-1.778e-01(5.721e-02)</td>
<td>0.73012(0.06758)</td>
<td>-1.550e-01(2.940e-01)</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>-3.238e-02(5.886e-02)</td>
<td>-3.238e-02(5.886e-02)</td>
<td>-2.982e-05(3.435e-05)</td>
<td>9.335e-01(1.386e-01)</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>-3.238e-02(5.886e-02)</td>
<td>-3.238e-02(5.886e-02)</td>
<td>-2.982e-05(3.435e-05)</td>
<td>9.335e-01(1.386e-01)</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>-3.238e-02(5.886e-02)</td>
<td>-3.238e-02(5.886e-02)</td>
<td>-2.982e-05(3.435e-05)</td>
<td>9.335e-01(1.386e-01)</td>
</tr>
</tbody>
</table>

Dispersion parameter: 0.4759 0.4770 0.3375 0.3277

df: 126 127 127 124

Residual deviance: 146.53 146.54 137.03 136.82

Chi-square: 0.8981 0.8868 0.7439 0.7366

MAD: 23.9126 23.9138 20.2127 19.0526

MSPE: 1154.922 1146.265 812.4387 757.5182

AIC: 1225.3 1223.6 1173.3 1171

Table 5.8: Generalized Poisson-Lognormal Models for 4SG

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Model Form IX</th>
<th>Model Form X</th>
<th>Model Form XI</th>
<th>Model Form XII</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>2.448e+00(2.222e-01)</td>
<td>2.284e+00(2.063e-01)</td>
<td>-5.19445(1.50188)</td>
<td>-3.779e+00(3.359e+00)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>2.579e-05(8.351e-06)</td>
<td>2.167e-05(8.265e-06)</td>
<td>-0.14766(0.05037)</td>
<td>-1.316e-01(5.173e-02)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>1.052e-04(1.324e-05)</td>
<td>9.533e-05(1.251e-05)</td>
<td>0.26073(0.14999)</td>
<td>5.638e-01(8.176e-01)</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
</tr>
</tbody>
</table>

df: 125 127 127 127

Residual deviance: 114.5922 114.6250 81.4387 75.5182

Chi-square: 0.8981 0.8868 0.7439 0.7366

MAD: 23.9126 23.9138 20.2127 19.0526

MSPE: 1295.692 1350.764 1353.509 1352.252

AIC: 1229.6 1229.6 1171.2 1174.8

BIC: 1246.9 1241.1 1185.6 1197.8

Table 5.9: Generalized Poisson-Inverse Gaussian Models for 4SG

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Model Form XIII</th>
<th>Model Form XIV</th>
<th>Model Form XV</th>
<th>Model Form XVI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>2.670e+00(2.087e-01)</td>
<td>2.674e+00(2.085e-01)</td>
<td>-4.7728(1.47638)</td>
<td>-3.864e+00(4.201e+00)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>2.411e-05(8.030e-06)</td>
<td>2.392e-05(8.082e-06)</td>
<td>-0.1384(0.05041)</td>
<td>-1.273e-01(5.712e-02)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>1.090e-04(1.295e-05)</td>
<td>1.083e-04(1.286e-05)</td>
<td>0.762(0.07234)</td>
<td>3.753e-02(4.918e-01)</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>-8.807e-02(6.571e-02)</td>
<td>-1.033e-01(5.976e-02)</td>
<td>-3.327e-05(6.595e-05)</td>
<td>8.486e-01(1.494e-01)</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>-2.584e-02(5.207e-02)</td>
<td>– – –</td>
<td>– – –</td>
<td>1.486e-05(4.181e-05)</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
<td>– – –</td>
</tr>
</tbody>
</table>

df: 125 126 126 126

Residual deviance: 1220.596 1220.00 116.131 1162.744

Iterations: 5 3 5 5

MAD: 27.512 27.8321 20.903 20.3775

MSPE: 1783.791 1782.113 857.584 828.8007

AIC: 1229.6 1229.6 1171.2 1174.8

BIC: 1246.9 1241.1 1185.6 1197.8
Table 5.10: GOF Measurements of Unadjusted Models for 4SG

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>MAD</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model Form I</td>
<td>Model Form II</td>
<td>Model Form III</td>
</tr>
<tr>
<td>Poisson GLM</td>
<td>3109.1</td>
<td>22.3635</td>
<td>937.5214</td>
</tr>
<tr>
<td>NB GLM</td>
<td>1225.3</td>
<td>22.3836</td>
<td>935.8763</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>1246.9</td>
<td>23.9126</td>
<td>1146.2650</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>1232.6</td>
<td>23.8682</td>
<td>1295.6920</td>
</tr>
<tr>
<td></td>
<td>3108.5</td>
<td>23.9138</td>
<td>1146.2650</td>
</tr>
<tr>
<td></td>
<td>2618.3</td>
<td>24.0411</td>
<td>1350.764</td>
</tr>
<tr>
<td></td>
<td>2583.1</td>
<td>24.8321</td>
<td>1782.113</td>
</tr>
</tbody>
</table>

5.2.1.2 Spatially Adjusted Models for 4SG

Before including spatial effects into the crash prediction models, it is sensible to check whether the spatial auto-correlation is in fact displayed in the model residuals. The spatial weight matrix must be constructed before conducting the tests. Let $W = (w_{ij} : i,j = 1,...,n)$ denote the spatial weight matrix summarizing the spatial relationships between $n$ intersections in the study region. The “self influence” effect is excluded here; therefore, $w_{ii} = 0$ for all $i = 1,...,n$. The spatial weights of the 80 sample sites create an 80-by-80 matrix. Three forms of neighboring proximity are discussed for 4SG intersections:

- Distance-based spatial proximity;
- Adjacency-based spatial proximity; and
- Corridor-based spatial proximity.
a. Spatially adjusted models using distance-based weight matrix (Weight I)

Spatial weights are calculated based on distances between intersections (i.e., \( w_{ij} = d_{ij}^{-1} \), where \( d_{ij} \) is the distance between intersections \( i \) and \( j \)). Figure 5.4 shows the coordinates of the sample intersections used in the analysis. Model form VIII is adopted at this stage. Positive spatial auto-correlation is detected not only in fitted values of \( \mu_i (MC = 0.2076) \) but also in residuals of this model specification \( (MC = 0.1381) \).

![Figure 5.4: Coordinates of 4SG Intersections in the Study Region](image)

ESFs are constructed to account for the spatial auto-correlation in the data. By forward regression, eight eigenvectors are selected as spatial variables. The first spa-
tially adjusted model (SA-I) is identified with the following form.

\[ \mu_i = \exp(\beta_{X0} + \beta_{X1}x_{N.AW.LTL} + \beta_2 F_R + \beta_{E5}E V_5 + \beta_{E15}E V_{15} + \beta_{E12}E V_{12} + \beta_{E6}E V_6 + \beta_{E16}E V_{16} + \beta_{E14}E V_{14} + \beta_{E8}E V_8) \left[ T_{1,i}^{\beta_{X3}} \exp(\beta_{X4}T_{2,i}) \right] \left[ T_{2,i}^{\beta_{X5}} \exp(\beta_{X6}T_{1,i}) \right] \]

(5.2)
Table 5.11: Parameter Estimates of Spatial Adjusted Model SA-I for 4SG

<table>
<thead>
<tr>
<th>Parameter (Model Variable)</th>
<th>Spatial Adjusted Model I (SA-I)</th>
<th>Unadjusted Model VIII (UA-VIII)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.  Std.Error  p</td>
<td>Coef.  Std.Error  p</td>
</tr>
<tr>
<td>( \beta_{X0} ) (Intercept)</td>
<td>-2.015e+00  2.014e+00  0.0056</td>
<td>-2.437e+00  2.495e+00  0.0033</td>
</tr>
<tr>
<td>( \beta_{X1} ) (NAWLT)</td>
<td>-1.086e-01  4.549e-02  0.0006</td>
<td>-1.926e-01  5.001e-02  0.0001</td>
</tr>
<tr>
<td>( \beta_{X2} ) (( FR_i ))</td>
<td>2.963e-01  6.840e-01  3.7e-14</td>
<td>-1.170e-01  7.576e-01  0.0088</td>
</tr>
<tr>
<td>( \beta_{X3} ) (( T_{1,i} ))</td>
<td>-5.232e-02  2.410e-01  0.0006</td>
<td>-1.550e-01  2.940e-01  0.0480</td>
</tr>
<tr>
<td>( \beta_{X4} ) (( T_{2,i} ))</td>
<td>-3.084e-05  3.031e-05  0.0001</td>
<td>-2.982e-05  3.435e-05  0.0038</td>
</tr>
<tr>
<td>( \beta_{X5} ) (( T_{2,i} ))</td>
<td>7.174e-01  1.269e-01  1.24e-10</td>
<td>9.335e-01  1.386e-01  1.65e-11</td>
</tr>
<tr>
<td>( \beta_{X6} ) (( T_{1,i} ))</td>
<td>1.858e-05  1.467e-05  0.0399</td>
<td>1.871e-05  1.665e-05  0.0267</td>
</tr>
<tr>
<td>( \beta_{E5} ) (( EV_{5} ))</td>
<td>1.405e+00  5.100e-01  0.0059</td>
<td>–  –  –</td>
</tr>
<tr>
<td>( \beta_{E15} ) (( EV_{15} ))</td>
<td>-1.331e+00  5.194e-01  0.0104</td>
<td>–  –  –</td>
</tr>
<tr>
<td>( \beta_{E12} ) (( EV_{12} ))</td>
<td>1.378e+00  5.164e-01  0.0076</td>
<td>–  –  –</td>
</tr>
<tr>
<td>( \beta_{E6} ) (( EV_{6} ))</td>
<td>1.561e+00  6.220e-01  0.0121</td>
<td>–  –  –</td>
</tr>
<tr>
<td>( \beta_{E16} ) (( EV_{16} ))</td>
<td>1.289e+00  5.745e-01  0.0249</td>
<td>–  –  –</td>
</tr>
<tr>
<td>( \beta_{E14} ) (( EV_{14} ))</td>
<td>-1.093e+00  5.512e-01  0.0474</td>
<td>–  –  –</td>
</tr>
<tr>
<td>( \beta_{E9} ) (( EV_{9} ))</td>
<td>-1.230e+00  5.431e-01  0.0236</td>
<td>–  –  –</td>
</tr>
<tr>
<td>( \beta_{E8} ) (( EV_{8} ))</td>
<td>8.954e-01  5.123e-01  0.0805</td>
<td>–  –  –</td>
</tr>
</tbody>
</table>

Residual deviance  136.75  136.82
Over-dispersion parameter  0.2332  0.3277
  MAD  16.1871  19.0526
  MSPE  564.2973  757.5182
  AIC  1149.2  1171
The spatial adjusted model (SA-I) fits the data much better than the unadjusted model VIII as seen in Figure 5.5 and Table 5.11. The spatial auto-correlation test was performed for model SA-I and no significant spatial coefficient was found in the adjusted model. It can be seen that the spatial adjusted model accounts for the spatial correlations.

