2018

Three essays on crash frequency analysis

Chenhui Liu
Iowa State University

Three essays on crash frequency analysis

by

Chenhui Liu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Civil Engineering (Transportation Engineering); Statistics

Program of Study Committee:
Anuj Sharma, Co-major Professor
Lily Wang, Co-major Professor
Jing Dong
Peter Savolainen
Jarad Niemi

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

To my family, my home, and my dream.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

NOMENCLATURE

AADT	Annual Average Daily Traffic
BYM	Besag-York-Mollie
CAR	Conditional Autoregressive
CBD	Central Business District
DIC	Deviance Information Criterion
ICAR	Intrinsic Conditional Autoregressive
INLA	Integrated Nested Laplace Approximation
MBYM	Multivariate Besga-York-Mollie
MVN	Multivariate Normal
MVNB	Multivariate Negative Binomial
MVP	Multivariate Poisson
MVPLN	Multivariate Poisson Log-Normal
MVRPZINB	Multivariate Random Parameters Zero-Inflated Negative Binomial
MVRPZIP	Multivariate Random Parameters Zero-Inflated Poisson
MVZINB	Multivariate Zero-Inflated Negative Binomial
MVZIP	Multivariate Zero-Inflated Poisson
NB	Negative Binomial
NFC	National Functional Classification
RW	Random Walk
MRW	Multivariate Random Walk

URPZINB	Univariate Random Parameters Zero-Inflated Negative Binomial
URPZIP	Univariate Random Parameters Zero-Inflated Poisson
VMT	Vehicle Miles Travelled
ZIP	Zero-Inflated Poisson

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my major professors, Dr. Anuj Sharma and Dr. Lily Wang, for their insightful guidance, warm encouragement, and generous support throughout the research and my graduate studies. Dr. Sharma always embraces new ideas and encourages me to explore new transportation areas. Dr. Wang opens the door of statistics to me. I would also like to give great thanks to my committee members, Dr. Dong, Dr. Savolainen, and Dr. Niemi, for their guidance and support throughout the course of this research.

In addition, I would like to acknowledge the Iowa Department of Transportation and the Nebraska Department of Roads for providing the data for the research. My thanks go out to all the friends and colleagues in Nebraska and Iowa for providing a friendly, comfortable, and productive environment during my graduate studies. Without their support and encouragement, I would not have finished this study.

Lastly, I would like to thank my parents and grandparents for their selfless love and support to my study at home and abroad. I would like to give special thanks to my older brother for taking care of our family during my study in the United States. Your support is the driving force for me to move ahead.

ABSTRACT

Road crashes have been one of the major leading causes of deaths and injuries in the United States, and also bring huge financial expenses. The general theme of this dissertation is to use advanced statistical models to better understand the characteristics of crash frequency in Iowa and Nebraska, and identify important factors influencing crash frequency. It is expected that the findings of these studies could be utilized in safety improvement programs to improve traffic safety in future.

This dissertation includes three published essays. The first essay explores the spatio-temporal effects in traffic crash trend analysis under univariate cases at the macro level, where spatial and temporal effects are found to be essential in crash frequency analysis. The second essay extends the univariate spatio-temporal models into multivariate crash data, where multivariate spatio-temporal models are proved to be necessary in multivariate crash frequency analysis. The third essay examines the effects of traffic operational and roadway geometric factors on three kinds of crash types on urban midblock segments at the micro level, where segment-specific effects of these factors are revealed.

CHAPTER 1. GENERAL INTRODUCTION

Traffic safety has been a concern of this planet since the invention of vehicles. Motor crashes have been one of the major sources of fatalities and injuries in the United Stated (US). After multiple years' decline, the fatalities increased in 2015 in US. 35,092 people died on highways in motor vehicle traffic crashes in 2015 in US, a 7.2% increase than in 2014. That is, nearly 100 people die from vehicle related accidents every day (White House, 2016). Traffic safety studies generally can be divided into two categories in terms of research objects: 1) Microscopic level: focus on analyzing individual crashes, usually involving in analyzing crash occurrence or not, or crash severity; 2) Macroscopic level: focus on analyzing crash frequency data, which are usually obtained by summarizing individual crash data over space and time. The microscopic crash data are more informative than macroscopic crash data because they provide more information to be utilized for crash analysis (Savolainen et al., 2011). However, due to privacy concern, incomplete data record, and other reasons, individual crash data are often unavailable. Transportation agencies usually only offer crash frequency data of their jurisdictions by month or year publicly. That is, in many cases, only crash frequency data are available in traffic safety study. The Fatality Analysis Reporting System (FARS) provides individual fatal crash data of US from 1994, but not other crash data. Additionally, crash frequency modeling could provide very insightful results on the effects of macroscopic factors, such as policy, law, weather, economy, infrastructure construction, or highway improvement programs, on traffic safety. Thus, crash frequency analysis plays a critical role in traffic safety study.

Crash frequency data are non-negative integers, thus they are often analyzed with the Poisson model (Lord and Mannering, 2010). In addition, they often have some unique features to be considered in analysis.

- Over-dispersion/Under-dispersion

Crash frequency data are often over-dispersed, i.e. the variance being larger than the mean, where the Quasi-Poisson, negative binomial/Poisson-gamma model, and Poisson-lognormal model are often used (Lord and Mannering, 2010). Lord and Mannering (2010) also discussed the under-dispersion cases, which were much less common for crash frequency data.

- Zero inflation

Many crash data are often zero-inflated, i.e. the proportions of zero crashes are larger than what it is supposed to be under assumed distributions. Zero inflation is more common for severe crashes, such as fatal crashes. Thus, the zero-inflated/hurdle count data models are often adopted (Lord and Mannering, 2010; Mannering and Bhat, 2014).

- Crash-Between Correlation

Traffic crashes can be divided into multiple classes by different criteria, such as injury severity, collision manner, victim, wrecker type, and other indicators (Lord and Mannering, 2010). For example, crashes are often classified into five categories by injury severity: fatal (K), incapacitating injury (A), non-incapacitating injury (B), possible injury (C), and no injury (O), i.e. the "KABCO" injury scale (AASHTO, 2009). These different injury severity crashes may have some correlations (Lord and Mannering, 2010; Savolainen et al., 2011). It is understandable that the locations with many fatal crashes also very likely have many injury crashes. When multiple crashes are analyzed at the same time, the multivariate count data models may be more desired than univariate ones, since the multivariate analysis could borrow information from each other component to get more accurate prediction and estimation results (Savolainen et al., 2011).

- Spatial Correlation

Crash frequency data are always presented with location tags, such as intersection, segment, city, county, state, and so on. Tobler's first law of geography says that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). It also applies to traffic crashes. For example, two adjacent intersections on the same arterial may share some common crash features, since they have similar traffic characteristics. Two adjacent counties may share some common crash features, since they may have similar terrain, weather, culture, population, economy, and laws. Thus, the spatial correlation of traffic crashes may not be ignored.

- Temporal Correlation

Similar to the spatial correlation, crash frequency data are also presented with time tags, such as hour, day, week, month, quarter, or year. Thus, they may also show some time series correlation in long term. For example, Iowa always has more crashes in winter months than in summer months. Thus, the temporal correlation should also be considered in crash frequency modeling.

Generally speaking, crash frequency data are often presented as multivariate spatio-temporal count data, with possible over-dispersion and zero-inflation, which should be considered in crash frequency modeling. Ignoring these correlations may produce biased and less-efficient estimation results (Savolainen et al., 2011).

This dissertation includes three essays. Chapter 2 starts with the analysis of yearly county-level fatal crash frequencies of Iowa from 2006 to 2015. Multiple Bayesian spatio-temporal Poisson models are built to account for possible correlations of crashes over space and time. The integrated nested Laplace approximation (INLA) is introduced to estimate these

Bayesian models. Both spatial effects and temporal effects are found to be essential for crash frequency analysis, while the spatial effects play a more important role than the temporal effects for this case. The counties in the central north and south of Iowa are found to tend to have fewer crashes than other counties in space. Multiple temporal models, including the 1st order random walk (RW1) model, the 1st order autoregressive (AR1) model, and the linear temporal model, are compared. The linear temporal model is found to be superior to other models. Fatal crashes are found to show a decreasing trend in Iowa but with varying decreasing rates by counties. In addition, it is found that spatial and temporal effects could take zero inflation and over dispersion of crashes into account well, and makes it no more need of zero-inflated models.

Chapter 3 extends the univariate spatio-temporal analysis into multivariate cases, where the yearly county-level fatal crashes, major injury crashes, and minor injury crashes of Iowa from 2006 to 2015 are analyzed simultaneously. It is found that the multivariate spatio-temporal model has a greater performance than the univariate ones. Income and weather are found to have insignificant effects on these crashes in long term, while the unemployment rate is found to have significant negative effects on major injury and minor injury crashes. Significantly spatial correlations are found to exist both for each crash type and across crash types, where the counties in the central north and south of Iowa tend to have fewer crashes than other counties in space. Each crash type generally shows some decreasing trends over time. However, their temporal correlations across crash types are found to be insignificant. In addition, all these crashes are found to be positively correlated to each other, but major injury crashes and minor injury crashes show a closer relationship than fatal crashes. Based on the estimated results, all the counties are ranked by the crash cost rates with the posterior expected rank to identify high-risk counties.

Chapter 2 and Chapter 3 focus on analysis of crash frequency by jurisdictions at the macro level, while Chapter 4 analyzes the crash frequency by segments at the micro level. Using the yearly sideswipe (same direction), rear-end, and other crash frequency data of 1506 segments in Lincoln and Omaha from 2003 to 2012. Traffic operation and roadway geometry characteristics were investigated to identify significant influencing factors. Due to the concern of unobserved heterogeneity produced by correlations across crash types and segments, excess zeros, and over dispersion in crash data, the multivariate zero-inflated negative binomial regression (MVRPZINB) model is adopted. The MVRPZINB model provides a better fit in terms of both deviance information criteria (DIC) and root mean square error (RMSE) values for all three crash types than other common models. The model comparison shows that none of the four types of unobserved heterogeneities was negligible. The MVRPZINB model reveals 9 out of 18 covariates to be able to significantly influence crash frequency of the studied midblock segments. It is found that number of lanes, AADT per lane, and segment length might have non-positive effects on crash frequencies for some segments. Thus, it should be careful to use them as exposure variables in future studies. The segments with the speed limit of 45 mph tend to have fewer crashes than those with lower speed limits, and the segments in Omaha tend to have fewer crashes than those in Lincoln. It is also found that the presence of shoulder, presence of on-street parking, presence of one-way setting, and lane width do not have significant influences on crash frequencies. In addition, the MVRPZINB model makes it possible to identify the segment-specific effects of various factors on crash frequencies. These findings are informative for transportation agencies to take correct and efficient measures to improve traffic safety.

The dissertation closes with an overall summary of our findings and a general discussion of future extensions.

References

AASHTO, 2009. Highway Safety Manual. National Research Council (US). Transportation Research Board. Task Force on Development of the Highway Safety Manual, American Association of State Highway, Transportation Officials. Joint Task Force on the Highway Safety Manual, et al.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43 (5), 1666–1676.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography 46 (2), 234–240.

White House, 2016. 2015 Traffic Fatalities Data Has Just Been Released: A Call to Action to Download and Analyze. https://obamawhitehouse.archives.gov/blog/2016/08/29/2015-traffic-fatalities-data-has-just-been-released-call-action-download-and-analyze (accessed 2.20.17).

CHAPTER 2. EXPLORING SPATIO-TEMPORAL EFFECTS IN TRAFFIC CRASH TREND ANALYSIS

A paper published on the Analytic Methods in Accident Research

Abstract

Unobserved heterogeneity produced by spatial and temporal correlations of crashes often needs to be captured in crash frequency modeling. Although many studies have included either spatial or temporal effects in crash frequency modeling, only a limited number of studies have considered both. This study addresses the limitations of existing studies by exploring multiple models that best fit the spatial and temporal correlations. In this study, we used Bayesian spatio-temporal models to investigate regional crash frequency trends, and explored the effects of omitting spatial or temporal trends in spatio-temporal correlated data. The fast Bayesian inference approach, integrated nested Laplace approximation, was used to estimate parameters. It was found that fatal crashes showed decreasing trends in all Iowa counties from 2006 to 2015, but the decreasing rates varied by counties. Among all the covariates investigated, only vehicle miles traveled (VMT) was significant. None of the socio-economic or weather indicators were found to be significant in the presence of VMT. Both spatial and temporal effects were found to be important, and they were responsible for both over dispersion and zero inflation in the crash data. In addition, spatial effects played a more important role than did temporal effects in the studied dataset, but temporal component selection was still important in spatio-temporal modeling.

Keywords: spatio-temporal modeling, Bayesian, Integrated Nested Laplace Approximation, conditional autoregressive, unobserved heterogeneity

2.1 Introduction

Traffic crashes have been one of the major sources of fatalities and injuries in the United States. Crash frequency models often are used to identify the factors influencing the propensity of traffic crashes. The most common crash frequency model is the Poisson model. When crashes show over dispersion, quasi-Poisson, Poisson log-normal model (PLN), and negative binomial (NB) models are often adopted. Unobserved heterogeneity is often an issue in crash frequency analysis, because many crash-related factors are often not observed by the analyst (Mannering et al., 2016). The excess zeros in crash data can be a result of unobserved heterogeneity (Mullahy, 1997), often causing zero-inflated and hurdle models to be adopted (Lord et al., 2005; Lord and Mannering, 2010; Malyshkina and Mannering, 2010; Mannering et al., 2016; Mannering and Bhat, 2014). In addition, the zero-state Markov switching model, which allows observations to switch between zero and normal-count states over time, has been proven to be a viable alternative to zero-inflated models (Malyshkina and Mannering, 2010). Because crash data are often aggregated over time and space, spatial and temporal correlations are often also responsible for a portion of unobserved heterogeneity, as crashes that occur close in space or time are very likely to share some unobserved characteristics (Lord et al., 2005; Lord and Mannering, 2010; Mannering et al., 2016; Mannering and Bhat, 2014; Savolainen et al., 2011). However, these spatial and temporal correlations are often overlooked in existing studies, and neglecting them may produce inefficient or biased estimated results (Mannering et al., 2016; Mannering and Bhat, 2014; Savolainen et al., 2011).

The spatial correlation of traffic crashes may exist on a macro- or microscopic spatial scale. At a macroscopic level, factors such as census tract (Wang and Kockelman, 2013), traffic analysis zone (Matkan and Mohaymany, 2013), ZIP code level (Ponicki et al., 2013), census block group (Noland et al., 2013), census ward (Boulieri et al., 2016; Quddus, 2008), county

(Aguero-Valverde and Jovanis, 2006; Eckley and Curtin, 2013; Song et al., 2006), and state/province (Erdogan, 2009; Truong et al., 2016), as well as similarity of economic and social activities, culture, land use, and enforcements within a given region, may explain the spatial correlation in traffic crashes. At a microscopic level, crashes occurring at nearby intersections (Abdel-Aty and Wang, 2006; Ahmed and Abdel-Aty, 2015; Guo et al., 2010; Liu et al., 2015; Mitra et al., 2007; Pulugurtha and Sambhara, 2011; Wang and Abdel-Aty, 2006; Xie et al., 2014) or adjacent road segments (Aguero-Valverde, 2011; Aguero-Valverde and Jovanis, 2008; Jiang et al., 2014; Wang et al., 2011, 2009; Zeng and Huang, 2014) may be correlated as a result of geometric or traffic flow similarities (Levine et al., 1995).

Temporal correlation captures the variability of traffic crashes with temporal scales such as year (Andrey, 2010; Boulieri et al., 2017; Brijs et al., 2008; El-Basyouny et al., 2014; Malyshkina and Mannering, 2010; Matkan and Mohaymany, 2013; Wang et al., 2011; Wang and Abdel-Aty, 2006; Yannis et al., 2011), month (Hu et al., 2013; Quddus, 2008), week (Kilamanua et al., 2011; Liu et al., 2015; Malyshkina et al., 2009; Sukhai et al., 2011), day (Brijs et al., 2008), and hour (Kilamanua et al., 2011; Liu et al., 2015). Temporal correlation reflects the influence of different traffic-related factors, such as economy, weather, environment, law, and travel demand, which often exhibit some temporal trends or periodicities.

Depending on the study site, one of three scenarios is feasible: (a) the crash data may show both spatial and temporal effects, (b) these effects may exist individually, or (c) neither of them may exist. When spatial and temporal effects co-exist, their interaction (i.e. spatio-temporal effects) also needs to be considered. Although many studies have included either spatial effects or temporal effects in crash frequency modeling, only a limited number of studies have considered both of them. Miaou et al. (2003) first introduced the spatio-temporal modeling

approach to traffic crash modeling in analyzing yearly county-level crash rates in Texas from 1992 to 1999 using multiple spatio-temporal models. Wang and Abdel-Aty (2006) analyzed spatial and temporal correlations for rear-end crashes at signalized intersections in Florida. However, they built separate models for spatial effects and temporal effects. Jiang et al. (2014) considered both spatial and temporal correlations in analyzing the crashes on urban four-lane divided arterial segments in the central Florida area. However, they assumed that the spatial and temporal effects followed normal distributions without presenting any data-driven evidence to support their assumption. Truong et al. (2016) analyzed yearly crash fatalities of 63 provinces in Vietnam from 2012 to 2014 using the conditional autoregressive (CAR) spatio-temporal autocorrelation technique. The CAR spatio-temporal model performed better than the random effects NB model and random parameters NB model did in terms of both goodness of fit and crash prediction. Aguero-Valverde and Jovanis (2006) had similar findings.