Figure 5.5: Comparison of Spatially Adjusted Model SA-I and Unadjusted Model UA-VIII for 4SG
b. Spatially adjusted models using adjacency-based weight matrix (Weight II)

This weight definition emphasizes the correlations between intersections that are di-
rectly connected. Intersections sharing the same segment are assigned spatial weight
of 1, and 0 is assigned to other sites. Positive spatial auto-correlation is detected
in fitted values of $\mu_i$ ($MC = 0.7582$) and also in model residuals ($MC = 0.5134$).
The spatial coefficients are much higher using the contiguity based weight than the
distance-based weight. The spatial adjusted model (SA-II) has the form of (5.3).
Only one eigenvector was selected to account for spatial effects.

$$
\mu_i = \exp(\beta X_0 + \beta X_1 x_{NAWLTL_i} + \beta X_2 FR_i + \beta X_6 EV_6)
\left[ T^{\delta_{X_1}}_{1,i} \exp(\beta X_4 T_{2,i}) \right] \left[ T^{\delta_{X_2}}_{2,i} \exp(\beta X_6 T_{1,i}) \right]
$$

(5.3)

According to Table 5.12 and Figure 5.6, model SA-II with binary spatial weight
does not perform as well as model SA-I with a distance-based weight matrix. The
binary weight matrix purely considers the connections between adjacent intersections
and ignores the spatial correlations between the site of interest and other sites on the
road network. Even though the connections with remote sites may not be as strong
as adjacent ones, there are still potential correlations that are significant.
Table 5.12: Parameter Estimates of Spatial Adjusted Model SA-II for 4SG

<table>
<thead>
<tr>
<th>Parameter (Model Variable)</th>
<th>Spatial Adjusted Model II (SA-II)</th>
<th></th>
<th></th>
<th>Unadjusted Model VIII (UA-VIII)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std.Error</td>
<td>p</td>
<td>Coef.</td>
<td>Std.Error</td>
<td>p</td>
</tr>
<tr>
<td>$\beta_{X0}$ (Intercept)</td>
<td>-2.280e+00</td>
<td>2.346e-01</td>
<td>0.0003</td>
<td>-2.437e+00</td>
<td>2.495e+00</td>
<td>0.0033</td>
</tr>
<tr>
<td>$\beta_{X1}$ (NAWLTL)</td>
<td>-1.872e-01</td>
<td>4.894e-02</td>
<td>0.0001</td>
<td>-1.926e-01</td>
<td>5.001e-02</td>
<td>0.0001</td>
</tr>
<tr>
<td>$\beta_{X2}$ ($F_{R_i}$)</td>
<td>3.988e-02</td>
<td>7.581e-01</td>
<td>0.0095</td>
<td>-1.170e-01</td>
<td>7.756e-01</td>
<td>0.0088</td>
</tr>
<tr>
<td>$\beta_{X3}$ ($T_{1,i}$)</td>
<td>-1.366e-01</td>
<td>2.780e-01</td>
<td>0.0033</td>
<td>-1.550e-01</td>
<td>2.940e-01</td>
<td>0.0480</td>
</tr>
<tr>
<td>$\beta_{X4}$ ($T_{2,i}$)</td>
<td>-3.237e-05</td>
<td>3.353e-05</td>
<td>0.0001</td>
<td>-2.982e-05</td>
<td>3.435e-05</td>
<td>0.0038</td>
</tr>
<tr>
<td>$\beta_{X5}$ ($T_{2,i}$)</td>
<td>8.877e-01</td>
<td>1.360e-01</td>
<td>6.66e-11</td>
<td>9.335e-01</td>
<td>1.386e-01</td>
<td>1.65e-11</td>
</tr>
<tr>
<td>$\beta_{X6}$ ($T_{1,i}$)</td>
<td>1.853e-05</td>
<td>1.604e-05</td>
<td>0.0305</td>
<td>1.871e-05</td>
<td>1.665e-05</td>
<td>0.0267</td>
</tr>
<tr>
<td>$\beta_{E6}$ ($EV_{6}$)</td>
<td>1.242e+00</td>
<td>5.739e-01</td>
<td>0.0304</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Residual deviance</td>
<td>136.98</td>
<td>–</td>
<td>–</td>
<td>136.82</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Over-dispersion parameter</td>
<td>0.3119</td>
<td>–</td>
<td>–</td>
<td>0.3277</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MAD</td>
<td>18.6096</td>
<td>19.0526</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSPE</td>
<td>703.4888</td>
<td>757.5182</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>1169.5</td>
<td>1171</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
c. Spatially adjusted models using corridor-based weight matrix

In recent publications, the common assumption for analyzing intersections along corridors is that the intersections within the same corridor have spatial weight of 1 while other sites have 0 weight. In other words, intersections are only correlated to the ones within the same corridor. This underlying assumption has potential issues that may impact the model fitting. Whether the connections between intersections within and outside of corridors are trivial needs to be explored. The actual spatial process should be mimicked appropriately. Figure 5.7 presents the coordinates of the sample corridors employed in the analysis.
The following three possible spatial correlations of intersections by corridors are analyzed next:

- Distance-based spatial weight by corridors;
- Adjacency-based spatial weight by corridors; and
- Combined spatial weight by corridors.

c1. Distance-based spatial weight by corridors (Weight III)
In the first scenario considered, only intersections within the same corridor are correlated, and the spatial weight is modeled using the inverse distance between sites of
interest.

\[ w_{ij} = \begin{cases} 
  d^{-1}_{ij} & \text{if } i, j \in k \\
  0 & \text{O/W}
\end{cases} \]

Again, significant spatial correlations were observed in the data with \( MC = 0.2769 \) in the fitted values and \( MC = 0.1879 \) in the residuals. The spatial adjusted model (SA-III) is developed and presented in Equation (5.4) in which 11 eigenvectors are selected to model the spatial effects as follows:

\[
\mu_i = \exp(\beta_0 + \beta_{x_1} x_{N.AW.LTL} + \beta_2 F R_i + \beta_{E14} E V_{14} + \beta_{E7} E V_{7} + \beta_{E6} E V_{6} + \beta_{E17} E V_{17} + \\
\beta_{E11} E V_{11} + \beta_{E4} E V_{4} + \beta_{E12} E V_{12} + \beta_{E26} E V_{26} + \beta_{E10} E V_{10} + \beta_{E2} E V_{2} + \beta_{E20} E V_{20}) \\
\left[ T_{1,i}^{\beta_{x_3} \text{EXP}(\beta_{x_4} T_{2,i})} \right] \left[ T_{2,i}^{\beta_{x_5} \text{EXP}(\beta_{x_6} T_{1,i})} \right]
\]

(5.4)

c2. Adjacency-based spatial weight by corridors (Weight IV)

In the second scenario considered, only intersections within the same corridor are correlated, and the weight is modeled using the binary values 0 and 1.

\[ w_{ij} = \begin{cases} 
  1 & \text{if } i, j \in k \\
  0 & \text{O/W}
\end{cases} \]

\( MC = 0.7164 \) for fitted values and \( MC = 0.4792 \) for residuals were obtained. The spatial adjusted model (SA-IV) is presented as:

\[
\mu_i = \exp(\beta_0 + \beta_{x_1} x_{N.AW.LTL} + \beta_2 F R_i + \beta_{E5} E V_{5} + \beta_{E11} E V_{11} + \beta_{E2} E V_{2} + \\
\beta_{E17} E V_{17} + \beta_{E12} E V_{12}) \left[ T_{1,i}^{\beta_{x_3} \text{EXP}(\beta_{x_4} T_{2,i})} \right] \left[ T_{2,i}^{\beta_{x_5} \text{EXP}(\beta_{x_6} T_{1,i})} \right]
\]

(5.5)

Estimated model parameters shown in Table 5.13 are quite different for two spatially adjusted models. Nevertheless, there was little difference found between the
model fitting of model SA-III and model SA-IV and they both provide significantly better model fitting and predicting compared to the unadjusted models. The distance-based weight matrix provides slightly better fit than the binary-based model because of lower AIC value and model errors.

Figure 5.8: Comparison of Spatially Adjusted Models SA-III and SA-IV for 4SG
Table 5.13: Parameter Estimates of Spatial Adjusted Models SA-III and SA-IV for 4SG

<table>
<thead>
<tr>
<th>Parameter (Model Variable)</th>
<th>Spatial Adjusted Model III (SA-III)</th>
<th>Parameter (Model Variable)</th>
<th>Spatial Adjusted Model IV (SA-IV)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std.Error</td>
<td>p</td>
</tr>
<tr>
<td>$\beta_{X0}(\text{Intercept})$</td>
<td>-2.052e+00</td>
<td>1.823e+00</td>
<td>0.0026</td>
</tr>
<tr>
<td>$\beta_{X1}(\text{NAWLTL})$</td>
<td>-8.126e-02</td>
<td>4.657e-02</td>
<td>0.0809</td>
</tr>
<tr>
<td>$\beta_{X2}(\text{FR}_i)$</td>
<td>9.239e-04</td>
<td>6.881e-01</td>
<td>0.0010</td>
</tr>
<tr>
<td>$\beta_{X3}(T_{1,i})$</td>
<td>-1.629e-01</td>
<td>2.269e-01</td>
<td>0.0472</td>
</tr>
<tr>
<td>$\beta_{X4}(T_{2,i})$</td>
<td>-3.702e-05</td>
<td>3.122e-05</td>
<td>0.0023</td>
</tr>
<tr>
<td>$\beta_{X5}(T_{2,i})$</td>
<td>8.346e-01</td>
<td>1.290e-01</td>
<td>9.98e-11</td>
</tr>
<tr>
<td>$\beta_{X6}(T_{1,i})$</td>
<td>2.726e-05</td>
<td>1.472e-05</td>
<td>0.0639</td>
</tr>
<tr>
<td>$\beta_{E14}(\text{EV}_{14})$</td>
<td>1.733e+00</td>
<td>5.207e-01</td>
<td>0.0008</td>
</tr>
<tr>
<td>$\beta_{E7}(\text{EV}_7)$</td>
<td>-1.241e+00</td>
<td>5.456e-01</td>
<td>0.0229</td>
</tr>
<tr>
<td>$\beta_{E6}(\text{EV}_6)$</td>
<td>1.150e+00</td>
<td>5.085e-01</td>
<td>0.0237</td>
</tr>
<tr>
<td>$\beta_{E17}(\text{EV}_{17})$</td>
<td>-1.310e+00</td>
<td>5.170e-01</td>
<td>0.0113</td>
</tr>
<tr>
<td>$\beta_{E11}(\text{EV}_{11})$</td>
<td>-1.188e+00</td>
<td>5.539e-01</td>
<td>0.0319</td>
</tr>
<tr>
<td>$\beta_{E4}(\text{EV}_4)$</td>
<td>1.407e+00</td>
<td>5.445e-01</td>
<td>0.0097</td>
</tr>
<tr>
<td>$\beta_{E2}(\text{EV}_2)$</td>
<td>9.147e-01</td>
<td>5.426e-01</td>
<td>0.0918</td>
</tr>
<tr>
<td>$\beta_{E20}(\text{EV}_{20})$</td>
<td>-7.435e-01</td>
<td>5.241e-01</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

Residual deviance: 136.25 136.29
Over-dispersion parameter:
- MAD: 15.1297 16.4666
- MSPE: 539.9289 603.0610
- AIC: 1152.1 1153.1
c3. Combined weight by corridors (Weight V)

A trial is made to customize the spatial weight matrix to improve overall model fitting. The spatial weight matrix is designed to highlight the inner correlations among intersections within the same corridor, as well as take the potential correlations between intersections and exterior sites into account.

\[ w_{ij} = \begin{cases} 
1 & \text{iff } i, j \in k \\
\frac{1}{d_{ij}} & \text{O/W}
\end{cases} \]

By creating the new spatial matrix, the spatial coefficient was found significant in the fitted values of model VIII \((MC = 0.2028)\) and in the model residuals \((MC = 0.125)\). The spatially adjusted model (SA-V) is presented in (5.6).

Figure 5.9: Comparison of Spatially Adjusted Model SA-V and Unadjusted Model UA-VIII for 4SG
\[ \mu_i = \exp(\beta_{X0} + \beta_{X1}x_i + \beta_2 FR_i + \beta_{E10} EV_{10} + \beta_{E22} EV_{22} + \beta_{E7} EV_7 + \beta_{E6} EV_6 + \beta_{E16} EV_{16} + \beta_{E3} EV_3 + \beta_{E15} EV_{15} + \beta_{E4} EV_4 + \beta_{E3} EV_5) \left[ T_{1,i}^{\beta_{X3}} EXP(\beta_{X3}T_{1,i}) \right] \left[ T_{2,i}^{\beta_{X4}} EXP(\beta_{X6}T_{1,i}) \right] \]

Table 5.14: Parameter Estimates of Spatial Adjusted Model SA-V for 4SG

<table>
<thead>
<tr>
<th>Parameter (Model Variable)</th>
<th>Spatial Adjusted Model V (SA-V)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef. Estimate</td>
</tr>
<tr>
<td>(\beta_{X0}) (Intercept)</td>
<td>-2.371e+00</td>
</tr>
<tr>
<td>(\beta_{X1}) (NAWLTL)</td>
<td>-3.530e-02</td>
</tr>
<tr>
<td>(\beta_{X2}) (FR_i)</td>
<td>3.668e-01</td>
</tr>
<tr>
<td>(\beta_{X3}) (T_1,i)</td>
<td>-9.312e-02</td>
</tr>
<tr>
<td>(\beta_{X4}) (T_2,i)</td>
<td>-3.539e-05</td>
</tr>
<tr>
<td>(\beta_{X5}) (T_3,i)</td>
<td>7.524e-01</td>
</tr>
<tr>
<td>(\beta_{X6}) (T_4,i)</td>
<td>2.862e-05</td>
</tr>
<tr>
<td>(\beta_{E10}) (EV_10)</td>
<td>-2.131e+00</td>
</tr>
<tr>
<td>(\beta_{E22}) (EV_22)</td>
<td>-1.495e+00</td>
</tr>
<tr>
<td>(\beta_{E7}) (EV_7)</td>
<td>1.261e+00</td>
</tr>
<tr>
<td>(\beta_{E6}) (EV_6)</td>
<td>-1.296e+00</td>
</tr>
<tr>
<td>(\beta_{E16}) (EV_16)</td>
<td>1.034e+00</td>
</tr>
<tr>
<td>(\beta_{E3}) (EV_3)</td>
<td>1.097e+00</td>
</tr>
<tr>
<td>(\beta_{E15}) (EV_15)</td>
<td>-8.199e-01</td>
</tr>
<tr>
<td>(\beta_{E4}) (EV_4)</td>
<td>9.184e-01</td>
</tr>
<tr>
<td>(\beta_{E5}) (EV_5)</td>
<td>6.680e-01</td>
</tr>
<tr>
<td>Residual deviance</td>
<td>138.75</td>
</tr>
<tr>
<td>Over-dispersion parameter</td>
<td>0.1778</td>
</tr>
<tr>
<td>MAD</td>
<td>13.7487</td>
</tr>
<tr>
<td>MSPE</td>
<td>407.5019</td>
</tr>
<tr>
<td>AIC</td>
<td>1122.8</td>
</tr>
</tbody>
</table>

5.2.1.3  Comparison of Developed Models for 4SG

The standardized AIC defined as AIC/n where n is the sample size of the focused data set is used to compare model goodness-of-fit. For the 4SG data set, the standardized AIC value decreased from 9.66 to 8.94 when including traffic volumes and other explanatory variables. When spatial effects are included in the model, it further
drops to 8.76 (as shown in column four in Table 5.15). Model SA-V with corridor-based spatial weights produces the lowest AIC/n which indicates that this model provides the best fit. The standardized AIC value dropped to 8.57 for model SA-V. The contribution of the spatial component associated with the ESF parts in the model is calculated as the reduction from each spatially adjusted model. For example, the contribution of model SA-V is calculated as the ratio of the difference between model SA-V and UA-VIII and the difference between model SA-V and NBGLM* (i.e., the model with only a constant and random error). Therefore, model SA-V accounts for approximately 34% of the total spatial effects while model SA-II (with binary spatial weights) accounts for 11% of the spatial effects.

The spatial correlation tests were conducted for the residuals of each model. Significant spatial correlations were found in the first two models but no significant MC values were found when the spatial effects are considered in the spatially adjusted models except for model SA-II. The dispersion parameters for the developed models also demonstrated a downward trend which indicates that model SA-V provides the best fit.

<table>
<thead>
<tr>
<th>Table 5.15: Comparison of Model GOF Statistics for 4SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>****</td>
</tr>
<tr>
<td>Contribution of spatial component</td>
</tr>
<tr>
<td>Dispersion parameter</td>
</tr>
<tr>
<td>MC in residuals</td>
</tr>
</tbody>
</table>

*NBGLM here indicates the model with only a constant and random error.

**Significant Moran’s I Coefficient
5.2.2 Four-Way Stop Controlled Intersections (4ST)

A similar modeling procedure is conducted for 4ST intersection data set. Related results are presented in Table 5.16-5.18. Only traffic volumes on major and minor approaches were found to be significant in the developed model. According to Table 5.16 and 5.17, spatial adjusted models fit the data much better than unadjusted ones. In addition, the model with corridor-based weights (Weight V in Table 5.17) perform better than other models. It was found that the Poisson-Inverse Gaussian model fits this specific data set better than other models. The best fit model found is:

$$
\mu_i = \exp(\beta X_0 + \beta E_5 E V_5 + \beta E_4 E V_4 + \beta E_7 E V_7 + \beta E_2 E V_2) T^{\beta X_1}_{1,i} T^{\beta X_2}_{2,i} \quad (5.7)
$$

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>MAD</th>
<th>MSPE</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson GLM</td>
<td>446.55</td>
<td>4.7397</td>
<td>76.9140</td>
<td>6.3019</td>
</tr>
<tr>
<td>NB GLM</td>
<td>293.20</td>
<td>4.7442</td>
<td>121.9212</td>
<td>6.2173</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>290.50</td>
<td>4.3231</td>
<td>105.5332</td>
<td>5.5086</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>290.17</td>
<td>4.1933</td>
<td>10.2258</td>
<td>5.3996</td>
</tr>
<tr>
<td>Poisson GLM</td>
<td>447.04</td>
<td>5.9157</td>
<td>76.7069</td>
<td>6.2173</td>
</tr>
<tr>
<td>NB GLM</td>
<td>289.67</td>
<td>6.2573</td>
<td>190.3096</td>
<td>1.4888</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>288.00</td>
<td>4.3673</td>
<td>132.2472</td>
<td>2.0875</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>288.52</td>
<td>4.2579</td>
<td>10.1664</td>
<td>1.0984</td>
</tr>
<tr>
<td>Poisson GLM</td>
<td>415.16</td>
<td>5.7677</td>
<td>71.6641</td>
<td>5.5086</td>
</tr>
<tr>
<td>NB GLM</td>
<td>283.22</td>
<td>5.5349</td>
<td>76.4916</td>
<td>1.1190</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>280.50</td>
<td>4.3819</td>
<td>78.0783</td>
<td>2.8322</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>277.37</td>
<td>4.4068</td>
<td>10.4269</td>
<td>1.0625</td>
</tr>
<tr>
<td>Poisson GLM</td>
<td>400.05</td>
<td>1.8692</td>
<td>66.1672</td>
<td>5.3996</td>
</tr>
<tr>
<td>NB GLM</td>
<td>281.27</td>
<td>1.8513</td>
<td>102.6541</td>
<td>1.1764</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>278.30</td>
<td>1.8857</td>
<td>79.5956</td>
<td>2.1974</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>281.23</td>
<td>1.9000</td>
<td>10.4721</td>
<td>1.2147</td>
</tr>
</tbody>
</table>
Table 5.17: GOF Measurements of Spatially Adjusted Models for 4ST

<table>
<thead>
<tr>
<th></th>
<th>Weight I</th>
<th>Weight II</th>
<th>Weight III</th>
<th>Weight IV</th>
<th>Weight V</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>262.0849</td>
<td>265.386</td>
<td>255.2360</td>
<td>257.0516</td>
<td>220.1809</td>
</tr>
<tr>
<td>MAD</td>
<td>3.7418</td>
<td>10.5263</td>
<td>2.0120</td>
<td>2.5263</td>
<td>0.6500</td>
</tr>
<tr>
<td>MSPE</td>
<td>59.4963</td>
<td>75.2062</td>
<td>32.7500</td>
<td>40.2056</td>
<td>9.5277</td>
</tr>
<tr>
<td>AD</td>
<td>2.8591</td>
<td>5.2179</td>
<td>2.5411</td>
<td>2.8800</td>
<td>1.0518</td>
</tr>
</tbody>
</table>

Table 5.18: Parameter Estimates of Spatial Adjusted Model for 4ST

| Parameter (Model Variable) | Coef. Estimate | Std.Error | Pr(>|z|) |
|----------------------------|----------------|-----------|---------|
| $\beta_{X_0}$ (Intercept) | -4.1170        | 1.0014    | 0.0001787 |
| $\beta_{X_1}(T_{1,i})$   | 0.3530         | 0.1522    | 0.0253363 |
| $\beta_{X_2}(T_{2,i})$   | 0.3414         | 0.1205    | 0.0070445 |
| $\beta_{E_5}(EV_5)$      | 2.7052         | 0.7744    | 0.0011386 |
| $\beta_{E_4}(EV_4)$      | 2.8384         | 1.0590    | 0.0104620 |
| $\beta_{E_7}(EV_7)$      | -1.3536        | 0.6953    | 0.0582643 |
| $\beta_{E_2}(EV_2)$      | -1.5732        | 0.8839    | 0.0423211 |