The CAR model (Besag, 1974; Besag et al., 1991) often is used for modeling areal data in spatial statistics. Several researchers (Aguero-Valverde and Jovanis, 2006; Boulieri et al., 2017; Truong et al., 2016; Wang et al., 2011) have used the CAR model to illustrate spatial correlations paired with different temporal models. However, they all showed only one temporal model, despite the fact that the choice of a particular temporal model was also very important (Miaou et al., 2003). In this study, we used the spatio-temporal crash frequency model to identify the long-term county-level fatal crash frequency trends in Iowa. Multiple temporal components were built and contrasted to choose the most appropriate model. A fast Bayesian estimation tool, integrated nested Laplace approximation (INLA), was used to estimate these spatio-temporal models.

The workflow of the data analysis is as follows:

- First, we discuss whether crashes have over dispersion and zero inflation.
- Second, we examine spatial correlations and temporal correlations of crashes.
- Third, we evaluate the necessity of including the spatial component, temporal component, and spatio-temporal component in modeling, and we also discuss the temporal component selection.
- Finally, after determining the final model, the estimation results are discussed.

The rest of paper is organized as follows. Section 2 comprises a discussion of the traffic crash data used for this study. Section 3 presents the statistical models and estimation methods used in this study. Section 4 includes the analyses and discussions of the observed results. A conclusion and future recommendations are provided in section 5.

2.2 Data Description

Traffic crash data for Iowa's 99 counties from 2006 to 2015 were obtained from the Iowa Department of Transportation. Based on their severity, the crashes were divided into five categories: fatal, major injury, minor injury, possible injury/unknown, and property damage only. Fatal crashes were analyzed for this study, as they usually cause much more severe outcomes than do other types of crashes. The vehicle miles traveled (VMT) data for each county in each year from 2006 to 2015 were downloaded from the website of the Iowa Department of Transportation (2016). In addition, population and unemployment rate data were downloaded from the website of Iowa Community Indicators Program (2016), and per capita personal income data were downloaded from the website of the U.S. Bureau of Economic Analysis (2016). Because weather has been shown to significantly influence crash frequencies in many studies (Brijs et al., 2008; Golob and Recker, 2003; Knapp et al., 2000; Maze et al., 2005), rainfall amounts, snowfall amounts, and the number of days with a minimum temperature higher than

32°F (TH32) were downloaded from the website of the Iowa Environmental Mesonet (2017). These data were collected based on the daily climate observations from the National Weather Service's Cooperative Observer Program. A summary of the variables is given in Table 2-1.

The variance of fatal crashes was larger than the mean, which implied that over-dispersion was occurring. The proportion of zero crashes, used to preliminarily check whether or not zero-inflated models are needed, is shown in the last column of Table 2-1. The zero proportion of fatal crashes was 0.113, much larger than 0.034, which was the supposed probability value of zero under a Poisson distribution with the mean being 3.383. This implies that the zero-inflated model may be considered.

Table 2-1 *Descriptive statistics of collected variables*

Variables	Mean	Std. Error	Median	Min.	Max.	Zero-proportion
Fatal crash frequency	3.383	3.818	2.000	0.000	35.000	0.113
VMT (1,000,000 miles)	0.320	0.487	0.186	0.047	4.215	—
Population (10,000)	3.076	5.273	1.571	0.380	46.771	—
Unemployment rate (%)	4.846	1.347	4.600	2.000	10.200	—
Income ($10,000)	3.877	0.666	3.877	2.247	6.464	—
Rainfall (inch)	38.390	8.570	38.610	17.850	64.990	—
Snowfall (inch)	34.560	14.377	35.000	0.000	85.100	—
TH32 (days)	222.600	15.733	221.000	174.000	272.000	—

Note: VMT, vehicle miles traveled; TH32, number of days with minimum temperature higher than 32°F.

Unobserved heterogeneity caused by spatial and temporal correlations of data often can be found by visualizing the data and corroborated with statistical methods. The yearly average fatal crash frequencies for each county in Iowa is shown in Figure 2-1. As expected, fatal crash data revealed a cluster of high numbers of crashes in the central counties around the yellow-shaded area, where the largest city of Iowa, Des Moines, is located. Fatal crash data also revealed a cluster of low numbers of crashes in the northern and southwestern parts of Iowa (deep-shaded areas). Next, statistical analysis was performed to investigate the presence of spatial correlations.

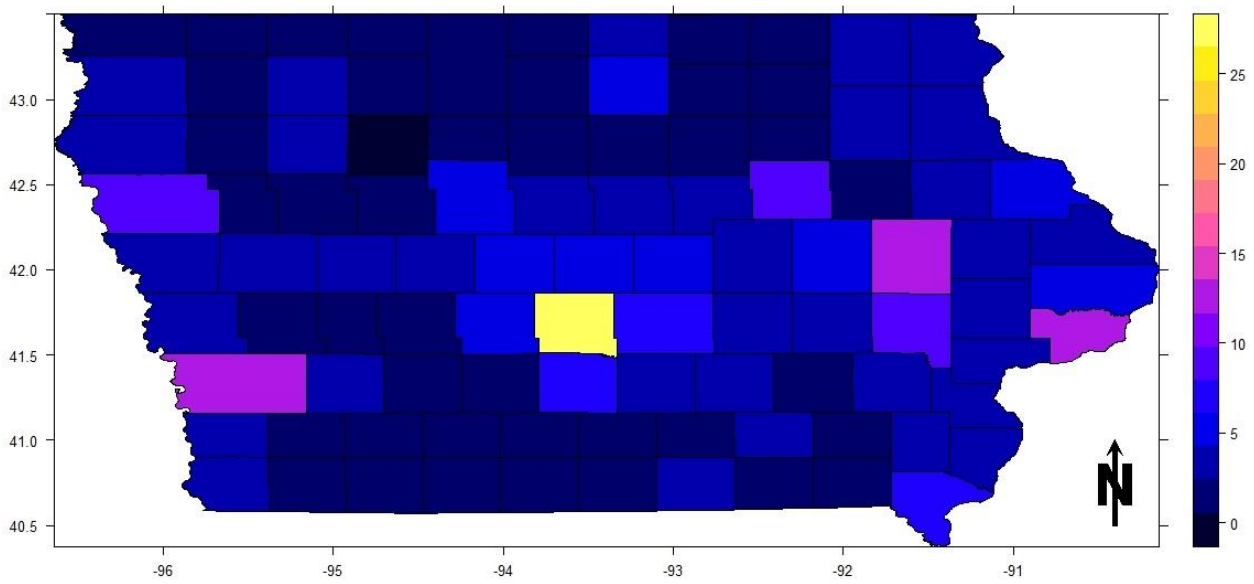

Figure 2-1 *County-level yearly average fatal crash counts of Iowa (2006-2015)*

Moran's *I* statistic is commonly used to test spatial correlations in traffic crash analysis (Guo et al., 2010; Quddus, 2008; Xie et al., 2014; Zeng and Huang, 2014). The global Moran's *I* is defined as (Anselin, 1988):

$$I = \frac{n\sum_i\sum_i \omega_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_{i\neq j}\omega_{ij}\sum_i(y_i-\bar{y})^2}, \quad (2.1)$$

where n is the total number of observations, y_i and y_j are the values of observation i and observation j, $\bar{y}$ is the average value of observations, and ω_{ij} is the spatial weight between observations i and j.

Negative Moran's *I* values indicate negative spatial autocorrelation, positive values indicate positive spatial autocorrelation, and zero indicates no spatial autocorrelation. The *z*-score of Moran's *I* shows if the spatial autocorrelation is significant.

The global Moran's *I* statistics of fatal crashes in each year from 2006 to 2015 were calculated using the "spdep" package (Bivand and Piras, 2015) in the R platform (R Core Team, 2016) with queen continuity spatial weights, whereby counties with a shared border or vertex

were considered as neighbors. When areas were neighbors, the spatial weights were 1; otherwise, they were 0. The results are shown in Table 2-2.

Table 2-2 *Global Moran's I statistics of fatal crashes in each year*

Year	Moran's *I*	*P*-value
2006	1.986	0.024*
2007	2.091	0.018*
2008	1.520	0.064
2009	1.661	0.048*
2010	2.486	0.006*
2011	1.919	0.027*
2012	1.240	0.108
2013	2.387	0.008*
2014	1.241	0.107
2015	2.300	0.011*

Note: *significant at $P = 0.05$.

Significant spatial autocorrelations for fatal crashes existed in 7 out of 10 years at a 95% confidence level and at a 90% confidence level for the remaining 3 years. Thus, fatal crashes were highly likely to be spatially correlated at the county level in Iowa. These trends may be site specific. For example, Aguero-Valverde and Jovanis (2006) found the county-level yearly fatal crashes of Pennsylvania to not be significantly correlated. This suggests that the presence and type of spatial correlation is site and data sensitive. Therefore, no prior assumptions should be made about the presence or absence of spatial correlation, and it is recommended to statistically test the presence of spatial correlation prior to modeling.

Temporal correlation was not directly tested, as there were only 10 time points in this dataset. However, as shown in Figure 2-2, the yearly fatal crash counts of Iowa from 2006 to 2015 revealed a linearly decreasing trend, which needed to be considered when building the model.

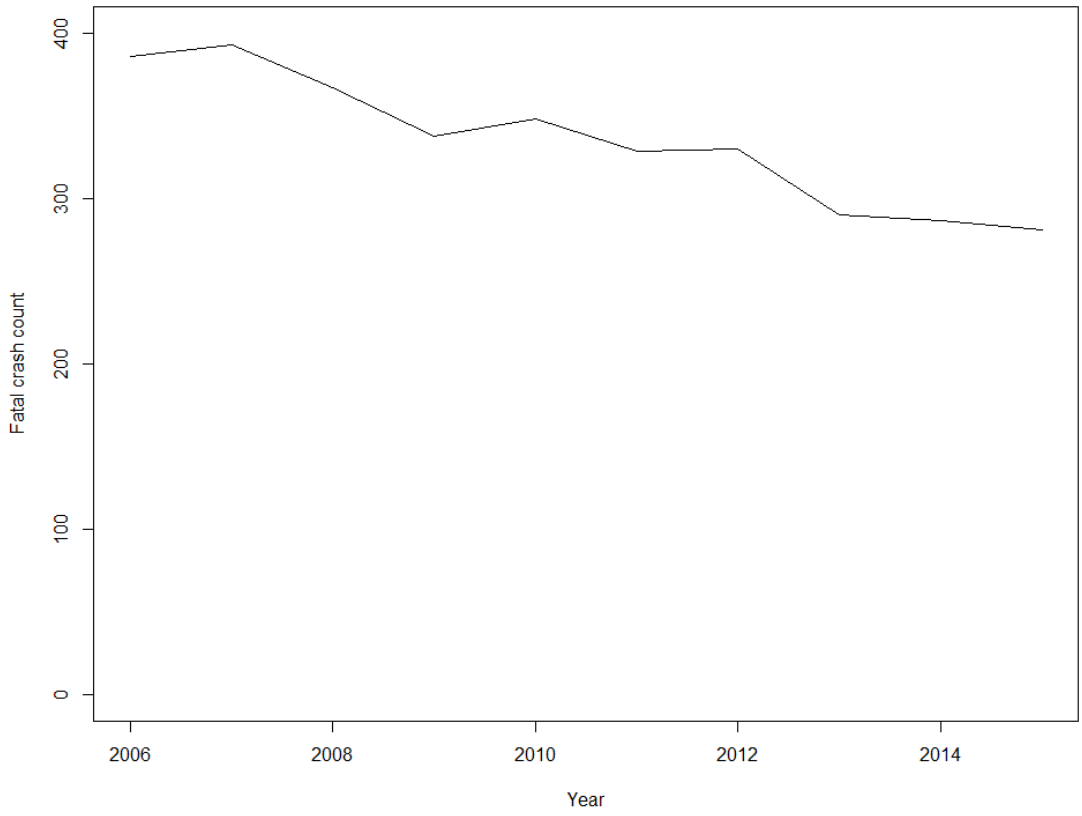

Figure 2-2 *Iowa state-level yearly fatal crash counts (2006-2015)*

2.3 Methodology

2.3.1 Statistical Framework

The statistical framework uses a Bayesian hierarchical architecture, including both the spatial and temporal random effect components. The statistical model is presented in equations 2 and 3:

$$y_{it} \sim Poisson(\lambda_{it}) \tag{2.2}$$

$$\log(\lambda_{it}) = \alpha + \beta * X_{it} + \upsilon_i + \nu_i + \varphi_t + \eta_{it}, \tag{2.3}$$

where i is the county number, 1,2,… 99; t is the year, 1 (2006), 2 (2007), …,10 (2015); y_{it} is the crash count of county i in year t; λ_{it} is the mean crash frequency of county i in year t; α is the intercept term; β is the regression coefficient vector; X_{it} is the covariate vector of county i in year t; υ_i is the structured spatial random effect of county i; ν_i is the unstructured spatial

random effect of county i; φ_t is the temporal random effect in year t; and η_{it} is the spatio-temporal interaction effect.

The spatial and temporal components helped us to identify the underlying unobserved heterogeneity across county and year. For this study, we analyzed three kinds of spatio-temporal models that had the same spatial component but different temporal components.

2.3.1.1 Spatial Component

The spatial component, i.e. $\upsilon_i + \nu_i$, was assumed to follow the Besag-York-Mollie (BYM) model (Besag et al., 1991). The BYM model has been widely used in traffic accident analysis (Aguero-Valverde and Jovanis, 2006; Boulieri et al., 2017; Wang et al., 2013; Xie et al., 2014) and has been recommended for traffic crash analyses (Boulieri et al., 2017). For the BYM model, the structured spatial effect, υ_i, is modeled using an intrinsic conditional autoregressive (ICAR) structure, and the unstructured spatial effect, ν_i, follows a normal distribution.

$$\upsilon_i|\upsilon_{j\neq i} \sim N\left(\frac{\sum_{j\in N(i)}\upsilon_j}{\#N(i)}, \frac{{\tau_\upsilon}^{-1}}{\#N(i)}\right) \tag{2.4}$$

$$\nu_i \sim N(0, {\tau_\nu}^{-1}), \tag{2.5}$$

where $N(i)$ are the neighbors of county i, $\#N(i)$ are the number of neighbors of county i, and τ_υ and τ_ν are precisions.

The ICAR part accounts for possible spatial correlations between counties, and the unstructured part is responsible for county individual heterogeneity.

2.3.1.2 Temporal Component

Three temporal models, including the linear temporal model, the 1st order autoregressive (AR1) model, and the 1st order random walk (RW1) model, were considered.

The linear temporal model is defined in equations 6 and 7 (Bernardinelli et al., 1995):

$$\varphi_t = (\beta_2 + \delta_i) * t \tag{2.6}$$

$$\delta_i \overset{iid}{\sim} N(0, {\tau_\delta}^{-1}), \tag{2.7}$$

where β_2 is the global time trend; δ_i is the interaction between time and county i, $\delta_i < 0$ implies that the area-specific trend is smaller than the mean trend, whereas $\delta_i > 0$, implies that the area-specific trend is larger than the mean trend; and τ_δ is the precision.

δ_i could reflect the degree to which spatial effects and temporal effects have interactions (Blangiardo et al., 2013).

The AR1 model is defined in equations 8, 9, and 10:

$$\varphi_t \sim \begin{cases} N\left(0, \left(\tau_\varphi(1-\rho^2)\right)^{-1}\right) & for\ t = 1 \\ \rho\varphi_{t-1} + \varepsilon_t & for\ t = 2,3,..10 \end{cases} \tag{2.8}$$

$$|\rho| < 1 \tag{2.9}$$

$$\varepsilon_t \sim N(0, {\tau_\varepsilon}^{-1}), \tag{2.10}$$

where ρ is a correlation parameter, ε_t is the white noise, and τ_ε is a precision.

The RW1 model is defined in equations 11 and 12:

$$\varphi_{t+1} = \varphi_t + \gamma_t \tag{2.11}$$

$$\gamma_t \overset{iid}{\sim} N\left(0, {\tau_\gamma}^{-1}\right), \tag{2.12}$$

where γ_t is the white noise and τ_γ is a precision.

2.3.1.3 Spatio-Temporal Component

The spatio-temporal component, η_{it}, is assumed to follow a zero-mean normal distribution.

$$\eta_{it} \overset{iid}{\sim} N\left(0, {\tau_\eta}^{-1}\right). \tag{2.13}$$

where τ_η is a precision.

Due to the presence of η_{it}, this statistical model becomes the Poisson log-normal model.

In addition, the performance of the best spatio-temporal model, which is the linear temporal component model as proven later, is compared against several traditional models discussed below.

2.3.1.4 Other Comparison Models

2.3.1.4.1 Spatial Effects and Temporal Effects Assessment

Three models, one with no spatial or temporal effects, one with only spatial effects, and one with only temporal effects, were compared against the best spatio-temporal model to assess the importance of explicitly accounting for spatial and temporal effects.

2.3.1.4.2 Poisson Model vs. Zero-Inflated Poisson (ZIP) model

As shown in Table 2-1, fatal crashes had zero inflation. Thus, the ZIP model was also built for comparison. It should be noted that for zero-inflated crash data, the zero-state Markov switching model has been shown to be superior to the zero-inflated model (Malyshkina and Mannering, 2010). However, the zero-state Markov switching model is not discussed here, as the focus is on explaining zero inflation caused by spatial or temporal correlations and hence can be explicitly explained using a ZIP model. All combinations of spatial, temporal, and base case models explored in this study are listed in Table 2-3.

Table 2-3 *Summary of models developed for fatal crash frequency analysis*

No	Model code	Spatial effect	Temporal effect	Spatio-temporal effect	Base model
1	$S_0T_0ST_0P$	—	—	—	Poisson
2	$S_{BYM}T_0ST_0P$	BYM	—	—	Poisson
3	$S_0T_LST_0P$	—	Linear	—	Poisson
4	$S_{BYM}T_LST_0P$	BYM	Linear	—	Poisson
5	$S_{BYM}T_LST_1P$	BYM	Linear	η_{it}	Poisson
6	$S_{BYM}T_{AR1}ST_1P$	BYM	AR1	η_{it}	Poisson
7	$S_{BYM}T_{RW1}ST_1P$	BYM	RW1	η_{it}	Poisson
8	$S_{BYM}T_LST_1ZIP$	BYM	Linear	η_{it}	ZIP

Note: 0, component not included; 1, component included; L, linear temporal; BYM, Besag-York-Mollie; AR1, 1st order autoregressive; RW1, 1st order random walk; ZIP, zero-inflated Poisson; "—" means non-existent.

2.3.2 Integrated Nested Laplace Approximation (INLA)

Bayesian models are usually solved with Markov chain Monte Carlo (MCMC) simulations. However, when the models are very complex without close-form posterior density available, as in this case, the MCMC method can be very time consuming if both spatial and temporal effects are included. Rue and Martino (2009) proposed the INLA method to numerically approximate the full Bayesian inference for latent Gaussian models. INLA can produce much faster results than can the MCMC approach for Bayesian models without compromising accuracy (Martins et al., 2013), as it can accurately derive the posterior densities by numerical approximation and significantly decrease the MCMC simulation workload.

Assume y is the response vector, θ is the target parameter vector, and ψ is the hyper-parameter vector. The posterior probability densities of parameter elements and hyper-parameter elements in Bayesian models are (Blangiardo et al., 2013):

$$p(\theta_i|y) = \int p(\psi|y)p(\theta_i|\psi,y)d\psi \tag{2.14}$$

$$p(\psi_k|y) = \int p(\psi|y)d\psi_{-k}, \tag{2.15}$$

where i is the ith observation; θ_i is the ith parameter; ψ_k is the kth hyper-parameter; and ψ_{-k} is the complement hyper-parameter set to ψ_k.

The INLAs for the posterior densities of interest can be written as (Blangiardo et al., 2013; Rue et al., 2009):

$$p(\psi|y) = \frac{p(\theta,\psi|y)}{p(\theta|\psi,y)} \propto \frac{p(\psi)p(\theta|\psi)p(y|\theta)}{p(\theta|\psi,y)} \approx \frac{p(\psi)p(\theta|\psi)p(y|\theta)}{\tilde{p}(\theta|\psi,y)}|_{\theta=\theta^*(\psi)} =: \tilde{p}(\psi|y) \tag{2.16}$$

$$p(\theta_i|\psi,y) = \frac{p((\theta_i,\theta_{-i})|\psi,y)}{p(\theta_{-i}|\theta_i,\psi,y)} \approx \frac{p(\theta,\psi|y)}{\tilde{p}(\theta_{-i}|\theta_i,\psi,y)}|_{\theta_{-i}=\theta^*{}_{-i}(\theta_i,\psi)} =: \tilde{p}(\theta_i|\psi,y), \tag{2.17}$$

where $\tilde{p}(\psi|y)$ is the Gaussian approximation of $p(\theta|\psi,y)$ and $\theta^*(\psi)$ is its mode and $\tilde{p}(\theta_{-i}|\theta_i,\psi,y)$ is the simplified Laplace approximation based on the Taylor's series expansion of the Laplace approximation of $\tilde{p}(\theta_i|\psi,y)$.

As compared to the Gaussian approximation, the simplified Laplace approximation in equation 17 provides a good balance between speed and accuracy.

INLA first obtains the marginal joint posterior of $\tilde{p}(\psi|y)$ to locate the mode by grid search. Then, for each ψ^* with the corresponding weight ${w_\psi}^*$, the conditional posteriors $\tilde{p}(\theta_i|\psi^*, y)$ are also obtained by grid search. Finally, the marginal posteriors $\tilde{p}(\theta_i|y)$ are obtained by numerical integration:

$$\tilde{p}(\theta_i|y) \approx \sum_{\psi^* \in G} \tilde{p}(\theta_i|\psi^*, y)\tilde{p}(\psi^*|y){w_\psi}^*. \quad (2.18)$$

More details about INLA can be found elsewhere (Blangiardo et al., 2013; Hu et al., 2013; Martins et al., 2013; Rue et al., 2009).

All eight models listed in Table 2-3 were implemented in the R environment (R Core Team, 2016) using the 'INLA' package (Lindgren and Rue, 2015; Martins et al., 2013; Rue et al., 2009). The regression coefficients β were assigned independent normal distributions $N(0, 1000)$. Six hyper-parameters are defined in this study, i.e. the precision parameters τ_υ, τ_ν, τ_δ, τ_ε, τ_γ, and τ_η. The logarithm of these values were assigned to follow the log-Gamma distribution $logGamma(1,0.0005)$ (Blangiardo et al., 2013).

2.3.3 Model Comparison and Checking

The deviance information criterion (DIC) was used as a measure of assessing different Bayesian models (Spiegelhalter et al., 2002). DIC is defined as

$$DIC = D(\bar{\theta}) + 2p_D = \bar{D} + p_D, \quad (2.19)$$

where $D(\bar{\theta})$ is the deviance using the posterior mean values of the estimated parameters ($\bar{\theta}$), $\bar{D}$ is the posterior mean of deviances, and p_D is the effective number of parameters.

Similar to Akaike's information criterion (AIC), DIC considers both the Bayesian measure of fit or adequacy and the complexity of the model (Spiegelhalter et al., 2002). Models

with smaller DIC values are expected to perform better. Roughly, differences of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are substantial, and differences less than 5 might mean that the models are not significantly different (MRC Biostatistics Unit, 2004).

However, DIC may under-penalize complex models with many random effects (Plummer, 2008), such as CAR models. Thus, the conditional predictive ordinate (CPO) (Pettit, 1990) and the cross-validated probability integral transform (PIT) (Dawid, 1984) were also calculated for model assessment. Both of them are leave-one-out cross validation scores.

$$CPO_i = \pi(y_i|\boldsymbol{y}_{-i}) \tag{2.20}$$

$$PIT_i = p(Y_i \leq y_i|\boldsymbol{y}_{-i}), \tag{2.21}$$

where ${y_i}^{obs}$ is the ith observation and $\boldsymbol{y}_{-i}$ represents all the observations except the ith one.

The negative mean logarithmic CPO was calculated as a measure of the predictive quality of the model (Gneiting and Raftery, 2007; Roos and Held, 2011).

$$\overline{CPO} = -\frac{1}{n}\sum_i^n \log(CPO_i) \tag{2.22}$$

Stone (1977) proved that the $\overline{CPO}$ was asymptotically equivalent to AIC. Thus, $\overline{CPO}$ can be used for model choice, and a lower value of $\overline{CPO}$ indicates a better model.

A large or small PIT value indicates possible outliers, and the PIT values of a well-calibrated model should be uniformly distributed. Thus PIT histograms can be used to assess the calibration of a model (Czado et al., 2009). For count data, an adjusted PIT should be used instead to make the predictive distribution continuous (Czado et al., 2009).

$$Adjusted\ PIT_i = PIT_i - \frac{1}{2}CPO_i \tag{2.23}$$

In addition, root mean square error (RMSE) and mean absolute error (MAE) were also calculated to evaluate the adequacy of model fit.

$$RMSE = \sqrt{\frac{1}{n_0}\sum_{j=1}^{n_0}\left(O_j - P_j\right)^2} \tag{2.24}$$

$$MAE = \frac{1}{n_0}\sum_{j=1}^{n_0}\left|O_j - P_j\right|, \tag{2.25}$$

where O_j is the jth observation value, P_j is the predicted ith value from the model, and n_0 is the number of observations.

Similar to DIC, smaller MAE and RMSE values are desired.

2.3.4 Spatial Fraction Analysis

For the spatio-temporal analysis, one point of interest was to identify the contribution of the structured spatial effects ${\sigma_v}^2$ over the total marginal spatial variability ${\sigma_v}^2 + {\sigma_\nu}^2$ (Boulieri et al., 2017). The spatial fraction of interest is given by

$$frac_v = \frac{{\sigma_v}^2}{{\sigma_v}^2 + {\sigma_\nu}^2} = \frac{1/\tau_v}{1/\tau_v + 1/\tau_\nu}, \tag{2.26}$$

where ${\sigma_v}^2$ is the variance of the structured spatial effects, ${\sigma_\nu}^2$ is the variance of the unstructured spatial effects, and τ_v and τ_ν are the corresponding precisions.

When the spatial fraction is close to 1, the structured spatial effects explain most of the variability of the model. Otherwise, the unstructured spatial random effects play the main role.

2.4 Results and Discussions

All eight models listed in Table 2-3 were implemented in INLA. On an Intel(R) Xeon(R) CPU at 3.70 GHz with 16 GB random access memory, it took a total of 73.074 sec to run these eight models. As a comparison, it took INLA 13.609 sec to estimate the $S_{BYM}T_LST_1P$ model, whereas it took OpenBUGS (Sturtz et al., 2005) 1,053 sec to estimate the same model with the MCMC simulation settings of three simulation chains, 5,000 burn-in samples, and 5,000 adopted samples with a thin interval set at 2. The computation time was greatly reduced using INLA, and the computation time is expected to be saved more with the increase of data and parameters.

The DIC, $\overline{CPO}$, RMSE, and MAE values of the eight models listed in Table 2-3 are shown in Table 2-4. These four measures help in identifying the best spatio-temporal model. The following observations can be made from data shown in Table 2-4.

Table 2-4 *DIC, $\overline{CPO}$, and RMSE, MAE values for all the models*

No	Model	DIC	$\overline{CPO}$	RMSE	MAE
1	$S_0T_0ST_0P$	4282.01	2.172	2.652	1.810
2	$S_{BYM}T_0ST_0P$	3791.62	1.920	1.851	1.350
3	$S_0T_LST_0P$	3860.00	1.953	1.987	1.421
4	$S_{BYM}T_LST_0P$	3749.39	1.896	1.757	1.315
5	$S_{BYM}T_LST_1P$	3746.13	1.894	1.757	1.314
6	$S_{BYM}T_{AR1}ST_1P$	3750.60	1.899	1.762	1.316
7	$S_{BYM}T_{RW1}ST_1P$	3752.33	1.896	1.765	1.319
8	$S_{BYM}T_LST_1ZIP$	3749.35	1.895	1.756	1.314

Note: 0, component not included; 1, component included; L, linear temporal component; BYM, Besag-York-Mollie; AR1, 1st order autoregressive; RW1, 1st order random walk; ZIP, zero-inflated Poisson; "—" means non-existent.

2.4.1 Choice of the Temporal Component

The DIC values do not show significant differences among the $S_{BYM}T_LST_1P$, $S_{BYM}T_{AR1}ST_1P$, and $S_{BYM}T_{RW1}ST_1P$ models, but the $S_{BYM}T_LST_1P$ model with the linear temporal component had the lowest $\overline{CPO}$, RMSE, and MAE values. In addition, the adjusted PIT histogram of the $S_{BYM}T_LST_1P$ model is shown in Figure 2-3, where the adjusted PIT values show a very good uniform distribution. That is, the $S_{BYM}T_LST_1P$ model was well calibrated for the data. Thus, the $S_{BYM}T_{\mathrm{L}}ST_1P$ model was considered as the best fit in this case; that is, fatal crash frequencies had some linear changing trend. Although these models did not show large differences, the results still implied the necessity of temporal component selection, especially considering different models would lead to different interpretations of the data. For example, the linear temporal component implies that the number of fatal crashes would change linearly in the future, but the same conclusion may not be drawn from the RW1 temporal component.

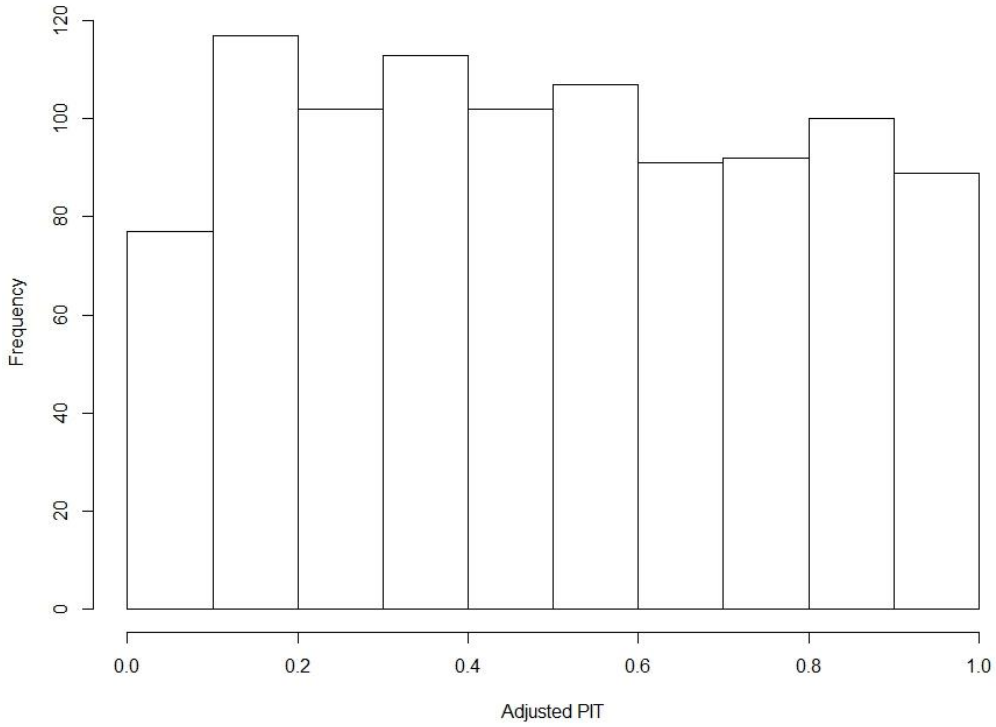

Figure 2-3 *Histogram of the adjusted PIT values of the* $S_{BYM}T_LST_1P$ *model*

2.4.2 Necessity of Including Spatial, Temporal, and Spatio-Temporal Effects

The $S_{BYM}T_LST_0P$ model performed much better than did the $S_0T_0ST_0P$, $S_{BYM}T_0ST_0$, and $S_0T_LST_0P$ models in terms of all four measures. This means that, in this case, both spatial and temporal effects played important roles in unobserved heterogeneity and thus needed to be considered. Meanwhile, because the $S_{BYM}T_0ST_0P$ model had much lower DIC, $\overline{CPO}$, RMSE, and MAE values than did the $S_0T_LST_0P$ model, spatial effects had a greater influence than did temporal effects in this case. This finding indicates that fatal crashes have very strong correlations across counties in Iowa. Only 10 years of data were used for this study, and it may not be a long enough time span for crashes to show a big change over time. If more years of data were available or monthly data had been analyzed, the temporal effects may have played a more

important role. The $S_{BYM}T_LST_1P$ model was slightly better than $S_{BYM}T_LST_0P$ model, which meant the spatio-temporal interaction effects were very weak.

2.4.3 Zero-Inflation of Crashes

The $S_{BYM}T_LST_1P$ model had nearly the same performance as the $S_{BYM}T_LST_1ZIP$ model did in terms of all four measures. In addition, the zero-inflation probability value, which showed the probability of zero crashes being from the zero state, was only 0.0046 for the $S_{BYM}T_LST_1ZIP$ model. This means that there was no longer a need to consider zero inflation after including spatial and temporal effects, as the zero inflation of fatal crashes could be well explained by spatial and temporal effects. This finding provides a new point of view for the explanation of where zero inflation comes from in crash data.

Because the $S_{BYM}T_LST_1P$ model had the best performance of all eight models, it was used in the following analysis. The estimated parameters, their standard errors, and 95% credible intervals are shown in Table 2-5. As expected, VMT had significant positive effects. However, all the other variables were statistically insignificant. It is thought that population, employment rate, and income indicators in Iowa had been relatively consistent from 2006 to 2015 because Iowa was a typical farming state and there were no significant changes in these variables. Thus, these indicators did not show significant influences. In addition, although adverse weather may increase the number of crashes in the short term, the results here show that weather may not have a big influence on fatal crashes in the long term in Iowa.

Because only the VMT parameter was significant, the $S_{BYM}T_LST_1P$ model was rebuilt using only VMT. The results are shown in Table 2-6.

Table 2-5 *Estimated parameters of the $S_{BYM}T_LST_1P$ model with all covariates*

Parameter	Mean	Std. Err.	0.025 quantile	0.975 quantile
(Intercept)	0.427	0.431	–0.420	1.272
VMT	0.887	0.082	0.727	1.049
Population	–0.003	0.003	–0.010	0.003
Income	–0.014	0.036	–0.085	0.058
Unemployment rate	0.013	0.020	–0.027	0.053
Rainfall	–0.002	0.003	–0.007	0.003
Snowfall	0.000	0.002	–0.003	0.003
TH32	0.002	0.002	–0.001	0.006
Year	–0.041	0.006	–0.053	–0.029

Note: VMT, vehicle miles traveled; TH32, number of days with minimum temperature higher than 32°F.