5.2.3 Signalized T-Intersections (3SG)

The same process is repeated to select the best model to fit signalized T-intersections. The Poisson-Inverse Gaussian model fits better than others. It should be noted that among the investigated facility types 3SG data set has the lowest number of sites with available data. Nonetheless, the spatial adjusted Poisson-Inverse Gaussian models perform better than the other models. Figure 5.10 shows the plots of model residuals and Q-Q plot of standardized residuals and confirmed that the model fits this small data set relatively well. The spatially adjusted model with corridor-based weight matrix (weight V) provides better fit.

$$
\mu_i = \exp(\beta_{X_0} + \beta_1 R_i + \beta_{E_3} EV_3 + \beta_{E_2} EV_2) \\
\left[ T_{1,i}^{\beta_{X_3}} \exp(\beta_{X_4} T_{2,i}) \right] \left[ T_{2,i}^{\beta_{X_5}} \exp(\beta_{X_6} T_{1,i}) \right]
$$

(5.8)
Table 5.19: GOF Measurements of Unadjusted Models for 3SG

<table>
<thead>
<tr>
<th>Model Form</th>
<th>AIC</th>
<th>MAD</th>
<th>MSPE</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson GLM</td>
<td>309.43</td>
<td>13.8209</td>
<td>301.7731</td>
<td>14.3041</td>
</tr>
<tr>
<td>NB GLM</td>
<td>175.59</td>
<td>13.8209</td>
<td>301.7731</td>
<td>14.3012</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>177.2</td>
<td>13.3239</td>
<td>285.1330</td>
<td>13.8201</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>176.9677</td>
<td>13.3239</td>
<td>285.1330</td>
<td>13.8201</td>
</tr>
</tbody>
</table>

Table 5.20: GOF Measurements of Spatially Adjusted Models for 3SG

<table>
<thead>
<tr>
<th>Weight</th>
<th>AIC</th>
<th>MAD</th>
<th>MSPE</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight I</td>
<td>151.0666</td>
<td>1.7513</td>
<td>1.4383</td>
<td>2.2056</td>
</tr>
<tr>
<td>Weight II</td>
<td>160.0767</td>
<td>7.2053</td>
<td>32.2036</td>
<td>3.2077</td>
</tr>
<tr>
<td>Weight III</td>
<td>159.5961</td>
<td>1.6230</td>
<td>1.3020</td>
<td>1.5200</td>
</tr>
<tr>
<td>Weight IV</td>
<td>159.4156</td>
<td>2.5502</td>
<td>16.3205</td>
<td>4.1112</td>
</tr>
<tr>
<td>Weight V</td>
<td>132.0572</td>
<td>2.5359</td>
<td>1.1057</td>
<td>1.1015</td>
</tr>
</tbody>
</table>

Table 5.21: Parameter Estimates of Spatial Adjusted Model for 3SG

| Parameter (Model Variable) | Coef. Estimate | Std.Error | Pr(>|z|) |
|---------------------------|----------------|-----------|---------|
| $\beta_{X0}(\text{Intercept})$ | 30.1878743 | 9.425e+00 | 0.0084116 |
| $\beta_{X1}(FR_{i})$ | -5.3193672 | 1.086e+00 | 0.0004741 |
| $\beta_{X2}(T_{1,i})$ | 0.0000988 | 4.262e-05 | 0.0406951 |
| $\beta_{X3}(T_{2,i})$ | 0.0003219 | 9.307e-05 | 0.0053409 |
| $\beta_{X4}(T_{2,i})$ | 0.9009547 | 4.752e-01 | 0.0845302 |
| $\beta_{X5}(T_{1,i})$ | -3.7771173 | 1.042e+00 | 0.0040004 |
| $\beta_{E3}(EV_{3})$ | -1.9094649 | 4.109e-01 | 0.0007079 |
| $\beta_{E2}(EV_{2})$ | -0.9502768 | 3.275e-01 | 0.0143973 |
5.2.4 Stop Controlled T-Intersections (3ST)

For this specific data set, the spatially adjusted model fits the data significantly better than others. In terms of heterogeneity, the Poisson mixtures provide similar parameter estimates. The best model found is presented in (5.9) and the corresponding parameters can be found in Table 5.24.

\[
\mu_i = exp(\beta_{X_0} + \beta_{E_6}EV_6 + \beta_{E_4}EV_4 + \beta_{E_1}EV_1 + \beta_{E_3}EV_3)T_{1,i}^{\beta_{X_1}}T_{2,i}^{\beta_{X_2}} \quad (5.9)
\]
### Table 5.22: GOF Measurements of Unadjusted Models for 3ST

<table>
<thead>
<tr>
<th>AIC</th>
<th>Model Form I</th>
<th>Model Form II</th>
<th>Model Form III</th>
<th>Model Form IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson GLM</td>
<td>920.13</td>
<td>920.13</td>
<td>874.55</td>
<td>842.68</td>
</tr>
<tr>
<td>NB GLM</td>
<td>393.38</td>
<td>389.99</td>
<td>387.13</td>
<td>392.94</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>383.8</td>
<td>378.7</td>
<td>377.0</td>
<td>382.7</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>386.4715</td>
<td>381.6744</td>
<td>371.0604</td>
<td>385.0053</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAD</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson GLM</td>
<td>7.8923</td>
<td>7.8923</td>
<td>7.7587</td>
<td>7.3944</td>
</tr>
<tr>
<td>NB GLM</td>
<td>8.2039</td>
<td>8.6114</td>
<td>7.3857</td>
<td>7.4475</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>7.2867</td>
<td>7.3007</td>
<td>7.0657</td>
<td>7.0147</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>0.7936</td>
<td>0.7927</td>
<td>0.8180</td>
<td>0.8177</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MSPE</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson GLM</td>
<td>367.9543</td>
<td>367.9543</td>
<td>353.4705</td>
<td>331.2179</td>
</tr>
<tr>
<td>NB GLM</td>
<td>432.3537</td>
<td>479.3373</td>
<td>386.6308</td>
<td>388.1604</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>395.0357</td>
<td>405.9691</td>
<td>382.8795</td>
<td>381.663</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>1.1424</td>
<td>1.1155</td>
<td>1.1587</td>
<td>1.1660</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AD</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson GLM</td>
<td>12.9148</td>
<td>12.9148</td>
<td>12.0709</td>
<td>12.0384</td>
</tr>
<tr>
<td>NB GLM</td>
<td>1.1651</td>
<td>1.1001</td>
<td>1.1093</td>
<td>1.1684</td>
</tr>
<tr>
<td>PLN GLM</td>
<td>1.9776</td>
<td>1.6196</td>
<td>1.5892</td>
<td>1.9574</td>
</tr>
<tr>
<td>PIG GLM</td>
<td>1.0277</td>
<td>1.6727</td>
<td>1.5261</td>
<td>1.0201</td>
</tr>
</tbody>
</table>

### Table 5.23: GOF Measurements of Spatially Adjusted Models for 3ST

<table>
<thead>
<tr>
<th>Weight I</th>
<th>Weight II</th>
<th>Weight III</th>
<th>Weight IV</th>
<th>Weight V</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>369.3098</td>
<td>373.502</td>
<td>352.206</td>
<td>370.205</td>
</tr>
<tr>
<td>MAD</td>
<td>1.2052</td>
<td>32.2056</td>
<td>0.9523</td>
<td>6.2056</td>
</tr>
<tr>
<td>MSPE</td>
<td>1.0526</td>
<td>2.0279</td>
<td>1.0529</td>
<td>1.7900</td>
</tr>
<tr>
<td>AD</td>
<td>1.5793</td>
<td>8.8063</td>
<td>3.2069</td>
<td>1.0526</td>
</tr>
</tbody>
</table>

### Table 5.24: Parameter Estimates of Spatial Adjusted Model for 3ST

| Parameter (Model Variable) | Coef. Estimate | Std.Error | Pr(>|z|) |
|----------------------------|----------------|-----------|---------|
| $\beta_{X_0}(\text{Intercept})$ | -7.8417        | 2.9194    | 0.009683 |
| $\beta_{X_1}(T_{1,i})$ | 0.6913         | 0.2956    | 0.023246 |
| $\beta_{X_2}(T_{2,i})$ | 0.3928         | 0.1429    | 0.008211 |
| $\beta_{E_6}(EV_6)$ | -2.7559        | 0.8031    | 0.001185 |
| $\beta_{E_4}(EV_4)$ | 2.5434         | 1.0518    | 0.019136 |
| $\beta_{E_1}(EV_1)$ | -1.5677        | 0.8620    | 0.074723 |
| $\beta_{E_3}(EV_3)$ | -1.5934        | 1.0860    | 0.014844 |
5.3 Chapter Summary

This chapter presents the development of crash prediction models for intersections along designated corridors. Efforts were made in regards to data preparation and model development.

On one hand, a large number of study sites were selected so that a sufficient sample size was available for conducting valid statistical analyses. Site selection was a multi-level process, with each level focusing on eliminating potential issues and biases caused by specific factors. On the other hand, data acquisition involves collecting intersection-related crash data, traffic exposure data and intersection geometry data. Raw vehicular traffic volumes and crash records were provided by NDOT. Nevertheless, several vehicular volumes at minor approach of intersections were estimated based on similar sites. Sufficient data was collected for three types of urban intersections, four-way signalized intersections, four-way stop controlled intersections and stop controlled T-intersections. Limited data was assembled for signalized T-intersections.

While checking model assumptions and forms, the conflicts between major and minor approach traffic were fully discussed. Model forms that provide the best fit were found for each and every one of the data set. For the unfiltered models, spatial correlations were found in both model residuals and fitted values. ESF adjusted models were therefore proposed and compared with the unadjusted ones. No significant spatial correlations were found in the adjusted model. The results indicate that spatially adjusted crash prediction models successfully absorb latent auto-correlation and therefore, prevent parameter estimation bias. The Poisson-Inverse Gaussian model was found to provide slightly better fit for 4ST, 3SG and 3ST intersections. Model Form III was found to fit stop controlled intersections the best whereas model Form
IV was found to fit signalized intersections the best. Distance-based spatial proximity is recommended instead of adjacency-based neighboring proximity. The customized weights provide the best fit.
Chapter 6  Crash Prediction Models for Segments along Corridors

Segments are analyzed separately from intersections along urban corridors. This chapter develops crash prediction models for roadway segments using data from focused region and explores spatial correlations between segments on a road network. Efforts are made first to compile information from different data sources with intent to prepare combined segment data set for data processing and model development. The number of sites with available data is humongous (7186 segments with 474 miles) in the study region. However the data quality of minor arterial roads is rather poor. Moreover, both distance-based and adjacency-based spatial proximity are tested in developing spatially adjusted models for segments.

6.1 Data Set for Analysis

Similarly to the intersection data, original segment related data was received from NDOT. Data preparation was conducted using the ArcGIS software. Major issues found during the data processing include:

- Data quality needs to be assessed before processing. For example, several identical segments (with the same route master ID and mile posts) were found in the original database. This creates noises in the data set and creates errors when developing models especially the spatially filtered models.
• The HPMS database does not provide enough samples covering the study region. Therefore, some variables cannot be included in the model as explanatory variables because of the lack of data.

• Rather homogeneous segments were created in the HPMS database by NDOT engineers. However, small segments with lengths less than 0.1 mile account for approximately 40% in the database. According to the HSM, segments shorter than 0.1 mile will significantly impact model fitting. Thus, segment aggregation is necessary before data processing.

The linear referencing tool in ArcGIS was used to create segment event layers and afterwards, layers were spatially joint and related to integrate available segment data to create a database for segment analysis. Segments with Functional Class of 3 and 4, representing principal arterial and minor arterial roads respectively are included in the database. The basic statistics of the segment data is summarized in Table 6.1. The database contains a total of 641 urban segments (218 miles).

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Description</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>STDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{seg}$</td>
<td>5-year crash counts</td>
<td>0</td>
<td>554</td>
<td>33.5335</td>
<td>54.7049</td>
</tr>
<tr>
<td>$N_{seg}$</td>
<td>Averaged 5-year crashes</td>
<td>0</td>
<td>111</td>
<td>6.7067</td>
<td>10.9410</td>
</tr>
<tr>
<td>$T_i$</td>
<td>Segment AADT</td>
<td>100</td>
<td>145200</td>
<td>17590</td>
<td>19764</td>
</tr>
<tr>
<td>$L$</td>
<td>Segment length</td>
<td>0.1000</td>
<td>8.4167</td>
<td>0.3410</td>
<td>0.6792</td>
</tr>
<tr>
<td>$N_{tlane}$</td>
<td>Number of through lanes</td>
<td>2</td>
<td>7</td>
<td>3.7036</td>
<td>1.2448</td>
</tr>
<tr>
<td>$M$</td>
<td>Median width</td>
<td>0</td>
<td>75.64</td>
<td>8.0841</td>
<td>12.6916</td>
</tr>
<tr>
<td>$X_{mid}$</td>
<td>Latitude of midpoint</td>
<td>39.3952</td>
<td>39.898</td>
<td>39.5412</td>
<td>0.0620</td>
</tr>
<tr>
<td>$Y_{mid}$</td>
<td>Longitude of midpoint</td>
<td>-120</td>
<td>-119.578</td>
<td>-119.7968</td>
<td>0.0605</td>
</tr>
<tr>
<td>$X_{start}$</td>
<td>Latitude of start point</td>
<td>39.3918</td>
<td>39.8740</td>
<td>39.5404</td>
<td>0.0585</td>
</tr>
<tr>
<td>$Y_{start}$</td>
<td>Longitude of start point</td>
<td>-120.0001</td>
<td>-119.6269</td>
<td>-119.7979</td>
<td>0.0591</td>
</tr>
<tr>
<td>$X_{end}$</td>
<td>Latitude of end point</td>
<td>39.3961</td>
<td>39.8856</td>
<td>39.5421</td>
<td>0.0595</td>
</tr>
<tr>
<td>$Y_{end}$</td>
<td>Longitude of end point</td>
<td>-120.0001</td>
<td>-119.5978</td>
<td>-119.7965</td>
<td>0.0594</td>
</tr>
</tbody>
</table>

Table 6.1: Summary Statistics of Roadway Segment Data
6.2 Model Development

6.2.1 Model Assumptions and Forms

Two model forms are compared: Model Form I: \( \mu_i = \exp(\beta_0 + \beta_1 T_i + \sum \beta_i x_i) \) and Model Form II: \( \mu_i = \exp(\beta_0 + \sum \beta_i x_i) T_i^{\beta_1} L_i^{\beta_2} \). Similar to intersection analyses, over-dispersion is detected in the data set. The Poisson-Lognormal model performs slightly better for segment data than the Poisson-Gamma model and the Poisson-Inverse Gaussian model. One of the reasons could be that the Poisson-Lognormal has higher tail so that it could account for more randomness in this data set. The NB model effectively account for the over-dispersion within the data. Based on the comparison of AIC values, the NB model with form II is selected for further analysis.

<table>
<thead>
<tr>
<th></th>
<th>NB GLM</th>
<th>PLN GLM</th>
<th>PIG GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC Model Form I</td>
<td>5391.8</td>
<td>5385.0</td>
<td>5388.0</td>
</tr>
<tr>
<td>AIC Model Form II</td>
<td>5222.7</td>
<td>5235.6</td>
<td>5240.1</td>
</tr>
<tr>
<td>MAD Model Form I</td>
<td>39.6013</td>
<td>28.8787</td>
<td>26.3596</td>
</tr>
<tr>
<td>MAD Model Form II</td>
<td>23.04</td>
<td>21.7833</td>
<td>22.2946</td>
</tr>
<tr>
<td>MSPE Model Form I</td>
<td>18705.69</td>
<td>1159.02</td>
<td>3274.614</td>
</tr>
<tr>
<td>MSPE Model Form II</td>
<td>2335.0</td>
<td>1911.8</td>
<td>2074.594</td>
</tr>
<tr>
<td>AD Model Form I</td>
<td>1.1186</td>
<td>8.4613</td>
<td>2.4664</td>
</tr>
<tr>
<td>AD Model Form II</td>
<td>1.0953</td>
<td>8.2262</td>
<td>2.2332</td>
</tr>
</tbody>
</table>

6.2.2 Spatially Adjusted Models

Aguero-Valverde and Jovanis found that only first-order neighboring structure (i.e., directly connected) of roadway segments displayed significant contribution to the spatial models [4]. The same concept was applied by El-Basyouny and Sayed [29].
These two studies both used the adjacency-based spatial relation and binary codes (0 and 1) to create spatial weight matrices.

The spatial weight matrix created based on the inverse distance of segment midpoint (calculated using $X_{mid}$ and $Y_{mid}$) is first investigated in this section. However, the spatially adjusted model failed to converge. Then, a binary weight matrix is developed. The spatial weight of segment $i$ and $j$ is assigned 1 if they share the same starting or ending point. Otherwise, the weight is assigned as 0. In other words, the spatial weight between site $i$ and site $j$ is unity if one of the following scenarios occurs.

1. $X_{starti} = X_{startj} \& \ Y_{starti} = Y_{startj}$
2. $X_{starti} = X_{endi} \& \ Y_{starti} = Y_{endi}$
3. $X_{endi} = X_{endi} \& \ Y_{endi} = Y_{endi}$
4. $X_{endi} = X_{startj} \& \ Y_{endi} = Y_{startj}$

The spatially adjusted model is in the form of following.

$$\mu_i = \exp(\beta X_0 + \beta X_1 x_{Ntailc} + \beta E7 EV_7 + \beta E11 EV_{11} + \beta E12 EV_{12} + \beta E4 EV_4 + \beta E8 EV_8 + \beta E16 EV_{16} + \beta E9 EV_5 + \beta E6 EV_6 + \beta E19 EV_{19} + \beta E13 EV_{13} + \beta E9 EV_8) L_i^{x_2} I_i^{x_3}$$

\(6.1\)

The estimated parameters are summarized in Table 6.3. Model (6.1) can also analyze urban roadway segments if corridor boundaries have not been predetermined.

To account for the corridor effects, the spatial weight is defined as follows: segments within the same corridor have mutual weights of 1, and 0 with segments outside
Table 6.3: Parameter Estimates of Spatially Adjusted Model I for Roadway Segments

| Parameter (Model Variable) | Coef. Estimate | Std.Error | z value | Pr(>|z|) |
|----------------------------|----------------|-----------|---------|---------|
| $\beta_{X0}$ (Intercept)  | -5.555502      | 0.480439  | -11.563 | <2e-16  |
| $\beta_{X1}$ ($N_{lane}$) | 0.178620       | 0.040169  | 4.447   | 8.72e-06|
| $\beta_{X2}$ ($T_i$)      | 0.907576       | 0.058383  | 15.545  | <2e-16  |
| $\beta_{X3}$ ($L_i$)      | 0.406834       | 0.044138  | 9.217   | <2e-16  |
| $\beta_{E7}$ ($EV_7$)     | 5.900006       | 0.826368  | 7.140   | 9.35e-13|
| $\beta_{E11}$ ($EV_{11}$) | 3.664229       | 0.820455  | 4.466   | 7.97e-06|
| $\beta_{E12}$ ($EV_{12}$) | 2.440332       | 0.851557  | 2.866   | 0.00416 |
| $\beta_{E4}$ ($EV_{4}$)   | 2.210843       | 0.873658  | 2.531   | 0.01139 |
| $\beta_{E9}$ ($EV_{9}$)   | -2.136905      | 0.843849  | -2.532  | 0.01133 |
| $\beta_{E16}$ ($EV_{16}$) | -1.730224      | 0.787300  | -2.198  | 0.02797 |
| $\beta_{E5}$ ($EV_{5}$)   | -1.562599      | 0.822557  | -1.900  | 0.05747 |
| $\beta_{E6}$ ($EV_{6}$)   | 1.320499       | 0.780555  | 1.692   | 0.09069 |
| $\beta_{E19}$ ($EV_{19}$) | -1.640002      | 0.863520  | -1.899  | 0.05754 |
| $\beta_{E13}$ ($EV_{13}$) | 1.318381       | 0.808474  | 1.631   | 0.10002 |
| $\beta_{E8}$ ($EV_{8}$)   | -1.829000      | 0.852366  | -1.388  | 0.10065 |

Residual deviance 686.87
Over-dispersion parameter 0.5438
Adjusted Deviance 1.0980
MAD 5.1836
MSPE 343.84
AIC 5139.4

of the corridor. The corresponding model is in the form of following.

$$
\mu_i = \exp(\beta_{X0} + \beta_{X1}x_{N_{lane}} + \beta_{X2}x_M + \beta_{E12}EV_{12} + \beta_{E11}EV_{11} + \beta_{E24}EV_{24} + \beta_{E23}EV_{23} + \beta_{E4}EV_{4} + \beta_{E7}EV_{7} + \beta_{E22}EV_{22} + \beta_{E25}EV_{25} + \beta_{E19}EV_{19} + \beta_{E17}EV_{17} + \beta_{E15}EV_{15})T_i^{\beta_{X3}}L_i^{\beta_{X4}}
$$

(6.2)

No noticeable changes were observed between parameter estimates of key explanatory variables such as segment AADT (0.9075 in model I and 0.8443 in model II) and segment lengths (0.4068 in model I and 0.4041 in model II). Nonetheless, significant changes were observed in terms of the model goodness-of-fit statistics (AIC from 5139.4 to 5096.7). Apparently the spatial effect portion in model II accounts for more
randomness than model I. In model I, the spatial relations between segments that are not directly connected are ignored. In model II, the relations between segments within the same corridor are emphasized and the relations between segments and sites far away are ignored. Considering the overall model fitting, model II is recommended to model roadway segments along corridors. Corresponding parameter estimates are presented in Table 6.4.