Table 2-6 *Estimated parameters of the $S_{BYM}T_LST_1P$ model with only VMT*

	Intercept	VMT (β_1)	Year (β_2)	τ_υ	τ_ν	τ_δ	$frac_\upsilon$
Mean	0.923	0.887	–0.042	9.919	9.812	16424.166	0.497
Std. Err.	0.057	0.086	0.006	7.298	3.446	12860.000	—
0.025 quantile	0.810	0.714	–0.054	2.692	4.807	1926.189	—
0.975 quantile	1.032	1.046	–0.030	31.290	16.040	55533.450	—

Note: VMT, vehicle miles traveled; β_1, β_2, regression coefficients; $\tau_\upsilon, \tau_\nu, \tau_\delta$, precisions; $frac_\upsilon$ = spatial fraction.

2.4.4 Spatial Fraction Results

For the $S_{BYM}T_LST_1P$ model, the fraction of structured spatial effects was 0.497 (Table 2-6), which implied that the unstructured and structured spatial effects played nearly the same role in this case. That is, the unobserved heterogeneity in space existed both between counties and for individual counties. The exponential posterior means of the structured spatial effects of each county were shown in Figure 2-4; the counties with $\exp(\upsilon_i)$ lower than 1 tended to have fewer crashes and the counties with $\exp(\upsilon_i)$ greater than 1 tended to have more crashes. As shown in Figure 2-4, the counties located in northern and southwestern Iowa tended to have fewer fatal crashes. This finding is generally consistent with the empirically observed fatal crash distribution shown in Figure 2-1.

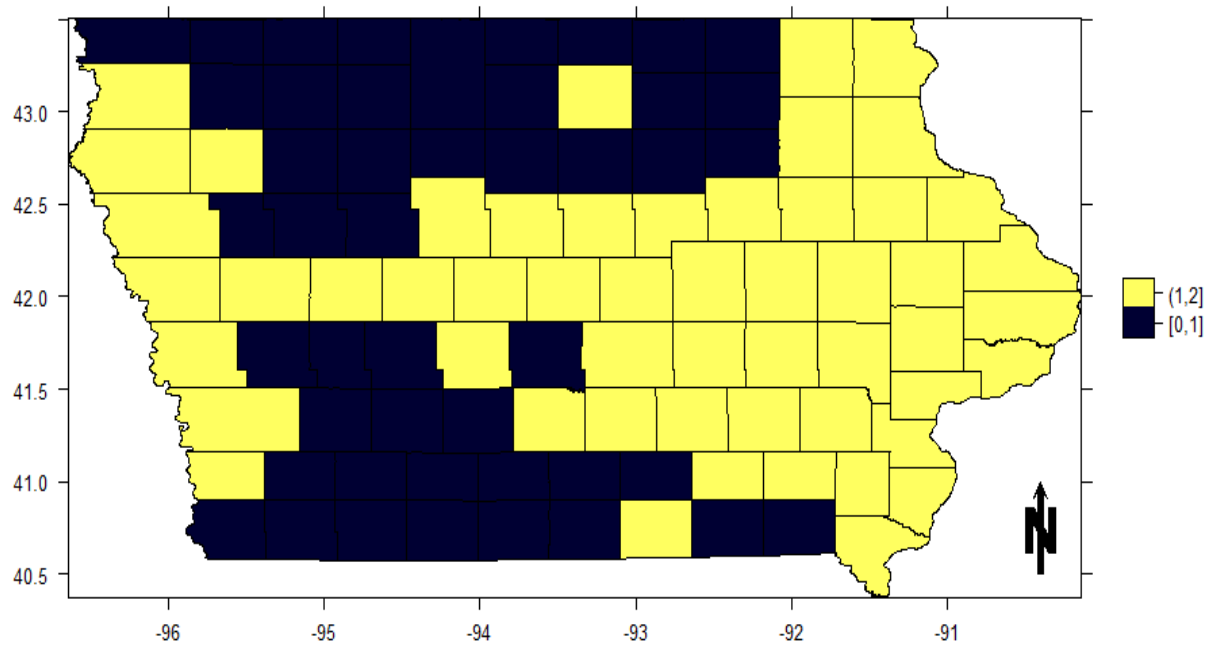

Figure 2-4 *Exponential posterior means of the structured spatial effect* ($exp(\upsilon_i)$)

Moran's *I* statistics of the residuals of the $S_{BYM}T_LST_1P$ model were calculated to see if they still had spatial correlations. As shown in Table 2-7, the p-values of residuals were significantly larger than 0.05 for any year except 2010, the *p*-value of which was very close to 0.05. Thus, the spatial component covered nearly all of unobserved heterogeneity in space. The results also verified the effectiveness of the spatial component.

Table 2-7 *Moran's I test results for the residuals of the* $S_{BYM}T_LST_1P$ *model*

Year	Moran's *I* statistic	*p*-value
2006	–1.036	0.850
2007	–0.156	0.562
2008	–0.792	0.786
2009	–0.535	0.704
2010	1.653	0.049
2011	0.292	0.385
2012	0.636	0.262
2013	0.460	0.323
2014	–0.876	0.809
2015	–1.387	0.917

2.4.5 Temporal Effects

The β_2 value of –0.042 with a 95% credible interval of [–0.054, –0.030] means that, on average, fatal crashes in Iowa significantly decreased from 2006 to 2015. The signs of δ_i values, a positive value meaning that the number of fatal crashes of county i decreased slower than the state average value and a negative value meaning that the number of fatal crashes of county i decreased faster than the state average value, are shown in Figure 2-5(a). The changing rates of fatal crashes for each county, i.e. $\beta_2 + \delta_i$, are shown in Figure 2-5(b). All $\beta_2 + \delta_i$ values were negative, which meant that the number of fatal crashes for all the counties showed decreasing trends from 2006 to 2015. The δ_i values for 50 out of the total of 99 counties were positive, whereas the δ_i values for the remaining 49 counties were negative; that is, the number of fatal crashes in 50 counties decreased slower than the mean trend of the whole state, whereas fatal crash numbers in the remaining 49 counties decreased faster than the mean trend. Thus, the first 50 counties should be the focus of future traffic safety improvement programs.

2.5 Conclusions and Future Research

Unobserved heterogeneity due to the correlations of crashes in space and time has been proven to be a big issue in many studies. However, only a limited number of studies have considered both of them in modeling crash frequency. This study explored spatial and temporal effects in crash frequency models to account for unobserved heterogeneity and accurately identified the long-term regional trends in the change of traffic crash frequencies. Focusing on the number of yearly fatal crashes at the county level in Iowa from 2006 to 2015, multiple spatio-temporal models with the same spatial component but different temporal components were developed using the Bayesian framework. INLA, a fast Bayesian model estimation methodology, was used to estimate parameters. The model with a linear temporal component was found to be

the most appropriate. Numbers of fatal crashes in all Iowa counties were found to show linearly decreasing trends but with different rates of decrease by counties. No explanatory factors, except VMT, were found to have a significant influence on fatal crash frequencies. Spatial and temporal effects were found to be responsible for both over dispersion and zero inflation of crash data, whereas spatial effects played a more important role than did temporal effects in this case.

In future research, the impact of a smaller time scale, such as season or month, should be explored, as this may offer more details about crash frequency changing trends and show the influences of periodic factors such as weather. Meanwhile, although zero inflation is not a problem anymore with the use of the spatio-temporal model for this dataset, this may not be true for other datasets. When the spatio-temporal model does not explain excess zeros completely, the zero-state Markov switching model may be combined with spatial effects to develop new spatio-temporal models. The zero-state Markov switching model could account for both zero inflation and temporal correlations, and it has been proven to be superior to traditional zero-inflated models (Malyshkina and Mannering, 2010). Finally, as Boulieri et al. (2017) has suggested, the multivariate space–time model considering factorial space and time interactions can be evaluated to better exploit spatial, temporal, and between-variable correlations, but this may need high performance computing along with complex modeling structure.

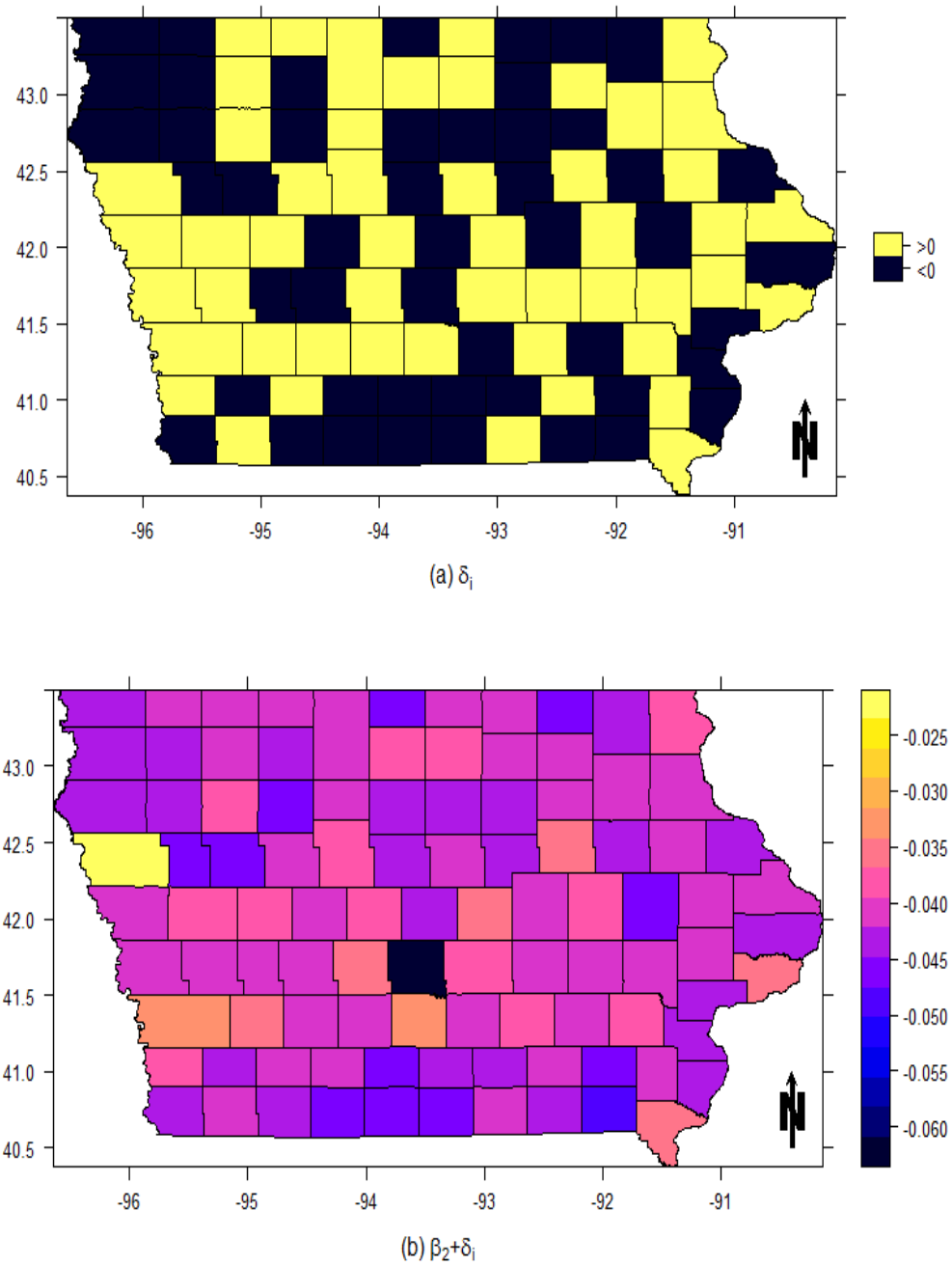

Figure 2-5 *Iowa county-level fatal crash yearly change trends from 2006 to 2015*

2.6 References

Abdel-Aty, M., Wang, X., 2006. Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. Transportation Research Record 1953, 98–111.

Aguero-Valverde, J., 2011. Direct spatial correlation in crash frequency models. 3rd International Conference on Road Safety and Simulation, Indianapolis, IN, USA.

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record 2061, 55–63.

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention 38 (3), 618–625.

Ahmed, M.M., Abdel-Aty, M., 2015. Evaluation and spatial analysis of automated red-light running enforcement cameras. Transportation Research Part C 50, 130–140.

Andrey, J., 2010. Long-term trends in weather-related crash risks. Journal of Transport Geography 18 (2), 247–258.

Anselin, L., 1988. Spatial Econometrics: Methods And Models. Springer, Netherlands.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M., 1995. Bayesian analysis of space-time variation in disease risk. Statistics in Medicine 14 (21–22), 2433–2443.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society Series B 36 (2), 192–263.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43 (1), 1–20.

Bivand, R., Piras, G., 2015. Comparing implementations of estimation methods for spatial econometrics. Journal of Statistical Software 63 (18), 1–36.

Blangiardo, M., Cameletti, M., Baio, G., Rue, H., 2013. Spatial and spatio-temporal models with R-INLA. Spatial and Spatio-temporal Epidemiology 7, 39–55.

Boulieri, A., Liverani, S., Hoogh, K. de, Blangiardo, M., 2017. A space–time multivariate Bayesian model to analyse road traffic accidents by severity. Journal of the Royal Statistical Society Series A 180 (1), 119–139.

Brijs, T., Karlis, D., Wets, G., 2008. Studying the effect of weather conditions on daily crash counts using a discrete time-series model. Accident Analysis and Prevention 40 (3), 1180–1190.

Czado, C., Gneiting, T., Held, L., 2009. Predictive model assessment for count data. Biometrics 65 (4), 1254–1261.

Dawid, A.P., 1984. Present position and potential developments: some personal views: statistical theory: the prequential approach. Journal of the Royal Statistical Society Series A 147 (2), 278–292.

Eckley, D.C., Curtin, K.M., 2013. Evaluating the spatiotemporal clustering of traffic incidents. Computers, Environment and Urban Systems 37, 70–81.

El-Basyouny, K., Barua, S., Islam, M.T., 2014. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis and Prevention 73, 91–99.

Erdogan, S., 2009. Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. Journal of Safety Research 40 (5), 341–351.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Golob, T.F., Recker, W.W., 2003. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. Journal of Transportation Engineering 129 (4), 342–353.

Guo, F., Wang, X., Abdel-Aty, M.A., 2010. Modeling signalized intersection safety with corridor-level spatial correlations. Accident Analysis and Prevention 42 (1), 84–92.

Hu, S., Ivan, J.N., Ravishanker, N., Mooradian, J., 2013. Temporal modeling of highway crash counts for senior and non-senior drivers. Accident Analysis and Prevention 50, 1003–1013.

Iowa Community Indicators Program, 2016. Iowa Community Indicators Program (ICIP). http://www.icip.iastate.edu/ (accessed 2.20.17).

Iowa Department of Transportation, 2016. Vehicle-miles traveled (VMT). http://www.iowadot.gov/maps/msp/vmt/vmt.html (accessed 10.20.16).

Iowa Environmental Mesonet, 2017. Iowa Environmental Mesonet (IEM). https://mesonet.agron.iastate.edu/ (accessed 2.20.17).

Jiang, X., Abdel-Aty, M., Alamili, S., 2014. Application of Poisson random effect models for highway network screening. Accident Analysis and Prevention 63, 74–82.

Kilamanua, W., Xia, J., Caulfieldb, C., 2011. Analysis of spatial and temporal distribution of single and multiple vehicle crash in Western Australia: a comparison study. 19th International Congress on Modelling and Simulation, Perth, Australia.

Knapp, K.K., Smithson, L.D., Khattak, A.J., 2000. The mobility and safety impacts of winter storm events in a freeway environment. Mid-Continent Transoprtation Symposium, Ames, IA, USA.

Levine, N.E.D., Kim, K.E., Nitz, L.H., 1995. Spatial analysis of Honolulu motor vehicle crashes: I. spatial patterns. Accident Analysis and Prevention 27 (5), 663–674.

Lindgren, F., Rue, H. avard, 2015. Bayesian spatial modelling with R-INLA. Journal of Statistical Software 63 (19), 1–26.

Liu, C., Gyawali, S., Sharma, A., Smaglik, E., 2015. A methodological approach for spatial and temporal analysis of red light running citations and crashes: a case-study in Lincoln, Nebraska. Transportation Research Board 94th Annual Meeting, Washington, D.C., USA.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37 (1), 35–46.

Malyshkina, N. V., Mannering, F.L., 2010. Zero-state Markov switching count-data models: an empirical assessment. Accident Analysis and Prevention 42 (1), 122–130.

Malyshkina, N. V, Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accident Analysis and Prevention 41 (2), 217–226.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1–16.

Martins, T.G., Simpson, D., Lindgren, F., Rue, H., 2013. Bayesian computing with INLA: new features. Computational Statistics and Data Analysis 67, 68–83.

Matkan, A., Mohaymany, A., 2013. Detecting the spatial–temporal autocorrelation among crash frequencies in urban areas. Canadian Journal of Civil Engineering 40 (3), 195–203.

Maze, T.H., Agarwal, M., Burchett, G., 2005. Whether weather matters to traffic demand, traffic safety, and traffic flow. Transportation Research Record 1948, 170–176.

Miaou, S.-P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping a space-time modeling approach. Journal of Transportation and Statistics 6 (1), 33–57.

Mitra, S., Washington, S., Schalkwyk, V., 2007. Important omitted spatial variables in safety models: understanding contributing crash causes at intersections. Transportation Research Board 86th Annual Meeting, Washington, D.C., USA.

MRC Biostatistics Unit, 2004. DIC: Deviance Information Criteria. https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-dic/ (accessed 8.23.17).

Mullahy, J., 1997. Heterogeneity, excess zeros, and the structure of count data models. Journal of Applied Econometrics 12 (3), 337–350.

Noland, R.B., Klein, N.J., Tulach, N.K., 2013. Do lower income areas have more pedestrian casualties? Accident Analysis and Prevention 59, 337–345.

Pettit, L.I., 1990. The conditional predictive ordinate for the normal distribution. Journal of the Royal Statistical Society Series B 52 (1), 175–184.

Plummer, M., 2008. Penalized loss functions for Bayesian model comparison. Biostatistics 9 (3), 523–539.

Ponicki, W.R., Gruenewald, P.J., Remer, L.G., 2013. Spatial panel analyses of alcohol outlets and motor vehicle crashes in California: 1999-2008. Accident Analysis and Prevention 55, 135–143.

Pulugurtha, S.S., Sambhara, V.R., 2011. Pedestrian crash estimation models for signalized intersections. Accident Analysis and Prevention 43 (1), 439–446.

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis and Prevention 40 (4), 1486–1497.

Quddus, M.A., 2008. Time series count data models: an empirical application to traffic accidents. Accident Analysis and Prevention 40 (5), 1732–1741.

R Core Team, 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roos, M., Held, L., 2011. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. Bayesian Analysis 6 (2), 259–278.

Rue, H., Martino, S., Chopin, N., 2009. Appoximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B 71 (2), 319–392.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43 (5), 1666–1676.

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97 (1), 246–273.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B 64 (4), 583–639.

Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society Series B 39 (1), 44–47.

Sturtz, S., Ligges, U., Gelman, A., 2005. R2OpenBUGS: a package for running OpenBUGS from R. Journal of Statistical Software 12 (3), 1–16.

Sukhai, A., Jones, A.P., Love, B.S., Haynes, R., 2011. Temporal variations in road traffic fatalities in South Africa. Accident Analysis and Prevention 43 (1), 421–428.

Truong, L.T., Kieu, L.-M., Vu, T.A., 2016. Spatiotemporal and random parameter panel data models of traffic crash fatalities in Vietnam. Accident Analysis and Prevention 94, 153–161.

U.S. Bureau of Economic Analysis, 2016. Personal income sumamry: personal income, population, per capita personal income. https://www.bea.gov/iTable/iTable.cfm?reqid=70&step=30&isuri=1&7022=20&7023=7&7024=non-industry&7033=-1&7025=4&7026=xx,19000&7027=2015,2014,2013,2012,2011,2010&7001=720&7028=3&7030=0&7031=19000&7040=-1&7083=levels&7029=20&7090=70#reqid=70&step=30&isuri=1& (accessed 2.20.17).

Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident Analysis and Prevention 43 (6), 1979–1990.

Wang, C., Quddus, M.A., Ison, S.G., 2009. Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. Accident Analysis and Prevention 41 (4), 798–808.

Wang, C., Quddus, M., Ison, S., 2013. A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. Transportmetrica A 9 (2), 124–148.

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis and Prevention 38 (6), 1137–1150.

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

Xie, K., Wang, X., Ozbay, K., Yang, H., 2014. Crash frequency modeling for signalized intersections in a high-density urban road network. Analytic Methods in Accident Research 2, 39–51.

Yannis, G., Antoniou, C., Papadimitriou, E., 2011. Autoregressive nonlinear time-series modeling of traffic fatalities in Europe. European Transport Research Review 3 (3), 113–127.

Zeng, Q., Huang, H., 2014. Bayesian spatial joint modeling of traffic crashes on an urban road network. Accident Analysis and Prevention 67, 105–112.

CHAPTER 3. USING THE MULTIVARIATE SPATIO-TEMPORAL BAYESIAN MODEL TO EXPLORE THE TRAFFIC CRASH FREQUENCY TREND IN LONG TERM

A paper published on the Analytic Methods in Accident Research

Abstract

Unobserved heterogeneity across space, time, and crash type is often non-negligible in crash frequency modeling. When multiple crash type with spatial and temporal features are analyzed, multivariate spaio-temporal Bayesian models should be considered. For this study, we analyzed the yearly county-level fatal, major injury, and minor injury crashes in Iowa from 2006 to 2015 using a multivariate spatio-temporal Bayesian model. The model adopts a multivariate spatial structure, a multivariate temporal structure, and a multivariate spatio-temporal interaction structure to account for possible correlations across injury severities over space, time, and spatio-temporal interaction, respectively. Income and weather indicators were found to be significant in the presence of vehicle miles traveled and unemployment rate. Both spatial and temporal effects were found to be important, and they played nearly the same roles for all three crash types in the studied dataset. Counties located in the north and southwest Iowa were found to tend to have fewer crashes than the remaining counties. All three crash types generally showed descending trends from 2006 to 2015. They also had significantly positive correlations between each other in space but not in time. The crude crash rates and the predicted crash rates were generally consistent for major injury and minor injury crashes but not for low-count fatal crashes. High-risk counties were identified using the posterior expected rank by the predicted crash cost rate, which was more able to truly represent the underlying traffic status than the rank by the crude crash cost rate.

Keywords: multivariate spatio-temporal, Bayesian, crash frequency, posterior expected rank, crash cost rate

3.1 Introduction

Traffic crashes have been one of the major sources of fatalities and injuries in the United States. Crash frequency analysis is often used to identify key factors influencing the propensity of crashes, which is important for policymakers as they propose interventions to prevent road traffic crashes. However, unobserved heterogeneity is often an issue in crash frequency analysis, because many crash-related elements are often unavailable. Neglecting unobserved heterogeneity may produce biased and inefficient results (Mannering et al., 2016).

Unobserved heterogeneity may come from many sources. Crashes are usually classified into multiple types by different criteria, and their underlying correlations may produce some unobserved heterogeneity across observations when they are analyzed simultaneously (Mannering et al., 2016; Mannering and Bhat, 2014). Thus, multivariate models, such as the multivariate Poisson log-normal (MVPLN) model, are often adopted (Aguero-Valverde and Jovanis, 2010; El-Basyouny et al., 2014; El-Basyouny and Sayed, 2009; Ma et al., 2008; Zhao et al., 2017). In addition, crash frequency data are always aggregated over space and time, which may also produce unobserved heterogeneity, as crashes that occur close in space or time are very likely to share some unobserved characteristics (Lord et al., 2005; Lord and Mannering, 2010; Mannering et al., 2016; Mannering and Bhat, 2014; Savolainen et al., 2011). Previous studies have shown that spatial correlations of traffic crashes may exist across states/provinces (Erdogan, 2009; Truong et al., 2016), counties (Aguero-Valverde and Jovanis, 2006; Eckley and Curtin, 2013; Song et al., 2006), census tracts (Wang and Kockelman, 2013), traffic analysis zones (Matkan and Mohaymany, 2013), intersections (Ahmed and Abdel-Aty, 2015; Liu et al., 2015) and segments (Aguero-Valverde, 2011; Aguero-Valverde and Jovanis, 2008; Jiang et al.,

2014; Wang et al., 2011, 2009; Zeng and Huang, 2014). The similarity of economy, culture, land use, weather, traffic laws, and driving behavior within a given region may explain the spatial correlations in traffic crashes. When multiple crash types with spatial correlations need to be analyzed, multivariate spatial models have been proved to be more powerful than univariate spatial models, as multivariate spatial models can account for correlations across crash types in space in addition to spatial correlations (Aguero-Valverde, 2013; Aguero-Valverde et al., 2016; Barua et al., 2016; Miaou and Song, 2005; Song et al., 2006; Wang and Kockelman, 2013). Temporal correlations of traffic crashes may exist across year (Andrey, 2010; Brijs et al., 2008; El-Basyouny et al., 2014; Matkan and Mohaymany, 2013; Wang et al., 2011; Wang and Abdel-Aty, 2006; Yannis et al., 2011), month (Hu et al., 2013; Quddus, 2008b), week (Kilamanua et al., 2011; Liu et al., 2015; Sukhai et al., 2011), and day (Brijs et al., 2008). Temporal correlations occur because many traffic-related factors, such as driver behavior, economy, weather, environment, law, and travel demand, often exhibit some temporal features. Similarly, when multiple crash types with temporal correlations need to be analyzed, multivariate temporal models should be considered, as they can account for correlations across crash types in time in addition to temporal correlations (Michalaki et al., 2016; Serhiyenko et al., 2014).

Crashes often have both spatial and temporal features. When only one crash type is analyzed, the univariate spatio-temporal modeling has been proved in some studies to be superior (Aguero-Valverde and Jovanis, 2006; Liu and Sharma, 2017; Miaou et al., 2003; Truong et al., 2016). When multiple crash types need to be analyzed, a multivariate spatio-temporal model may be needed. Ma et al. (2017) used the bivariate spatio-temporal model to analyze the daily non-injury and injury crash rates on 100 roadway segments of I70 in one year at the micro level, and Boulieri et al. (2017) used the bivariate spatio-temporal model to analyze the yearly low severity

and high severity accidents of 7932 electoral wards in England from 2005-2013 considering only vehicle miles traveled (VMT). Both studies showed the superiority of the bivariate spatio-temporal model to the univariate spatio-temporal model in terms of goodness of fit.

In this study, we used the multivariate spatio-temporal Bayesian model to analyze the yearly county-level fatal, major injury, and minor injury crash frequencies in Iowa. The goal of this study was to accurately identify the long-term effects of economy and weather on crash frequency in Iowa and to explore the spatial and temporal correlations of crashes. Additionally, the counties were ranked to identify high-risk areas for safety improvement programs, as funding available for safety improvements are often limited and proper ranking can significantly influence the appropriate distribution of safety funding toward areas with more critical needs. Raw crash data-based ranking is easy to use but crude and inefficient (Miaou and Song, 2005). In Bayesian cases, one statistical ranking method is the posterior expected rank (PER), i.e. the posterior mean of the rank by ranking indicators (Miaou and Song, 2005). When rankings are the main interest, the PER method is recommended (Shen and Louis, 1998). The most common ranking indicator is crash rate, but crash rate considering crash cost by injury severity, called the "crash cost rate" in the following analysis, is strongly recommended when injury severity and associated costs are the main concerns (Miaou and Song, 2005). Thus, the PER of the crash cost rate would be used to rank the studied areas based on the predicted results of the multivariate spatio-temporal Bayesian model in this study.

3.2 Data Description

Traffic crash data of Iowa's 99 counties from 2006 to 2015 were obtained from the Iowa Department of Transportation. Crashes were divided into five categories by severity: fatal, major injury, minor injury, possible injury/unknown, and property damage only. Fatal crashes, major injury crashes, and minor injury crashes were analyzed in this study, as these three types of

crashes often lead to significant economic loss and casualties. VMT data for each county in each year from 2006 to 2015 were downloaded from the website of the Iowa Department of Transportation (2016). In addition, unemployment rate data were downloaded from the website of Iowa Community Indicators Program (2016), and per capita personal income data were downloaded from the website of the U.S. Bureau of Economic Analysis (2016) of the U.S. Department of Commerce. Meanwhile, rainfall, snowfall, and the number of days with the minimum temperature higher than 32°F were downloaded from the website of the Iowa Environmental Mesonet (2017). These weather data are collected based on the daily climate observations from the National Weather Service's Cooperative Observer Program. A summary of the variables is given in Table 3-1. All three crash types have over-dispersion, as their variances are much larger than their means. Additionally, the highest correlation among the covariates was -0.338 (between snowfall and TH32). Thus, no explanatory variables showed strong positive or negative correlations.

Table 3-1 *Descriptive statistics of collected variables*

Variables	Min.	Median	Mean	Max.	Std. Error
Fatal crash	0	2	3.383	35	3.818
Major injury crash	0	8	13.680	245	22.042
Minor injury crash	1	23	49.060	894	93.742
VMT (1,000,000 miles)	0.047	0.186	0.320	4.215	0.487
Unemployment rate (%)	2.000	4.600	4.846	10.200	1.347
Income ($10,000)	2.247	3.877	3.877	6.464	0.666
Rainfall (inch)	17.850	38.610	38.390	64.99	8.570
Snowfall (inch)	0	35	34.560	85.100	14.377
TH32 (days)	174	221	222.6	272	15.733

Note: VMT, vehicle miles traveled; TH32, number of days with minimum temperature higher than 32°F.

The Pearson correlations of three types of crashes are shown in Table 3-2. All three crash types were highly positively correlated. That is, locations where many fatal/major injury/minor injury crashes were observed likely also had many crashes of the other two types.

Table 3-2 *Pearson correlation matrix of crashes*

	Fatal crash	Major injury crash
Major injury crash	0.837	
Minor injury crash	0.835	0.971

The yearly county-level average fatal, major injury, and minor injury crash counts in Iowa are shown in Figure 3-1. A cluster of high fatal crash frequencies can be observed in the central counties around the dark red-shaded area, where the largest city in Iowa, Des Moines, is located. A cluster of low crash frequencies can be observed in the northern and southwestern regions of Iowa (lightly shaded areas). A cluster of comparatively higher numbers of major injury crashes can also be observed in the central counties. However, no obvious clustering trends can be observed for minor injury crashes. Next, spatial correlations of crashes are examined statistically.

Moran's *I* statistic is commonly used to test spatial correlations in traffic crash analyses (Guo et al., 2010; Quddus, 2008; Xie et al., 2014; Zeng and Huang, 2014). The global Moran's *I* is defined as (Anselin, 1988):

$$I = \frac{n\sum_i\sum_i \omega_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_{i\neq j}\omega_{ij}\sum_i(y_i-\bar{y})^2} \quad (3.1)$$

where n is the total number of observations, y_i and y_j are the values of observation i and observation j, $\bar{y}$ is the average value of observations, and ω_{ij} is the spatial weight between observations i and j.

Negative Moran's *I* values indicate negative spatial autocorrelation, positive Moran's *I* values indicate positive spatial autocorrelation, and zero indicates no spatial autocorrelation. The *z*-score of Moran's *I* shows if the spatial autocorrelation is significant.

The global Moran's *I* statistics of crashes in each year from 2006 to 2015 were calculated using the "spdep" package (Bivand and Piras, 2015) in the R platform (R Core Team, 2016) with

queen continuity spatial weights, where counties with a shared border or vertex were considered neighbors. When areas were neighbors, the spatial weights were 1; otherwise, they were 0. The results are shown in Table 3-3.

Fatal crashes and major injury crashes showed significant spatial autocorrelations in seven and six out of 10 years, respectively, at a 95% confidence level, but minor injury crashes did not show any significant spatial autocorrelations at a 95% confidence level in any year. Additionally, the *P*-values of fatal crashes and major injury crashes were much smaller than those for minor injury crashes. Thus, fatal crashes and major injury crashes were highly likely to be spatially correlated as compared to minor injury crashes. These trends may be site-specific. As an example, Aguero-Valverde and Jovanis (2006) found injury crashes to have a significant spatial correlation and fatal crashes to not be significantly correlated in counties of Pennsylvania. Although minor injury crashes did not show significant spatial autocorrelations, it does not mean the absence of spatial autocorrelation for minor injury crashes; they may still have weak spatial correlations. The different strengths of spatial autocorrelations imply that the three crash types may have different spatial model parameters.

The temporal correlation was not directly tested, as there were only 10 time points for each crash type. However, as Figure 3-2 shows by the yearly state-level counts of all three crashes from 2006 to 2015, they all generally exhibited descending trends, with some dipping and heaving, and different descending rates.

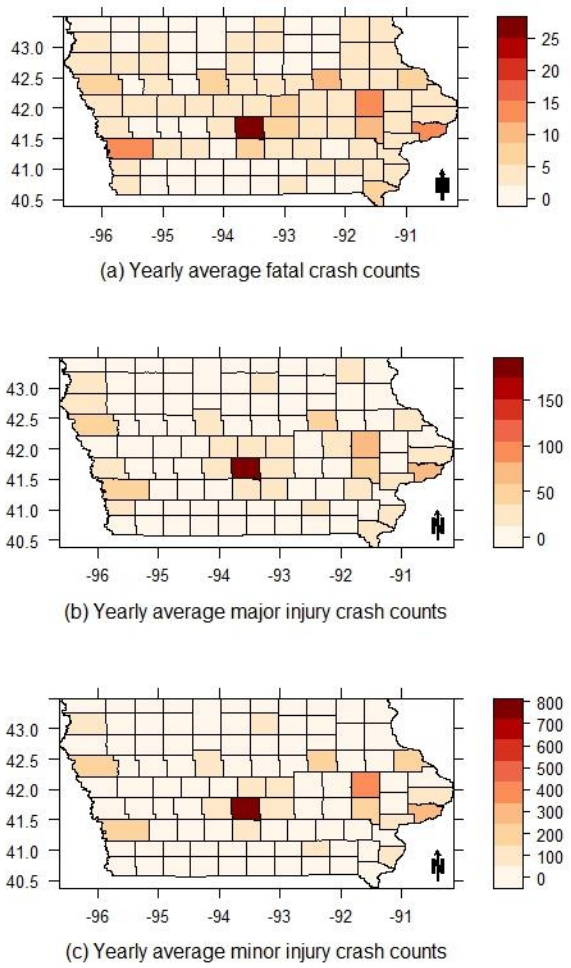

Figure 3-1 *County-level yearly average fatal, major injury, and minor injury crash counts (2006-2015)*

Table 3-3 *Global Moran's I statistics of crash counts in each year*

Year	Fatal crash		Major injury crash		Minor injury crash	
	Moran's *I*	*P*-value	Moran's *I*	*P*-value	Moran's *I*	*P*-value
2006	1.986	0.024*	1.752	0.041*	0.971	0.166
2007	2.091	0.018*	1.555	0.060	1.141	0.127
2008	1.520	0.064	0.688	0.246	0.871	0.192
2009	1.661	0.048*	1.181	0.119	0.764	0.222
2010	2.486	0.006*	1.586	0.056	1.106	0.134
2011	1.919	0.027*	1.883	0.031*	1.101	0.136
2012	1.240	0.108	2.017	0.022*	1.108	0.134
2013	2.387	0.009*	2.218	0.013*	1.555	0.060
2014	1.241	0.107	1.877	0.030*	1.252	0.105
2015	2.300	0.011*	2.770	0.003*	1.468	0.071

Note: * significant at $P = 0.05$.