Table 6.4: Parameter Estimates of Spatially Adjusted Model II for Segments within Corridors

| Parameter (Model Variable) | Coef. Estimate | Std.Error | z value | Pr(>|z|) |
|----------------------------|----------------|-----------|---------|----------|
| $\beta_{X0}$ (Intercept)   | -5.075979      | 0.462575  | -10.973 | <2e-16   |
| $\beta_{X1}$ ($N_{t lane}$) | 0.216802       | 0.037801  | 5.735   | 9.73e-09 |
| $\beta_{X2}$ ($M$)         | -0.007315      | 0.002768  | -2.643  | 0.008224 |
| $\beta_{X3}$ ($T_i$)       | 0.844287       | 0.055879  | 15.109  | <2e-16   |
| $\beta_{X4}$ ($L_i$)       | 0.404075       | 0.041032  | 9.848   | <2e-16   |
| $\beta_{E12}$ ($EV_{12}$)  | 5.088436       | 0.810803  | 6.276   | 3.48e-10 |
| $\beta_{E11}$ ($EV_{11}$)  | 4.232694       | 0.781447  | 5.416   | 6.08e-08 |
| $\beta_{E24}$ ($EV_{24}$)  | -3.191798      | 0.766107  | -4.166  | 3.10e-05 |
| $\beta_{E23}$ ($EV_{23}$)  | -2.764604      | 0.753242  | -3.670  | 0.000242 |
| $\beta_{E4}$ ($EV_{4}$)    | 2.993178       | 0.827129  | 3.619   | 0.000296 |
| $\beta_{E27}$ ($EV_{27}$)  | -2.770902      | 0.767630  | -3.610  | 0.000307 |
| $\beta_{E7}$ ($EV_{7}$)    | -2.749960      | 0.824048  | -3.337  | 0.000846 |
| $\beta_{E22}$ ($EV_{22}$)  | -2.572774      | 0.764259  | -3.366  | 0.000762 |
| $\beta_{E25}$ ($EV_{25}$)  | 1.746906       | 0.780515  | 2.238   | 0.025212 |
| $\beta_{E19}$ ($EV_{19}$)  | 1.654494       | 0.803271  | 2.060   | 0.039428 |
| $\beta_{E17}$ ($EV_{17}$)  | 1.748646       | 0.812641  | 2.152   | 0.031413 |
| $\beta_{E15}$ ($EV_{15}$)  | -1.375427      | 0.753921  | -1.824  | 0.068097 |

Residual deviance: 683.12
Over-dispersion parameter: 0.5063
Adjusted Deviance: 1.0947
MAD: 2.5099
MSPE: 293.99
AIC: 5096.7
6.3 Chapter Summary

Roadway segments are analyzed separately from intersections on a road network because of different facility characteristics. Vehicle drivers on roadway segments do not face as many potential conflicts as at intersections. Segments on the same corridor usually possess similar roadway characteristics, such as same number of through lanes, median type and width, shoulder type and width, similar traffic patterns (either random or platoons); therefore, they may have similar potential for crash occurrence. The spatially adjusted model considering corridor level spatial weight performs better than the unadjusted ones and is recommended for predicting safety at urban segments.
Chapter 7  Corridor Safety Evaluation

After a corridor is designated, the next step is to develop safety performance measures to identify corridors with the highest potential for safety improvements. This chapter contains the application of the developed models in Reno-Sparks, Nevada. The main purpose is to test whether the proposed safety measurement and procedures could produce sound results to support decision makings.

7.1  Corridor Safety Measurement (CSM)

The safety performance measure represents the degree of risk to roadway users, in terms of expected crash frequency. A Corridor Safety Measure (CSM) for corridor $i$ ($i = 1, 2, ..., n$) that comprises $J$ intersections and $K$ segments is:

$$CSM_i = \frac{\sum_{j=1}^{J} N_{int,j} + \sum_{k=1}^{K} N_{seg,k}}{L_i}$$  \hspace{1cm} (7.1)$$

$$N_{int,j} = w_{int,j} \times \hat{\mu}_j + (1 - w_{int,j}) \times Y_j$$  \hspace{1cm} (7.2)$$

$$w_{int,j} = \frac{1}{1 + (1/\phi_{int,j}) \times \sum \hat{\mu}_j}$$  \hspace{1cm} (7.3)$$

$$N_{seg,k} = w_{seg,k} \times \hat{\mu}_k + (1 - w_{seg,k}) \times Y_k$$  \hspace{1cm} (7.4)$$

$$w_{seg,k} = \frac{1}{1 + (1/\phi_{seg,k}) \times \sum \hat{\mu}_k}$$  \hspace{1cm} (7.5)$$
where, $N_{int_{ij}}$ represents the expected number of crashes at intersection $j$, $N_{seg_{ik}}$ represents the expected number of crashes at segment $k$ and $L_i$ is the length of corridor $i$. The corridor CSM was measured by the combined expected crashes on urban segments and intersections which occurred along specific corridor measured by unit length. $N_{int_{ij}}$ and $N_{seg_{ik}}$ are obtained from (4.44) in which $E(Y_{ij})$ and $E(Y_{ik})$ are the predicted numbers of crashes estimated by the crash prediction models developed in Chapter 5 and 6. The number of crashes expected to occur on a designated corridor is estimated by summing the expected crashes at intersections ($N_{int_i}$) and the crashes on segments ($N_{seg_i}$). For each designated corridor under consideration, the CSM index is estimated and applied either as the ranking tool for evaluating the corridor safety performance or a proactive method for urban safety corridor planning.

### 7.2 Case Demonstration

As presented in Chapter 3, principal arterial and minor arterial roads in Reno-Sparks were divided into 152 urban corridors. The CSM index is calculated for each designated corridor. Figure 7.1 displays the screening results in which the CSM index is divided into five categories with each category indicating a specified index range. Top ranked arterials with potential safety issues include: Kietzke Ln from the intersection of E 2nd Street to S Virginia Street, N McCarran Blvd from the intersection of Keystone Ave towards the interchange at I-80, large portion of S McCarran Blvd and Pyramid Way corridor. These arterial roads experience relatively higher traffic exposures and contains more signalized intersections. Most minor arterial roads are ranked low except E Plumb Ln and Sun Valley Blvd corridors. Hence, it is recommended that transportation agencies focus on principal arterial roads especially when the safety program funding is limited. Engineers should pay more attention to connected corridors with high CSMs such as S McCarran Blvd corridors shown in Figure
Tables 7.1 and 7.2 present the detailed corridor ranking information and associated CSMs based on both observed and expected crashes. The observed crashes and model-based screening provide very different results. Raw data based screening produces relatively high CSMs for sites with low raw crashes, short lengths and low traffic exposures. On the contrary, model-based screening provides results based on long term average number of crashes; therefore, the results are more convincing and reasonable.
Figure 7.1: Corridor Screening Results based on CSMs
Table 7.1: Top Ten High Crash Corridors in Reno-Sparks based on Observed CSMs

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Arterial From</th>
<th>To Arterial</th>
<th>Corridor Length (mi)</th>
<th>Average AADT</th>
<th>Observed Crashes (per year)</th>
<th>Expected Crashes (per year)</th>
<th>CSMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E 2nd St</td>
<td>Kuenzli St</td>
<td>0.5328</td>
<td>3740</td>
<td>56</td>
<td>105.1051</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>W 4th St</td>
<td>Ralston St</td>
<td>0.3597</td>
<td>13200</td>
<td>35</td>
<td>97.3033</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>N Virginia St</td>
<td>W 8th St</td>
<td>0.5762</td>
<td>9333</td>
<td>47</td>
<td>81.5689</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>W 4th St</td>
<td>Stoker Ave</td>
<td>1.0155</td>
<td>10980</td>
<td>76</td>
<td>74.8399</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>E 2nd St</td>
<td>N Virginia St</td>
<td>0.2823</td>
<td>7480</td>
<td>21</td>
<td>74.3889</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>W 4th St</td>
<td>SR659</td>
<td>0.8219</td>
<td>9967</td>
<td>49</td>
<td>59.6178</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Baring Blvd</td>
<td>Sparks Blvd</td>
<td>1.0453</td>
<td>9520</td>
<td>50</td>
<td>47.3322</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>N Virginia St</td>
<td>SR659</td>
<td>0.9195</td>
<td>10020</td>
<td>40</td>
<td>43.5019</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Vista Blvd</td>
<td>S Los Altos Pky</td>
<td>0.4091</td>
<td>15000</td>
<td>17</td>
<td>41.5546</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>N McCarran Blvd</td>
<td>Glendale Ave</td>
<td>0.4265</td>
<td>27400</td>
<td>17</td>
<td>39.8593</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.2: Top Ten High Crash Corridors in Reno-Sparks based on Expected CSMs

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Arterial From</th>
<th>To Arterial</th>
<th>Corridor Length (mi)</th>
<th>Average AADT</th>
<th>Observed Crashes (per year)</th>
<th>Expected Crashes (per year)</th>
<th>CSMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Kietzke Ln</td>
<td>E Plumb Ln</td>
<td>1.7014</td>
<td>22500</td>
<td>125</td>
<td>120</td>
<td>70.5302</td>
</tr>
<tr>
<td>2</td>
<td>Pyramid Way</td>
<td>Los Altos Pkwy</td>
<td>4.2052</td>
<td>42500</td>
<td>177</td>
<td>252</td>
<td>59.9258</td>
</tr>
<tr>
<td>3</td>
<td>N McCarran Blvd</td>
<td>Keystone Ave</td>
<td>2.3141</td>
<td>21500</td>
<td>96</td>
<td>120</td>
<td>51.8560</td>
</tr>
<tr>
<td>4</td>
<td>S McCarran Blvd</td>
<td>Mill St</td>
<td>4.3265</td>
<td>21000</td>
<td>178</td>
<td>220</td>
<td>50.8494</td>
</tr>
<tr>
<td>5</td>
<td>Kietzke Ln</td>
<td>Galletti Way</td>
<td>2.8563</td>
<td>13000</td>
<td>155</td>
<td>145</td>
<td>50.7650</td>
</tr>
<tr>
<td>6</td>
<td>S McCarran Blvd</td>
<td>S Virginia St</td>
<td>3.1055</td>
<td>21500</td>
<td>162</td>
<td>149</td>
<td>47.9794</td>
</tr>
<tr>
<td>7</td>
<td>E Plumb Ln</td>
<td>Virginia St</td>
<td>1.2027</td>
<td>26000</td>
<td>66</td>
<td>57</td>
<td>47.3934</td>
</tr>
<tr>
<td>8</td>
<td>Pyramid Way</td>
<td>N McCarran</td>
<td>2.7100</td>
<td>31500</td>
<td>120</td>
<td>116</td>
<td>42.8044</td>
</tr>
<tr>
<td>9</td>
<td>S McCarran Blvd</td>
<td>Skyline Blvd</td>
<td>3.6520</td>
<td>22500</td>
<td>112</td>
<td>125</td>
<td>34.2278</td>
</tr>
<tr>
<td>10</td>
<td>Sun Valley Blvd</td>
<td>E 7th St</td>
<td>2.0203</td>
<td>28500</td>
<td>85</td>
<td>69</td>
<td>34.1533</td>
</tr>
</tbody>
</table>
7.3 Chapter Summary