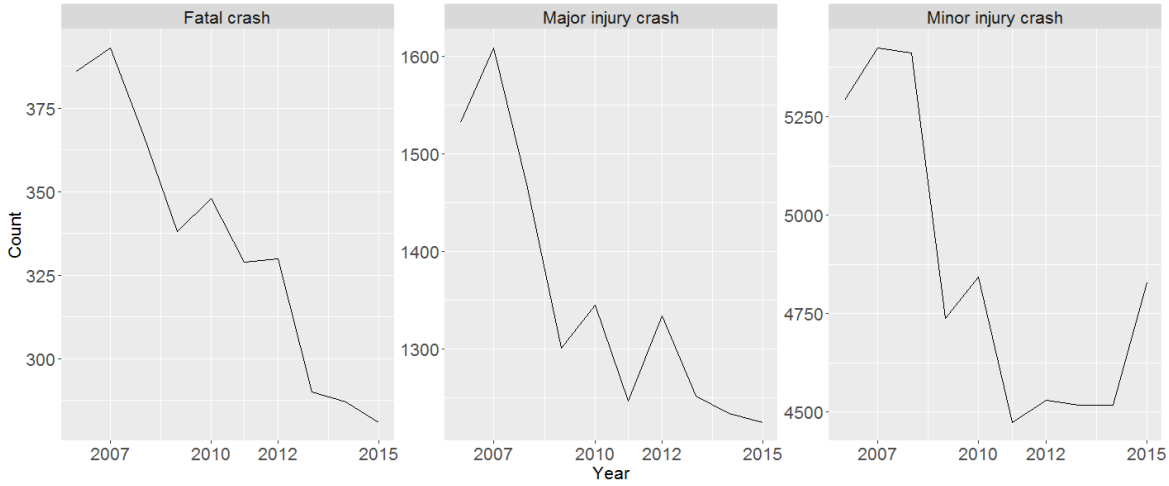

Figure 3-2 *Iowa state-level yearly crash counts (2006-2015)*

3.3 Methodology

3.3.1 Statistical Framework

The statistical framework used a Bayesian hierarchical architecture, including both the spatial, temporal, and spatio-temporal interaction components. The statistical model is presented in equations (2) and (3) (Ma et al., 2017):

$$y_{stk} \sim Poisson(\lambda_{stk}) \tag{3.2}$$

$$\log(\lambda_{stk}) = \alpha_k + X_{st}^{T} * \beta_k + \upsilon_{sk} + \nu_{sk} + \varphi_{tk} + \theta_{tk} + \eta_{stk} \tag{3.3}$$

where s is the space number, i.e. county number in this case, 1,2,… 99; t is the time point, i.e. year number in this case, 1 (2006), 2 (2007), …, 10 (2015); k is the crash injury severity number, 1 (fatal crash), 2 (major injury crash), 3 (minor injury crash); y_{stk} is the crash count of injury severity k of space s in time t; λ_{stk} is the mean crash frequency of injury severity k of space s in time t; α_k is the intercept term of crash type k; $\beta_k (= \beta_{k1}, \beta_{k2}, \dots, \beta_{km})$, is the m-dimensional regression coefficient vector of crash type k, and m is the number of covariates, i.e. 6 in this case; $X_{st} (= X_{st1}, X_{st2}, \dots, X_{stm})$ is the m-dimensional covariate vector of space s in time t; υ_{sk} is the structured spatial random effect of crash type k in space s; ν_{sk} is the unstructured

spatial random effect of crash type k in space s; φ_{tk} is the structured temporal random effect of crash type k in time t; θ_{tk} is the unstructured temporal random effect of crash type k in time t; and η_{stk} is the spatio-temporal interaction effect of crash type k in space s and time t.

The spatial component of each observation was consisted of two parts: $\upsilon_{sk} + \nu_{sk}$, while the temporal component also consisted of two parts: $\varphi_{tk} + \theta_{tk}$.

3.3.1.1 Spatial component

3.3.1.1.1 Univariate spatial model

The spatial component of each observation, $\upsilon_{sk} + \nu_{sk}$, was assumed to follow the Besag-York-Mollie (BYM) model (Besag et al., 1991). The BYM model has been proved to be powerful in traffic crash analysis (Aguero-Valverde and Jovanis, 2006; Boulieri et al., 2017; Ma et al., 2017; Wang et al., 2013; Xie et al., 2014). For the BYM model, the structured spatial effect, υ_{sk}, is modeled using an intrinsic conditional autoregressive (ICAR) structure, and the unstructured spatial effect, ν_{sk}, follows a normal distribution.

$$\upsilon_{sk}|\upsilon_{-sk} \sim N\left(\frac{\sum_{i\in N(s)}\upsilon_{ik}}{\#N(s)}, \frac{{\sigma^2}_{\upsilon}{}^{k}}{\#N(s)}\right) \tag{3.4}$$

$$\nu_{sk} \sim N\left(0, {\sigma^2}_{\nu}{}^{k}\right) \tag{3.5}$$

where $N(s)$ are the neighbors of space s; $\#N(s)$ are the number of neighbors of space s and ${\sigma^2}_{\upsilon}{}^{k}$ and ${\sigma^2}_{\nu}{}^{k}$ are two independent variances of crash injury severity k in space.

Two counties adjacent to each other were considered to be neighbors; otherwise, they were not neighbors. The ICAR part accounted for unobserved heterogeneity produced by possible spatial correlations between counties, and the unstructured part was responsible for county-specific heterogeneity. In the univariate BYM (UBYM) model, both the structured and unstructured spatial effects across crash injury severities were assumed to be independent for each observation.

3.3.1.1.2 Multivariate spatial model

The multivariate BYM (MBYM) model, shown in equations (6) and (7), is the extension of the BYM model in multivariate cases (Boulieri et al., 2017; Ma et al., 2017):

$$\upsilon_{s.}|\upsilon_{(i\neq s).} \sim N(\frac{\sum_{i\in N(s)}\upsilon_{i.}}{\#N(s)}, \frac{\Sigma_{\upsilon}}{\#N(s)}) \tag{3.6}$$

$$\nu_{s.} \sim N(0, \Sigma_{\nu}) \tag{3.7}$$

where $\upsilon_{s.} = (\upsilon_{s1}, \upsilon_{s2}, \upsilon_{s3})$ is the 3-dimensional structured spatial random effects of space s; $\nu_{s.} = (\nu_{s1}, \nu_{s2}, \nu_{s3})$ is the *3*-dimensional unstructured spatial random effects of space s; $N(s)$ are the neighbors of space s; $\#N(s)$ is the number of neighbors of space s; and Σ_{υ} and Σ_{ν} are two independent $3 * 3$ variance-covariance matrices in space.

The MBYM model consisted of a multivariate ICAR component and a multivariate normal (MVN) component. Different from the univariate BYM model, both the structured and unstructured spatial random effects of each observation are correlated across crash injury severities. Thus, they could account for possible unobserved heterogeneity across crash injury severities in space for each observation.

3.3.1.2 Temporal component

3.3.1.2.1 Univariate temporal Model

The structured temporal effect of each observation, φ_{tk}, was modeled with the 1st order random walk (RW1) structure. The unstructured temporal effect of each observation, θ_{tk}, followed a normal distribution. This temporal component was still called the RW1 model in the following analysis, although it actually consisted of an RW1 model and a random error term. The RW1 model was a special case of applying the ICAR model shown in Equation (3.4) in time. In the univariate RW1 model, both the structured and unstructured temporal effects across crash injury severities were assumed to be independent for each observation.

$$\varphi_{tk}|\varphi_{(-tk)}\sim\begin{cases} N\left(\varphi_{(t+1)k},{\sigma^2}_{\varphi}{}^{k}\right) & for\ t=1 \\ N\left(\frac{\varphi_{(t-1)k}+\varphi_{(t+1)k}}{2},\frac{{\sigma^2}_{\varphi}{}^{k}}{2}\right) & for\ t=2,3,\ldots,9 \\ N\left(\varphi_{(t-1)k},{\sigma^2}_{\varphi}{}^{k}\right) & for\ t=10 \end{cases} \quad (3.8)$$

$$\theta_{tk}\sim N(0,{\sigma^2}_{\theta}{}^{k}), \quad (3.9)$$

where ${\sigma^2}_{\varphi}{}^{k}$ and ${\sigma^2}_{\theta}{}^{k}$ are two independent variances of crash injury severity k in time.

3.3.1.2.2 Multivariate temporal model

The multivariate RW1 (MRW1) model is the extension of the RW1 model into multivariate cases (Boulieri et al., 2017; Ma et al., 2017) and is defined as:

$$\varphi_{t.}|\varphi_{(-t).}\sim\begin{cases} N\left(\varphi_{(t+1).},\Sigma_{\varphi}\right) & for\ t=1 \\ N\left(\frac{\varphi_{(t-1).}+\varphi_{(t+1).}}{2},\frac{\Sigma_{\varphi}}{2}\right) & for\ t=2,3,\ldots,9 \\ N\left(\varphi_{(t-1).},\Sigma_{\varphi}\right) & for\ t=10 \end{cases} \quad (3.10)$$

$$\theta_{t.}\sim N(0,\Sigma_{\theta}) \quad (3.11)$$

where $\varphi_{t.}=(\varphi_{t1},\varphi_{t2},\varphi_{t3})$ is the *3*-dimensional structured temporal random effects of time *t*; $\theta_{t.}=(\theta_{t1},\theta_{t2},\theta_{t3})$ is the *3*-dimensional unstructured temporal random effects of time t; and Σ_{φ} and Σ_{θ} are two independent $3*3$ variance-covariance matrices in time.

The MRW1 model consists of a multivariate RW1 component and an MVN component. Different from the univariate RW1 model, both the structured and unstructured temporal random effects of each observation were also correlated across crash injury severities. Thus, they could account for possible unobserved heterogeneity across crash injury severities in time for each observation.

3.3.1.3 Spatio-Temporal component

The spatio-temporal interaction effect of each observation across crash injury severities, $\eta_{(st).}$, was used to account for unobserved heterogeneity not explained by other components.

$$\eta_{(st).} \sim N(0, \Sigma_\eta) \tag{3.12}$$

where $\eta_{(st).} = (\eta_{st1}, \eta_{st2}, \eta_{st3})$ is the *3*-dimensional spatial-temporal interaction effect of space s in time t; and Σ_η is a $3 * 3$ variance-covariance matrix.

The structure of Σ_η could account for the rest possible correlations across crash injury severities for each observation. To select an appropriate model, there were four models built in this study, as shown in Table 3-4.

Table 3-4 *Summary of models developed in this study*

No	Model	Spatial component	Temporal component	Spatio-temporal component
1	$S_{BYM}T_{RW1}$	BYM	RW1	MVN
2	$S_{BYM}T_{MRW1}$	BYM	MRW1	MVN
3	$S_{MBYM}T_{RW1}$	MBYM	RW1	MVN
4	$S_{MBYM}T_{MRW1}$	MBYM	MRW1	MVN

Note: BYM, Besag-York-Mollie; MBYM, multivariate BYM; RW1, 1st order random walk; MRW1, multivariate RW1; MVN, multivariate normal.

3.3.2 Priors Settings

All four models were built within the Bayesian hierarchical structure. The priors of parameters were set as:

$$\alpha_k \sim Uniform(-\infty, +\infty) \tag{3.13}$$

$$\beta_{.j} \sim N(0, \Sigma_\beta) \tag{3.14}$$

$$\sigma^2{}_\upsilon{}^k, \sigma^2{}_\nu{}^k, \sigma^2{}_\varphi{}^k, \sigma^2{}_\theta{}^k \underset{\sim}{iid} Inverse - Gamma(1, 0.0005) \tag{3.15}$$

$$\Sigma_\beta, \Sigma_\upsilon, \Sigma_\nu, \Sigma_\varphi, \Sigma_\theta, \Sigma_\eta \underset{\sim}{iid} Inverse - Wishart(I_3, 3) \tag{3.16}$$

where $k(= 1,2,3)$ is the crash injury severity number; j(= 1,2,3,4,5,6) is the covariate number; $\beta_{.j}(= (\beta_{1j}, \beta_{2j}, \beta_{3j})^T)$ is the regression coefficient vector of the jth covariate across crash injury severities; Σ_β is the variance-covariance matrix of the regression coefficients of

covariates across crash injury severities; ${\sigma^2}_{\upsilon}{}^{k}$, ${\sigma^2}_{\nu}{}^{k}$, ${\sigma^2}_{\varphi}{}^{k}$, and ${\sigma^2}_{\theta}{}^{k}$ are independent variances of the structured spatial effects, unstructured spatial effects, structured temporal effects, and unstructured temporal effects in univariate models of crash injury severity k, respectively; Σ_{υ}, Σ_{ν}, Σ_{φ}, and Σ_{θ} are independent variance-covariance matrices of the structured spatial effects, unstructured spatial effects, structured temporal effects, unstructured temporal effects in multivariate models, respectively; Σ_{η} is the variance-covariance matrix of spatio-temporal interaction effects; and I_3 is the 3-dimension identity matrix.

The regression coefficients ($\beta_{.j}$) were given a multivariate normal prior to accommodate their possible correlations across crash severities. A flat prior was set for intercept terms ($\alpha_{.}$) to ensure identifiability of the model (MRC Biostatistics Unit, 2004). All the variances were set to have a minimally informative prior of an inverse Gamma distribution $Inverse - Gamma(1,0.0005)$ (Blangiardo et al., 2013), which also had been proved to be effective in our former study (Liu and Sharma, 2017). All the variance-covariance matrices were assigned an inverse-Wishart prior with the scale matrix being an identity matrix and the degree of freedom being 3 to provide weakly information.

3.3.3 Initial Values Settings

OpenBUGS, one popular Bayesian analysis software using Markov chain Monte Carlo (MCMC) simulation to estimate posterior distributions of parameters, was used in this study (Lunn et al., 2009). To start MCMC simulations, initial values have to be given for each unknown parameter and latent variables. Good initial values help MCMC simulation converge quickly to the true distributions of parameters, whereas bad initial values may make MCMC simulation converge slowly and even become stuck at some data points. When initial values are not given, OpenBUGS randomly generates initial values, which usually works after long MCMC

iterations. However, that was not the case in this study, as some variances and variance–covariance matrices of some chains were often stuck at some points using the randomly generated initial values of OpenBUGS. Thus, the posterior distributions of parameters did not converge well. Finally, we ran the MCMC simulation twice. The results of the first MCMC simulation were used as the initial values for the second MCMC simulation, which converged very well. Based on the first MCMC simulation result, initial values of the second MCMC simulation were set as: $\sigma^2{}_{\upsilon} = \left(\frac{1}{5}, \frac{1}{1500}, \frac{1}{1500}\right)$, $\sigma^2{}_{\nu} = \left(\frac{1}{1800}, \frac{1}{5}, \frac{1}{600}\right)$, $\sigma^2{}_{\varphi} = \left(\frac{1}{2000}, \frac{1}{2000}, \frac{1}{2000}\right)$, $\sigma^2{}_{\theta} = \left(\frac{1}{2000}, \frac{1}{2000}, \frac{1}{2000}\right)$, $\Sigma_{\upsilon}{}^{-1} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix}$, $\Sigma_{\nu}{}^{-1} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix}$, $\Sigma_{\varphi}{}^{-1} = \begin{bmatrix} 11 & 0 & 0 \\ 0 & 11 & 0 \\ 0 & 0 & 11 \end{bmatrix}$, and $\Sigma_{\theta}{}^{-1} = \begin{bmatrix} 18 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 18 \end{bmatrix}$.

3.3.4 Model Checking and Comparison

Deviance Information Criteria (DIC) is a generalized version of Akaike Information Criterion (AIC) for evaluating hierarchical models (Spiegelhalter et al., 2002). The deviance is defined as $D(\theta) = -2log\big(p(y|\theta)\big)$, where y is the data, θ is the unknown parameters, and $p(y|\theta)$ is the likelihood function. DIC is defined as (Spiegelhalter et al., 2002):

$$DIC = D(\bar{\theta}) + 2pD = \overline{D} + pD \tag{3.17}$$

where $\bar{\theta}$ is the posterior mean of the parameters; $D(\bar{\theta})$ is the deviance at the posterior mean of the parameters, a measure of data fit; pD is the effective number of the model, a measure of complexity computed as the difference between $\overline{D}$ and $D(\bar{\theta})$; and $\overline{D}$ is the mean of the sampled deviances from MCMC simulations.