In the final phase of this research, the crash prediction models were employed to develop the corridor safety measurement using (7.1). The CSM index is a composite of expected crashes at intersections and roadway segments. The observed CSMs were compared with the expected indices and dramatic difference was observed between two screening lists. The screening based on raw data tends to overestimate sites with low traffic and low number of observed crashes. The developed CSM plays a vital role in quantifying the overall risk of traveling on each designated corridor. Utilizing the proposed CSMs, agencies can make informed decision based on not only budget but also long term safety conditions.
Chapter 8  Summary and Conclusions

Although much research has been undertaken in the traffic safety field, the research on urban arterials, especially from a corridor perspective, is relatively limited. Urban corridors serve a high share of longer distance trips and daily vehicle miles of travel. For this reason, corridors are typically roadways with high traffic volumes and also high incidence of crashes. In this dissertation, efforts have been made to define urban corridor boundaries as well as quantify the long term performance of designated corridors.

8.1  Findings

Major findings and contributions of this research are summarized below.

- The concept of urban corridor is rather subjective. The spacing of corridors is closely related to the trip-end density characteristics of activity centers in urban areas. Transportation facilities close to each other tend to have similar roadway characteristics, traffic patterns and surrounding land-use intensity. Emphasizing this “closeness” may aggregate groups of “similar” sites on a road network. By means of spatial clustering, close signalized intersections with similar features are grouped together such that contiguous sites are treated as one entity. Given the contiguous nature of investigated sites, one can conclude that, the developed method yielded rational groupings of intersections and segments that can be
used, along with other factors, to define projects for safety programming.

- Traffic crashes are random and rare events. High random variability is often observed in real world crash database. Predicted annual number of crashes was estimated via geometric characteristics and traffic conditions. To understand and model the relations between crashes and roadway factors, certain assumptions are necessary. As a matter of course, mixtures of Poisson distributions are selected. The HSM recommends the Poisson-Gamma generalized linear models in developing crash prediction models. Such assumptions are plausible to a certain extent. Nonetheless, different assumptions may result in varied results and should be used judiciously.

- Significant spatial correlations were found in both intersection and segment level. The spatial effects can assist data analysts and engineers to understand and investigate the hidden reasons of traffic crashes. The results from this research proved that a higher proportion of uncertainty, or variability, was explained by the spatial correlation term. The uncertainty and variability may stem from omitted variables, varying degrees of measurement errors over space, mis-specification of functional forms for the mean of the response variable in the model, etc. Noticeable changes in estimated parameters after accounting for the spatial effects also proved the existence of bias associated with model forms and/or variable omission. The additional spatial adjustments increased the model prediction precision by borrowing strengths across various sites and including the possible variability across locations in space.

- Analyzing transportation facilities from a corridor perspective may uncover safety issues not identified through traditional hot spot analysis. The safety performance of the entire designated corridor can be assessed to provide a broader
perspective and to support comparisons with other corridors. The proposed corridor safety measurement index is designed to alert transportation agencies to potential safety issues on corridor level so that preventive countermeasures can be implemented. It could assist agencies in allocating limited funding, enhancing the safety of urban arterials and alleviating traffic congestion. The index can also be employed to urban network analysis so that the safety factor can be incorporated into the transportation planning analysis.

8.2 Future Extensions

The findings from this research indicate many areas of potential future research. The eigenvector spatial filtering method is proved to be an appealing method for accounting for the spatial relationships in analyzing crash data. Nevertheless, the method is still in its developing stage so that efforts can be made to extend the ESF method to the full Bayesian framework. Opportunities for other continued investigation lie in the development of data-driven neighboring structures with the goal to select the configuration of more complex spatial structures that result in the optimal performance of the prediction models. Other spatial techniques should be considered and compared with the proposed scheme.
Bibliography


[68] Shaw-Pin Miaou and Dominique Lord. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and bayes ver-


[77] Jutaek Oh, Craig Lyon, Simon Washington, Bhagwant Persaud, and Joe Bared. Validation of fhwa crash models for rural intersections: Lessons learned. *Trans-


Appendix A  Poisson Inverse-Gaussian Distribution

If $\epsilon_i$ follows an Inverse-Gaussian distribution, the following derivations can be obtained.

$$\gamma = 1; \quad \delta = 1/\tau$$

$$g(\epsilon_i|\gamma, \delta) = \left[ \frac{\delta}{2\pi \epsilon_i^3} \right]^{1/2} \exp \left\{ \frac{-\delta (\epsilon_i - \gamma)^2}{2\gamma^2 \epsilon_i} \right\}$$

Therefore,

$$g(\epsilon_i|1, 1/\tau) = \left[ \frac{1}{2\pi \tau \epsilon_i^3} \right]^{1/2} \exp \left\{ \frac{- (\epsilon_i - 1)^2}{2\tau \epsilon_i} \right\}$$

$$E(\epsilon_i) = 1; \quad Var(\epsilon_i) = \tau$$

The marginal distribution of $Y_i$ is derived as:

$$P(Y_i = y_i|\theta_i, \tau) = \int_0^\infty f(y_i|\theta_i, \epsilon_i)g(\epsilon_i)d\epsilon_i = \int_0^\infty \frac{(\theta_i \epsilon_i)^y e^{-\theta_i \epsilon_i}}{y!} \left[ \frac{1}{2\pi \tau \epsilon_i^3} \right]^{1/2} \exp \left\{ \frac{-(\epsilon_i - 1)^2}{2\tau \epsilon_i} \right\} d\epsilon_i$$

$$= \frac{\theta_i^y e^{1/\tau}}{y! (2\pi \tau)^{1/2}} \int_0^\infty \epsilon_i^{-3/2} e^{-\theta_i \epsilon_i} \left[ e^{\frac{1}{2\tau \epsilon_i}} \right] d\epsilon_i$$

The integral is of the form [36]:

$$\int_0^\infty e^{ax^h} \cdot x^{s-1} dx = \left( \frac{2}{h} \right) \left( \frac{b}{a} \right)^{s/2h} K_{s/h}(2\sqrt{ab})$$
where, $K_j(.)$ is the modified Bessel function of the second kind. For the Poisson-
Inverse Gaussian derivatives,

$$h = 1; \ a = -\left(\theta_i + \frac{1}{2\tau}\right); \ b = \frac{1}{2\tau}; \ s = y_i - \frac{1}{2}$$

Therefore, we have

$$P(Y_i = y_i|\theta_i, \tau) = \frac{\theta_i^{y_i} e^{1/\tau}}{y_i!(2\pi\tau)^{1/2}} \int_0^\infty e^{y_i-3/2} e^{-c_i(\theta_i + \frac{1}{2\tau}) - \frac{1}{2\tau} e_i} d\epsilon_i$$

$$= \frac{2\theta_i^{y_i} e^{1/\tau}}{y_i!(2\pi\tau)^{1/2}} \left(\frac{1}{2\tau \left(-\left(\theta_i + \frac{1}{2\tau}\right)\right)}\right)^{y_i-1/2} K_s \left(2\sqrt{\frac{1}{2\tau \left(-\left(\theta_i + \frac{1}{2\tau}\right)\right)}}\right)$$

Define,

$$c(\tau) = \left(\frac{2}{\pi\tau}\right)^{1/2} e^{\frac{1}{2\tau}} \text{ and } h(y) = c(\tau)(1 + 2\tau\theta) - y^{\frac{1}{2}} K_{y-1} \left(\frac{\sqrt{1 + 2\tau\theta}}{\tau}\right)$$

It can be shown that,

$$h(y + 1) = c(\tau)(1 + 2\tau\theta)^{\frac{2y+1}{4}} K_{y+\frac{1}{2}} \left(\frac{\sqrt{1 + 2\tau\theta}}{\tau}\right)$$

The Bessel functions have the following important properties that can be used to
derive the iteration functions.

$$K_{\frac{1}{2}}(a) = K_{-\frac{1}{2}}(a) = \left(\frac{\pi}{2a}\right)^{1/2} e^{-a}$$

$$K_{\frac{3}{2}}(a) = \left(1 + \frac{1}{a}\right) K_{\frac{1}{2}}(a)$$

Therefore, we obtain
\[
\frac{K_{y \frac{1}{2}}(a)}{K_{y - \frac{1}{2}}(a)} = 1 + \frac{1}{a} = 1 + \frac{\tau}{\sqrt{1 + 2\tau \theta}}
\]

\[
\frac{K_{y \frac{1}{2}}(a)}{K_{y - \frac{1}{2}}(a)} = \frac{h(y + 1) c(\tau) (1 + 2\tau \theta)^{-\frac{y}{2}}} {h(y) c(\tau) (1 + 2\tau \theta)^{-\frac{y+1}{2}}} = (1 + 2\tau \theta)^{\frac{1}{2}} M(y)
\]

\[
M(y) = \frac{h(y + 1)}{h(y)} \quad \text{and} \quad M(0) = (1 + 2\tau \theta)^{\frac{1}{2}}
\]

\[
h(0) = \left(\frac{2}{\pi \tau}\right)^{1/2} e\frac{\tau}{2} (1 + 2\tau \theta)^{-\frac{1}{2}} K_{-\frac{1}{2}} \left(\frac{\sqrt{1 + 2\tau \theta}}{\tau}\right)
\]

\[
= \left(\frac{2}{\pi \tau}\right)^{1/2} e\frac{\tau}{2} (1 + 2\tau \theta)^{-\frac{1}{2}} \left(\frac{\pi \tau}{2\sqrt{1 + 2\tau \theta}}\right)^{\frac{1}{2}} e^{-\frac{\sqrt{1 + 2\tau \theta}}{\tau}} = e^{\frac{\tau}{2} (1 - \sqrt{1 + 2\tau \theta})}
\]

\[
h(1) = M(0) h(0) = (1 + 2\tau \theta)^{-\frac{1}{2}} h(0)
\]