Bayesian models with smaller DIC values are desired. Models with smaller DIC values are expected to perform better. Roughly, differences of more than 10 might definitely rule out

the model with the higher DIC, differences between 5 and 10 are substantial, and differences less than 5 might mean that the models are not significantly different (MRC Biostatistics Unit, 2004).

Although DIC can be used for model comparison, it cannot evaluate the quality of fit of the model and observed data. The posterior predictive density is often used for checking the assumptions of a model and its goodness-of-fit. Assume there is a test statistic $D(y, \theta)$, which is a summary function. If the model is correct, we can use the posterior predictive distribution to generate replicated values y^{rep}, which are expected to be close to the observed data y^{rep}. The test statistic is used to check the assumption under investigation and measure discrepancies between the data and the model (Gelman et al., 1996). Based on the posterior predictive distribution, the posterior predictive p-value is defined as (Meng, 1994)

$$Posterior\ p-value\ =\ P(D(y^{rep},\theta) > D(y^{obs},\theta)|y^{obs}) \tag{3.18}$$

P-values around 0.5 indicate that the distributions of the replicated and observed data are close, whereas values close to zero or one indicate differences between them (Gelman et al., 1996). In this study, the mean values of crashes would be taken as the test statistic, as mean is the major parameter for a Poisson model.

3.3.5 Random Effects Analysis

3.3.5.1 Spatial fraction analysis

For spatial analysis, one point of interest is to identify the contribution of the structured spatial effects, ${\sigma_v}^2$, over the total marginal spatial variability, ${\sigma_v}^2 + {\sigma_v}^2$ (Boulieri et al., 2017). The spatial fraction of interest is given by

$$frac_v = \frac{{\sigma_v}^2}{{\sigma_v}^2+{\sigma_v}^2} \tag{3.19}$$

When it is close to 1, the structured spatial effects explain most of the variability of the model in space. Otherwise, the unstructured spatial random effects play the main role.

3.3.5.2 Temporal fraction analysis

Similarly, the temporal fraction is defined as the variance of structured temporal effects ${\sigma_\varphi}^2$ over the total marginal temporal variability ${\sigma_\varphi}^2 + {\sigma_\phi}^2$:

$$frac_\varphi = \frac{{\sigma_\varphi}^2}{{\sigma_\varphi}^2 + {\sigma_\theta}^2} \tag{3.20}$$

When it is close to 1, the structured temporal effects explain most of the variability of the model in time. Otherwise, the unstructured temporal random effects play the main role.

3.3.5.3 Spatial and temporal effects comparison

When both the spatial and temporal effects exist, it is also of interest to determine their relative importance. The relative importance of spatial effects is defined as the variance of spatial effects over the marginal variability of spatial and temporal effects:

$$frac_{\frac{S}{S+T}} = \frac{{\sigma_\upsilon}^2 + {\sigma_\nu}^2}{{\sigma_\upsilon}^2 + {\sigma_\nu}^2 + {\sigma_\varphi}^2 + {\sigma_\theta}^2} \tag{3.21}$$

3.3.6 PER by Total Crash Cost Rate

In the "safety improvement candidate location" methods of Iowa (Pawlovich, 2007), the costs of fatal, major injury, and minor injury crashes were set as 200, 100, and 10 units, respectively. They were adopted to calculate the total crash cost rate as shown in equation (22), where crash rate was the crash count per million VMT. The PER using the predicted total crash cost rate as well as the crude rank using the crude total crash cost rates would be computed and compared, respectively. The county ranked as 1st had the largest total crash cost rate.

$$Total\ crash\ cost\ rate = Fatal\ crash\ rate * 200 + Major\ injury\ crash\ rate * 100 + Minor\ injury\ crash\ rate * 10 \tag{3.22}$$

3.4 Results

All fours models were implemented in OpenBUGS in R (R Core Team, 2016) through "R2OpenBUGS" (Sturtz et al., 2005). OpenBUGS uses the Metropolis-Hastings algorithm to

sample data. Three simulation chains were run with 50,000 iterations for each chain, the first 25,000 samples discarded as burn-in and the remaining 25,000 samples retained to get the posterior distributions of parameters with a thinning interval of 5. Thus, 5,000 samples were recorded per chain. On an Intel(R) Xeon(R) CPU at 3.70 GHz with 16 GB random access memory, it took about 3.5 hours to run each model. The trace plots of estimated parameters showed that posterior samples converged well after the burn-in iterations. In addition, the Gelman and Rubin's convergence diagnostic, i.e. potential scale reduction factors of variables, were also calculated to check the convergence of multiple chains (Gelman and Rubin, 1992). The DIC values for the four models listed in Table 3-4 are shown in Table 3-5.

Table 3-5 *DIC values of four models*

No	Model	DIC
1	$S_{BYM}T_{RW1}$	14,330
2	$S_{BYM}T_{MRW1}$	8,519
3	$S_{MBYM}T_{RW1}$	10,970
4	$S_{MBYM}T_{MRW1}$	8,371

Note: DIC, Deviance Information Criteria.

Compared to the $S_{BYM}T_{RW1}$ model, the DIC values of both the $S_{MBYM}T_{RW1}$ and the $S_{BYM}T_{MRW1}$ models were much smaller. In addition, the DIC value of the $S_{MBYM}T_{MRW1}$ model was much smaller than that for the $S_{MBYM}T_{RW1}$ and the $S_{BYM}T_{MRW1}$ models. This implied that unobserved heterogeneity across crash injury severities existed in both space and time, thus the $S_{MBYM}T_{MRW1}$ model was preferred for this study. In addition, the posterior *p*-values of the mean values of fatal, major injury, and minor injury crashes were 0.500, 0.497, and 0.495, respectively, close to 0.5, which meant that the $S_{MBYM}T_{MRW1}$ model matched the data well. Mean and 95% credible interval (CI) values of estimated parameters of the $S_{MBYM}T_{MRW1}$ model are shown in Table 3-6.

Table 3-6 *Estimated parameters of the $S_{MBYM}T_{MRW1}$ model with all covariates*

	Fatal crash		Major injury crash		Minor injury crash	
Variables	Mean	95% CI	Mean	95% CI	Mean	95% CI
Intercept	0.168	(–1.591, 1.630)	2.185*	(1.212, 3.095)	3.028*	(2.505, 3.611)
Income	0.081	(–0.054, 0.228)	–0.036	(–0.119, 0.053)	0.018	(–0.052, 0.085)
Unemployment rate	0.037	(–0.029, 0.102)	–0.055*	(–0.103, –0.009)	–0.062*	(–0.092, –0.033)
Rainfall	–0.005	(–0.013, 0.003)	0.001	(–0.004, 0.006)	0.002	(–0.002, 0.005)
Snowfall	–0.002	(–0.006, 0.003)	–0.001	(–0.004, 0.002)	0.000	(–0.002, 0.001)
TH32	0.001	(–0.005, 0.008)	0.000	(–0.004, 0.004)	0.000	(–0.002, 0.002)
VMT	0.732*	(0.538, 0.903)	0.970*	(0.746, 1.160)	1.132*	(0.893, 1.301)
$\sigma^2{}_v$	0.305	(0.103, 0.644)	0.412	(0.116, 0.903)	0.491	(0.136, 1.088)
$\sigma^2{}_\nu$	0.124	(0.067, 0.203)	0.188	(0.100, 0.299)	0.231	(0.124, 0.362)
$frac_v$	0.681	(0.383, 0.890)	0.651	(0.314, 0.889)	0.643	(0.305, 0.887)
$\sigma^2{}_\varphi$	0.261	(0.081, 0.759)	0.253	(0.078, 0.708)	0.246	(0.078, 0.698)
$\sigma^2{}_\theta$	0.229	(0.074, 0.634)	0.214	(0.071, 0.576)	0.218	(0.072, 0.589)
$frac_\varphi$	0.527	(0.219, 0.816)	0.533	(0.228, 0.820)	0.525	(0.226, 0.809)
$frac_{\frac{S}{S+T}}$	0.487	(0.235, 0.718)	0.573	(0.314, 0.792)	0.618	(0.368, 0.813)
$\sigma^2{}_\eta$	0.047	(0.031, 0.067)	0.029	(0.022, 0.037)	0.018	(0.014, 0.022)

Note: CI, credible interval; TH32, number of days with minimum temperature higher than 32°F; VMT, vehicle miles traveled; $\sigma^2{}_v$, $\sigma^2{}_\nu$, $\sigma^2{}_\varphi$, $\sigma^2{}_\theta$, and $\sigma^2{}_\eta$ are variances; $frac_v$ is the spatial fraction; $frac_\varphi$ is the temporal fraction; $frac_{\frac{S}{S+T}}$ is the relative importance of spatial effects; *covariates significant at the 95% credible interval.

3.4.1 Regression Coefficients Results

The intercept term was insignificant for fatal crashes but was significant for major injury and minor injury crashes. As expected, VMT showed significant positive effects for all three crash types. In addition, both intercept and VMT coefficients increased as crash injury severity decreased, which was consistent with the magnitude of crash counts.

Income was statistically insignificant for all three crash types, although income had generally increased for counties in Iowa from 2006 to 2015. Unemployment rate did not have significant effects on fatal crash counts but did have significantly negative effects on major injury and minor injury crash counts; that is, the number of major and minor injury crashes decreased as the unemployment rate increased. The unemployment rate has been thought to have

mixed effects on traffic crash frequencies (Leigh and Waldon, 1991; Wagenaar, 1983). On one hand, high unemployment is associated with more mental stress in the population, related to both job loss and fear of job loss, which could lead to more aggressive driving patterns and more traffic crashes (Wagenaar, 1983). On the other hand, high unemployment also brings with it less driving and thus fewer traffic crashes (Leigh and Waldon, 1991; Wagenaar, 1983). The latter seemed to predominate in Iowa, which was consistent with what was found in Michigan, where unemployment had negative effects on crash counts (Wagenaar, 1983).

Rainfall, snowfall, and TH32 did not show significant effects on any crash type. Although these weather indicators had great variability within the time span studied, they were not related to traffic safety problems in the long term. Adverse weather, such as snowstorms and flooding, may result in more crashes in the short term but may also reduce people's travel in the following time, leading to lower crash numbers. The two opposite effects seemed to offset each other.

It should be noted that all regression coefficients were assumed to be fixed for this study as shown in equation (3). That is, the effects of covariates on crash frequencies were thought to be homogeneous over space and time. However, these effects might be heterogeneous in practice in the presence of spatial instability and temporal instability, where fixed parameters models might produce biased coefficient estimates and incorrect inferences (Mannering, 2018; Mannering et al., 2016). For example, snowfall might affect crash frequencies differently in rural areas and urban areas due to different travel demands and travel modes. Thus, spatio-temporal-varying parameter models might be considered to get more accurate results in future studies.

Because most covariates are found to be insignificant, the $S_{MBYM}T_{MRW1}$ model is re-run with only significant variables, and the results were shown in Table 3-7. The posterior p-values

of the mean values of fatal, major injury, and minor injury crashes for the new model are 0.493, 0.495, and 0.500 respectively, which meant it fitted the data well. Mean and 95% CI values of estimated parameters were found to be generally consistent with those shown in Table 3-6.

Table 3-7 *Estimated parameters of the $S_{MBYM}T_{MRW1}$ model with significant covariates*

Variables	Fatal crash		Major injury crash		Minor injury crash	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Intercept	0.685*	(0.395, 0.983)	2.073*	(1.726, 2.430)	3.153*	(2.850, 3.468)
Unemployment rate	–	–	–0.051*	(–0.095, –0.006)	–0.060*	(–0.092, –0.030)
VMT	0.794*	(0.594, 0.978)	1.039*	(0.798, 1.251)	1.222*	(0.954, 1.452)
$\sigma^2{}_{\upsilon}$	0.281	(0.098, 0.591)	0.366	(0.118, 0.847)	0.431	(0.136, 1.039)
$\sigma^2{}_{v}$	0.127	(0.070, 0.203)	0.192	(0.105, 0.297)	0.234	(0.130, 0.358)
$frac_{\upsilon}$	0.662	(0.373, 0.878)	0.622	(0.322, 0.878)	0.614	(0.311, 0.878)
$\sigma^2{}_{\varphi}$	0.248	(0.079, 0.695)	0.246	(0.078, 0.682)	0.244	(0.076, 0.707)
$\sigma^2{}_{\theta}$	0.213	(0.072, 0.565)	0.209	(0.071, 0.554)	0.200	(0.070, 0.519)
$frac_{\varphi}$	0.530	(0.225, 0.817)	0.533	(0.232, 0.818)	0.538	(0.241, 0.818)
$frac_{\frac{S}{S+T}}$	0.487	(0.246, 0.715)	0.563	(0.311, 0.781)	0.609	(0.356, 0.809)
$\sigma^2{}_{\eta}$	0.047	(0.031, 0.067)	0.029	(0.022, 0.037)	0.018	(0.014, 0.022)

Note: CI, credible interval; VMT, vehicle miles traveled; $\sigma^2{}_{\upsilon}$, $\sigma^2{}_{v}$, $\sigma^2{}_{\varphi}$, $\sigma^2{}_{\theta}$, and $\sigma^2{}_{\eta}$ are variances; $frac_{\upsilon}$ is the spatial fraction; $frac_{\varphi}$ is the temporal fraction; $frac_{\frac{S}{S+T}}$ is the relative importance of spatial effects; *covariates significant at the 95% credible interval.

3.4.1 Random Effects Analysis

3.4.1.1 Spatial random effects analysis

For the $S_{MBYM}T_{MRW1}$ model, the spatial fraction values of fatal, major injury, and minor injury crashes were 0.662, 0.622, and 0.614, respectively. This means that, for all three crash types, unobserved heterogeneity in space existed both between counties and within individual counties and the structured spatial effects played slightly more important roles than did the unstructured spatial effects. Shown in Figure 3-3 are the exponential posterior means of the structured spatial effects ($\exp(\upsilon_{sk})$) of each county for all three crash types; counties with $\exp(\upsilon_{sk})$ lower than 1 tended to have fewer crashes and counties with $\exp(\upsilon_{sk})$ greater than 1

tended to have more crashes. It is found that the counties located in the north and southwest regions of Iowa tended to have fewer fatal, major injury, and minor injury crashes. For fatal crashes, this finding is generally consistent with the empirically observed fatal crash distribution shown in Figure 3-1 (a). However, for major injury and minor injury crashes, it is not obvious to see these trends in Figure 3-1 (a) and (b). This finding is a good example showing that one main benefit of spatial analysis is to the identification of the underlying spatial clustering of crashes.

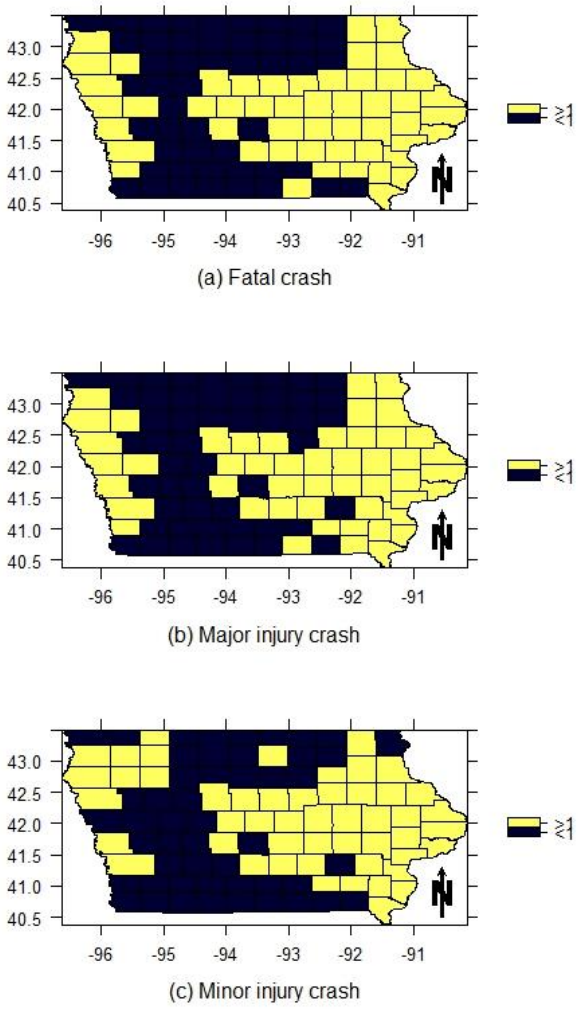

Figure 3-3 *Exponential posterior means of the structured spatial effect ($exp(\upsilon_{sk})$) of crashes in Iowa*

Moran's *I* statistics of residuals of the $S_{MBYM}T_{MRW1}$ model were calculated to see if they still had spatial correlations. As shown in Table 3-8, the residuals of fatal and major injury crashes did not show any significant spatial correlation at a 5% significance level for any year. In addition, the p-values were considerably larger than those shown in Table 3-3, which meant that unobserved heterogeneity in space was nearly completely covered by the spatial component. The *P*-values of Moran's *I* test for the residuals of minor injury crashes also increased considerably in most years, which meant that the weak spatial autocorrelations of minor injury crashes were also eliminated. However, there were some exceptions in 2006, 2007, and 2011 for minor injury crashes, when the raw crash data did not show significant spatial autocorrelations, whereas their residuals showed significant spatial autocorrelations. It is thought that minor injury crashes might have trivial spatial autocorrelation in these three years but did have non-trivial spatial correlations in other years. However, because the $S_{MBYM}T_{MRW1}$ model assigned fixed spatial random effects to the data for each year, the residuals would also have spatial effects as the complement in these three years. This needs further investigation to determine the true reason. This finding implies the importance of checking the necessity of adopting spatial models is crash analysis. We suggest making spatial tests before and after spatial analysis to justify the utilization of spatial models. In general, the spatial component covered nearly all unobserved heterogeneity of crashes in space. The results also generally verified the effectiveness of the spatial model.