Thereby, the probabilities of \(y_i\) can be calculated recursively as:

\[
P(Y_i = y_i | \theta_i, \tau) = \frac{\theta_i^{y_i}}{y_i!} h(y)
\]

\[
P(0 | \theta_i, \tau) = e^{\frac{\tau}{2} (1 - \sqrt{1 + 2\tau \theta})}
\]

\[
P(1 | \theta_i, \tau) = \theta_i (1 + 2\tau \theta)^{-1/2} P(0 | \theta_i, \tau)
\]

\[
P(y_i | \theta_i, \tau) = \frac{2\tau \theta}{1 + 2\tau \theta} \left(1 + \frac{3}{2y}\right) P(y_i - 1 | \theta_i, \tau) + \frac{\theta_i^2}{1 + 2\tau \theta y(y - 1)} P(y_i - 2 | \theta_i, \tau), \quad (y = 2, 3, \ldots)
\]

The mean and variance of P-IG distribution are:

\[
E(Y_i) = E(E(Y_i | \theta_i, \epsilon_i)) = E(\theta_i, \epsilon_i) = \theta_i
\]

\[
Var(Y_i) = Var(E(Y_i | \theta_i, \epsilon_i)) + E(Var(Y_i | \theta_i, \epsilon_i)) = \theta_i + \theta_i^2
\]

The maximum likelihood functions for the P-IG distribution are derived as follows:
\[ L(y_1, y_2, \ldots, y_n | \theta_i, \tau) = \prod_{i=1}^{n} P(y_i | \theta_i, \tau) = \prod_{i=1}^{n} \frac{\theta_i^{y_i}}{y_i!} \left( \frac{2}{\pi \tau} \right)^{1/2} e^{\tau} \left( 1 + 2\tau \theta_i \right)^{-y_i - 1/2} K_{y_i - 1/2} \left( \sqrt{1 + 2\tau \theta_i} \right) \]

The partial derivatives of the log-likelihood function with respect to the model parameters do not depend on the Bessel functions, rather they depend on \( M(y) \) which is recursively computed.
Appendix B  R Codes

# Read data
summary(Kietzke=read.csv("Kietzke.txt", header=TRUE, row.names=1))
summary(FOURSG=read.table(file='4SG.txt', header=TRUE,sep=','))
summary(FOURST=read.table(file='4ST.txt', header=TRUE,sep=','))
summary(THREESG=read.table(file='3SG.txt', header=TRUE,sep=','))
summary(THREEST=read.table(file='3ST.txt', header=TRUE,sep=','))
summary(SEG=read.table(file='SEG.txt', header=TRUE,sep=','))

# Generate spatial weight matrix for Kietzke Ln example
Kietzke.dists <- as.matrix(dist(cbind(Kietzke$Lon, Kietzke$Lat)))
Kietzke.dists.inv <- 1/ Kietzke.dists
diag(Kietzke.dists.inv) <- 0

# Calculate spatial autocorrelation parameter Moran’s I
Moran.I(Kietzke$variable, Kietzke.dists.inv)

# Cluster signalized intersections
# Elbow method
k=100
se=rep(0,k)
wss <- (nrow(Kietzke)-1)*sum(apply(Kietzke,2,var))
for (i in 2:k)wss[i] <- sum(kmeans(Kietzke, centers=i)$withinss)
plot(1:k, wss, type='b", xlab='Number of Clusters”, ylab='Within clusters sum of
squares")
# Gap statistic
k=100
Gap=rep(0,k)
se=rep(0,k)
for (i in 2:k)
{
    km=kmeans(Kietzke, centers=i)
    mem=km$cluster
    result=gap(Kietzke, class=mem)
Gap[i]=result[1]
se[i]=result[2]

plot(1:k, Gap, type='b”, xlab='Number of Clusters”)

# Cluster results
Kietzkecorridor <- kmeans(Kietzke, 3, 30)  #repeat the clustering process for 30 times
aggregate(Kietzke, by=list(Kietzkecorridor$cluster), FUN=mean)  #get cluster means
Kietzkecorridorresults <- data.frame(Kietzke, Kietzkecorridor$cluster)  #append corridor assignment

# Poisson GLM for 4SG  #same code frame is ued for other facility types
summary(poisson.glm1 <- glm(Crash.counts ~ AADTma + AADTmi + NAWLTL + NAWRT + PL, family='poisson', data= FOURSG))
summary(poisson.glm2 <- glm(Crash.counts ~ AADTma + AADTmi + NAWLTL, family='poisson', data= FOURSG))
summary(poisson.glm3 <- glm(Crash.counts ~ lnAADTma + lnAADTmi + NAWLTL, family='poisson', data= FOURSG))
summary(poisson.glm4 <- glm(Crash.counts ~ AADTma + lnAADTma + AADTmi + lnAADTmi + NAWLTL + FR, family='poisson', data= FOURSG))

# Check model goodness-of-fit for 4SG
with(poisson.glmX, cbind(res.deviance = deviance, df = df.residual, p = pchisq(deviance, df.residual, lower.tail=FALSE)))  #X=1,2,3,4
# Dispersion test for 4SG
dispersiontest(poisson.glmX, trafo=1)

# NB GLM for 4SG  #same code frame is ued for other facility types
summary(nb.glm1 <- glm.nb(Crash.counts ~ AADTma + AADTmi + NAWLTL + NAWRT + PL, data= FOURSG))
summary(nb.glm2 <- glm.nb(Crash.counts ~ AADTma + AADTmi + NAWLTL, data= FOURSG))
summary(nb.glm3 <- glm.nb(Crash.counts ~ lnAADTma + lnAADTmi + NAWLTL, data= FOURSG))
summary(nb.glm4 <- glm.nb(Crash.counts ~ AADTma + lnAADTma + AADTmi + lnAADTmi + NAWLTL + FR, data= FOURSG))

# Check model goodness-of-fit for 4SG
pchisq(summary(nb.glm1)$deviance, summary(nb.glm1)$df.residual)  #X=1,2,3,4

# PLN GLM for 4SG  #same code frame is ued for other facility types
# Define population size and parameter values for Lognormal
n=80
a=0.5
b=0.5
mu.pop=0
sigma.pop=1.0
# Set up data frame for the maximum likelihood routine and provide an initial estimate for mu and sigma
data.int=data.frame(FOURSG$Crash.counts, FOURSG$AADTma, FOURSG$AADTmi, FOURSG$NAWLTL, FOURSG$NAWRT, FOURSG$PL)
para=c(0,1.0,1,1,1,1,1,1)
# Load a function used to evaluate the likelihood
PoiNORML<-function(para, FOURSG)
{
    mu=para[1]*log(10)
    sigma=para[2]*log(10)
    a=para[3]; b=para[4]; c=para[5]; d=para[6]; e=para[7]; f=para[8]
    return(-sum(log(dpoilog(FOURSG$Crash.counts-a-(b*FOURSG$AADTma+
    c*FOURSG$AADTmi+d*FOURSG$NAWLTL+
    e*FOURSG$NAWRT+f*FOURSG$PL),mu,sigma))))
}
# end function
# Maximize the log likelihood
res<- optim(para, PoiNORML, FOURSG=data.int)

# PIG GLM for 4SG  %same code frame is ued for other facility types
summary(pig.glm1<- gamlss (Crash.counts ~ AADTma+ AADTmi+ NAWLTL+NAWRT +PL, family=PIG, data= FOURSG))
summary(pig.glm2<- gamlss (Crash.counts ~ AADTma+ AADTmi+ NAWLTL, family=PIG, data= FOURSG))
summary(pig.glm3<- gamlss (Crash.counts ~ lnAADTma+lnAADTmi+ NAWLTL, family=PIG, data= FOURSG))
summary(pig.glm4<- gamlss (Crash.counts ~ AADTma+ lnAADTma+ AADTmi+lnAADTmi+ NAWLTL+FR, family=PIG, data= FOURSG))

# ESF adjusted models for 4SG  %same coding frame is ued for other facility types
# Initial spatial autocorrelation tests
coords <- coordinates(FOURSG[19:20])
k0 <- knearest(coords, k=80)
k1 <- km2nb(k0)
plot(FOURSG[19:20], xlab = “Intersection Latitude”, ylab = “Intersection Longitude ”)
plot(k1, coords, add=TRUE)
all.linked <- max(unlist(nbdists(k1, coords)))
nb<- dnearest(as.matrix(FOURSG[19:20]), 0, all.linked)
summary(nb, coords)
dsts <- nbdist(nb, coords)
idw <- lapply(dsts, function(x) 1/(x))
w <- nb2listw(nb, glist=idw, style="W", zero.policy=FALSE)  %style="B" when binary weight is applied
col.mat <- listw2mat(w)
all.equal(col.mat, t(col.mat), check.attributes=FALSE)
moran.test(nb.glm4$fitted, w)
moran.test(nb.glm4$residuals, w)

# Corridor based spatial weight matrix
customizedw=read.table(file= 'customizedw.txt',header=TRUE,sep=',')
wcorridor<-as.matrix(customizedw[2:132])
w1<-mat2listw(wcorridor)
col.mat1<- listw2mat(w1)
moran.test(nb.glm4$fitted, w1)
moran.test(nb.glm4$residuals, w1)

# Calculate spatial eigenvectors
n <- length(nb)
M <- diag(n) - matrix(1,n,n)/n
B <- listw2mat(w)  %w1 is used for corridor analysis
MBM <- M %*% B %*% M
eig <- eigen(MBM, symmetric=T)
EV <- as.data.frame( eig$vectors[, eig$values/eig$values[1]>0.25])
colnames(EV) <- paste("EV",1:NCOL(EV), sep="")

# Run step-wise models with stepAIC
nb.full <- glm.nb(FOURSG$Crash.counts ~ FOURSG$AADTma+ FOURSG$lnAADTma+ FOURSG$AADTmi+FOURSG$lnAADTmi+ FOURSG$NAWLTL+FOURSG$FR+, data=EV)
nb.sf <- stepAIC(glm.nb(FOURSG$Crash.counts ~ FOURSG$AADTma+ FOURSG$lnAADTma+FOURSG$AADTmi+FOURSG$lnAADTmi+FOURSG$NAWLTL+ FOURSG$FR, data=EV), scope=list(upper=nb.full), direction="forward")

# Summarize the spatially adjusted model
summary(nb.sf <- glm.nb(FOURSG$Crash.counts ~ FOURSG$AADTma+ FOURSG$lnAADTma+ FOURSG$AADTmi+ FOURSG$lnAADTmi+ FOURSG$NAWLTL+ FOURSG$FR + EV5 + EV15 + EV12 + EV6 + EV16 + EV14 + EV9 + EV8, data=EV))

# Examine model fit
summary(glm.sf.bt <- lm(FOURSG$Crash.counts ~I(nb.sf$fitted)))
# Examine the residuals of the spatial filter model
nb.sf.res <- round(residuals(nb.sf, type="response"))
# Conduct spatial autocorrelation tests
moran.test(nb.sf.res, w)