Table 3-8 *Moran's I test results for the residuals of the $S_{MBYM}T_{MRW1}$ model*

Year	Fatal crash		Major injury crash		Minor injury crash	
	Moran's *I*	*P*-value	Moran's *I*	*P*-value	Moran's *I*	*P*-value
2006	-1.285	0.901	0.003	0.499	1.835	0.033*
2007	-0.195	0.577	-0.944	0.828	3.434	0.000*
2008	-0.912	0.819	-0.130	0.552	1.174	0.120
2009	-0.647	0.741	0.865	0.194	0.096	0.462
2010	0.660	0.255	1.323	0.093	-0.407	0.658
2011	0.099	0.461	0.489	0.313	3.586	0.000*
2012	0.232	0.408	-0.171	0.568	-0.789	0.785
2013	0.495	0.310	-0.021	0.508	-1.019	0.846
2014	-0.430	0.666	0.355	0.361	-0.669	0.748
2015	-1.409	0.921	-0.285	0.612	-1.579	0.943

Note: * significant at P = 0.05.

3.4.1.2 Temporal random effects analysis

For the $S_{MBYM}T_{MRW1}$ model, the temporal fractional values of fatal, major injury, and minor injury crashes were 0.530, 0.533, and 0.538, respectively. The structured temporal effects and the unstructured temporal effects played nearly the same roles for all three crashes. Thus, unobserved heterogeneity in time existed both between years and in individual years. Shown in Figure 3-4 are the exponential posterior means of the structured temporal effects ($\exp(\varphi_{tk})$) in each year for all three crash types. All three crash types generally showed descending trends from 2006 to 2015, whereas major injury and minor injury crashes had some fluctuations.

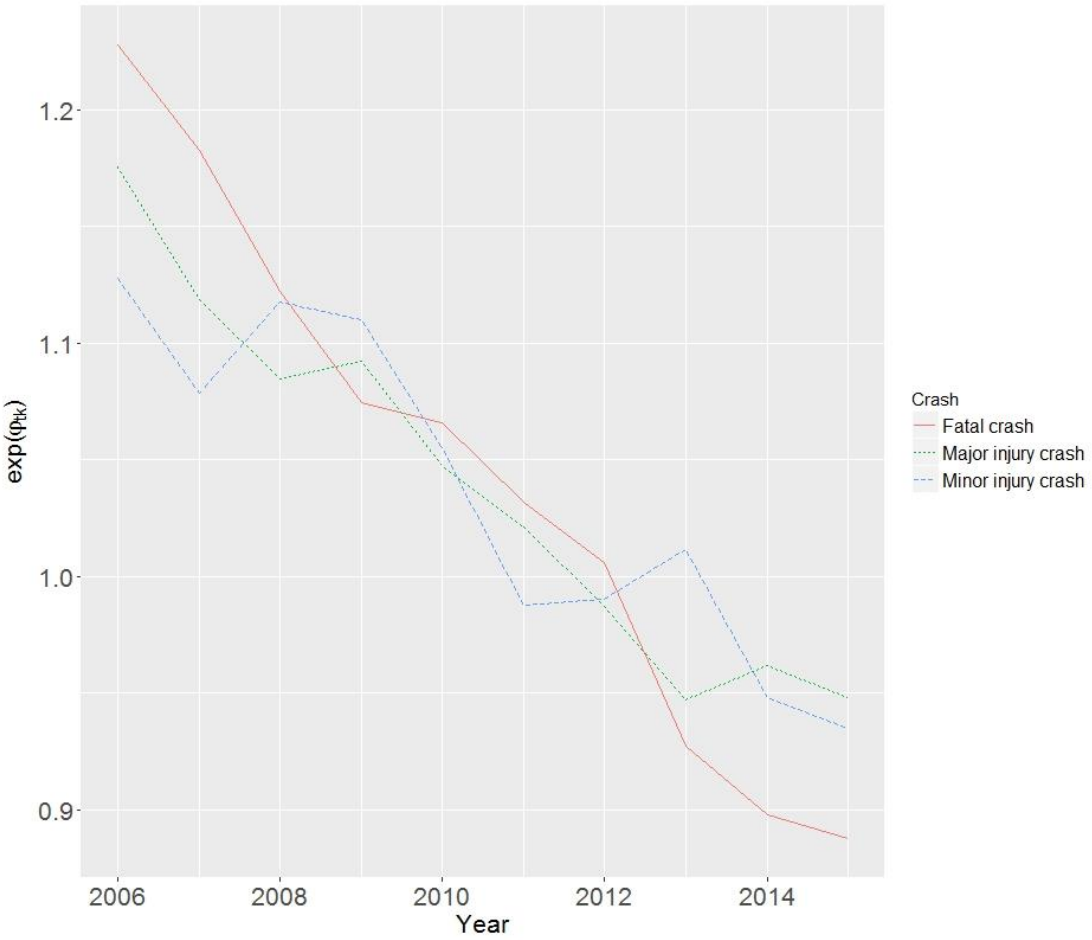

Figure 3-4 *Exponential posterior means of the structured temporal effects ($exp(\varphi_{tk})$) of the $S_{MBYM}T_{MRW1}$ model*

3.4.1.3 Spatial and temporal random effects comparison

The $frac_{\frac{S}{S+T}}$ values of fatal, major injury, and minor injury crashes were 0.487, 0.563, and 0.609, respectively. This means that the temporal effects played slightly more important roles for fatal crashes, whereas spatial effects played slightly more important roles for major injury and minor injury crashes. That is, the relative importance of spatial effects and temporal effects varied slightly by crash injury severity.

3.4.1.4 Unobserved heterogeneity across crash injury severities

The estimated variance-covariance matrices for all the random effects and the corresponding 95% credible intervals of the $S_{MBYM}T_{MRW1}$ model are shown in Table 3-9.

Table 3-9 *Estimated covariance matrices of the $S_{MBYM}T_{MRW1}$ model*

Structured spatial effects ($\upsilon_{s.}$)			
Σ_{υ}	Fatal crash	Major injury crash	Minor injury crash
Fatal crash	0.281 (0.098, 0.591)		
Major injury crash	0.235 (0.035, 0.594)	0.366 (0.118, 0.847)	
Minor injury crash	0.253 (0.035, 0.661)	0.322 (0.063, 0.850)	0.431 (0.136, 1.039)
Unstructured spatial effects ($\nu_{s.}$)			
Σ_{ν}	Fatal crash	Major injury crash	Minor injury crash
Fatal crash	0.127 (0.070, 0.203)		
Major injury crash	0.110 (0.046, 0.190)	0.192 (0.105, 0.297)	
Minor injury crash	0.121 (0.051, 0.207)	0.173 (0.082, 0.279)	0.234 (0.130, 0.358)
Structured temporal effects ($\varphi_{t.}$)			
Σ_{φ}	Fatal crash	Major injury crash	Minor injury crash
Fatal crash	0.248 (0.079, 0.695)		
Major injury crash	0.001 (-0.223, 0.231)	0.246 (0.078, 0.682)	
Minor injury crash	-0.003 (-0.235, 0.227)	-0.009 (-0.257, 0.208)	0.244 (0.076, 0.707)
Unstructured temporal effects ($\theta_{t.}$)			
Σ_{θ}	Fatal crash	Major injury crash	Minor injury crash
Fatal crash	0.213 (0.072, 0.565)		
Major injury crash	-0.002 (-0.193, 0.191)	0.209 (0.071, 0.554)	
Minor injury crash	-0.003 (-0.182, 0.173)	-0.007 (-0.186, 0.159)	0.200 (0.070, 0.519)
Spatio-temporal interaction effects ($\eta_{(st).}$)			
Σ_{η}	Fatal crash	Major injury crash	Minor injury crash
Fatal crash	0.047 (0.031, 0.067)		
Major injury crash	0.002 (-0.007, 0.010)	0.029 (0.022, 0.037)	
Minor injury crash	0.000 (-0.005, 0.006)	0.006 (0.002, 0.010)	0.018 (0.014, 0.022)

Note: values shown are the posterior mean with the 95% credible interval in parentheses; Σ_{υ}, Σ_{ν}, Σ_{φ}, Σ_{θ}, Σ_{η} are variance–covariance matrices of structured spatial effects, unstructured spatial effects, structured temporal effects, unstructured temporal effects, and spatio-temporal interaction effects, respectively.

For the $S_{MBYM}T_{MRW1}$ model, unobserved heterogeneity across crash injury severities had three sources: space, time, and spatio-temporal interaction. All the off-diagonal elements of Σ_{υ} and Σ_{ν} were significantly positive, which meant there were strong positive correlations across crash injury severities for both structured and unstructured spatial effects. That is, with the increase of fatal crash counts in one county, the major injury and minor injury crash counts of this county, and the fatal, major injury, and minor injury crash counts of its neighboring counties were also expected to increase. This proves the necessity of using multivariate spatial models

from another viewpoint. However, none of the off-diagonal elements of Σ_φ and Σ_θ were significantly different from zero, which implied that there were no strong correlations across crash injury severities for either structured or unstructured temporal effects. However, the DIC value of the $S_{MBYM}T_{MRW1}$ model was still much smaller than that for the $S_{MBYM}T_{RW1}$ model (shown in Table 3-5). This implies that, although crashes may not show strong correlations in time, their correlations may still not be ignored, as weak correlations may still explain some variability in the data. For the spatio-temporal interaction effects, major injury crashes showed significantly positive correlations with minor injury crashes, but fatal crashes did not show significant correlations with the other two crash types.

For each observation, because $\Sigma_\upsilon,\ \Sigma_\nu,\ \Sigma_\varphi,\ \Sigma_\theta,\ and\ \Sigma_\eta$ are independent, the Pearson's correlation coefficients of random effects across crash injury severities can be calculated as follows:

$$\rho_{12} = \frac{\Sigma_{\upsilon_{12}}+\Sigma_{\nu_{12}}+\Sigma_{\varphi_{12}}+\Sigma_{\theta_{12}}+\Sigma_{\eta_{12}}}{\sqrt{\sigma^2{\upsilon}^1+\sigma^2{\nu}^1+\sigma^2{\varphi}^1+\sigma^2{\theta}^1+\sigma^2{\eta}^1}\sqrt{\sigma^2{\upsilon}^2+\sigma^2{\nu}^2+\sigma^2{\varphi}^2+\sigma^2{\theta}^2+\sigma^2{\eta}^2}} \tag{3.22}$$

$$\rho_{13} = \frac{\Sigma_{\upsilon_{13}}+\Sigma_{\nu_{13}}+\Sigma_{\varphi_{13}}+\Sigma_{\theta_{13}}+\Sigma_{\eta_{13}}}{\sqrt{\sigma^2{\upsilon}^1+\sigma^2{\nu}^1+\sigma^2{\varphi}^1+\sigma^2{\theta}^1+\sigma^2{\eta}^1}\sqrt{\sigma^2{\upsilon}^3+\sigma^2{\nu}^3+\sigma^2{\varphi}^3+\sigma^2{\theta}^3+\sigma^2{\eta}^3}} \tag{3.23}$$

$$\rho_{23} = \frac{\Sigma_{\upsilon_{23}}+\Sigma_{\nu_{23}}+\Sigma_{\varphi_{23}}+\Sigma_{\theta_{23}}+\Sigma_{\eta_{23}}}{\sqrt{\sigma^2{\upsilon}^2+\sigma^2{\nu}^2+\sigma^2{\varphi}^2+\sigma^2{\theta}^2+\sigma^2{\eta}^2}\sqrt{\sigma^2{\upsilon}^3+\sigma^2{\nu}^3+\sigma^2{\varphi}^3+\sigma^2{\theta}^3+\sigma^2{\eta}^3}} \tag{3.24}$$

where ρ_{12} is the Pearson correlation coefficient of random effects between fatal and major injury crashes, ρ_{13} is the Pearson correlation coefficient of random effects between fatal and minor injury crashes, and ρ_{23} is the Pearson correlation coefficient of random effects between major injury and minor injury crashes.

The posterior means and 90% credible intervals of Pearson correlation coefficients of random effects are shown in Table 3-10. The Pearson correlation coefficient between any two

crash types was significantly positive at a 90% credible interval, but the Pearson correlation coefficient between major injury and minor injury crashes was generally larger than the other two values. That is, major injury and minor injury crashes had a stronger correlation compared to fatal crashes, which was consistent with the Pearson correlation coefficients of crash counts shown in Table 3-2.

Table 3-10 *Pearson's correlation coefficients of random effects across crash injury severities*

Pearson correlation	Mean	90% CI
ρ_{12}	0.357	(0.047, 0.605)
ρ_{13}	0.366	(0.066, 0.605)
ρ_{23}	0.453	(0.145, 0.689)

Note: CI, credible interval; ρ_{12}, ρ_{13}, ρ_{23}, Pearson correlation coefficients between fatal and major injury crashes, between fatal and minor injury crashes, and between major injury and minor injury crashes, respectively.

3.4.2 Site Ranking Results Analysis

The crude crash rates and the predicted crash rates for all three crash types, which were calculated by dividing the crash counts by VMT, are shown in Figure 3-5. A linear regression model was built to check their correlation.

The R^2 value was 0.929, which means that the crude crash rates were generally consistent with the predicted crash rates. Specifically, for major injury and minor injury crashes, these two rates were very consistent, but for fatal crashes, they were inconsistent. Major injury and minor injury crash counts were very large, but fatal crash counts were very small, as shown in Table 3-1. Thus, occurrences of fatal crashes were more stochastic than major injury and minor injury crashes. It is thought that the multivariate structure could borrow information from major injury and minor injury crashes to estimate fatal crashes stably (Boulieri et al., 2017). Thus, the predicted data from the SMBYMTMRW1 model are expected to be smoother for unstable low-

frequency fatal crashes, and could represent the underlying true distribution of fatal crashes better than the crude data could.

The crash cost rates directly influenced the ranking results shown in Figure 3-6, where x-axis showed the crude rank by the crude crash cost rate and y-axis showed the PER by the predicted crash cost rate. The two ranking methods produced consistent results for major injury and minor injury crashes but had large differences for fatal crashes, which led to different ranking results for total crashes.

The top 10 risky counties using the two ranking methods are shown in Figure 3-7. Of the counties ranked by these two methods, seven appeared in the top 10 for both methods, whereas three counties appeared only in the top 10 of one or the other method; Lyon, Hamilton, and Mahaska Counties were in the top 10 list using the predicted crash cost rate PER but not in the crude crash cost rate ranking. Moreover, the rank orders of the seven counties appearing in both top 10 lists were also very different. For example, the highest ranked county by the crude crash cost rate, Marion County, was ranked only eighth by the PER of the predicted crash cost rate. The big differences between the two ranking methods show the importance of the multivariate spatio-temporal Bayesian model, which is expected to better identify the underlying true status quo of traffic safety. The top 10 risky counties shown in Figure 3-7 (b) should be the focus of future safety improvement programs.

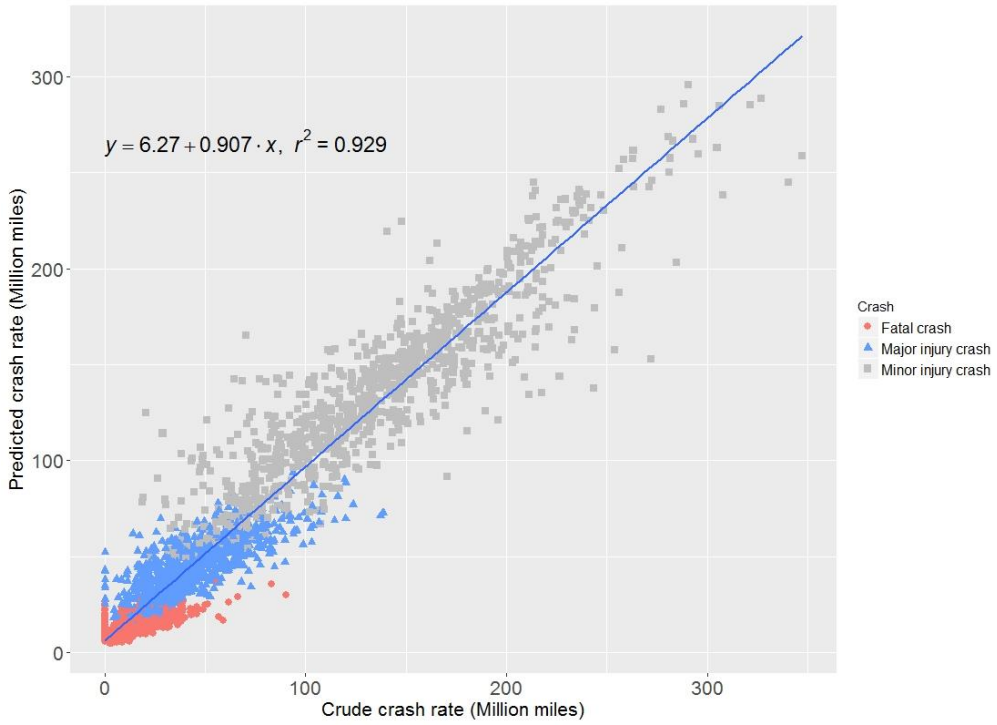

Figure 3-5 *Crude crash rate versus predicted crash rate of fatal, major injury, and minor injury crashes*

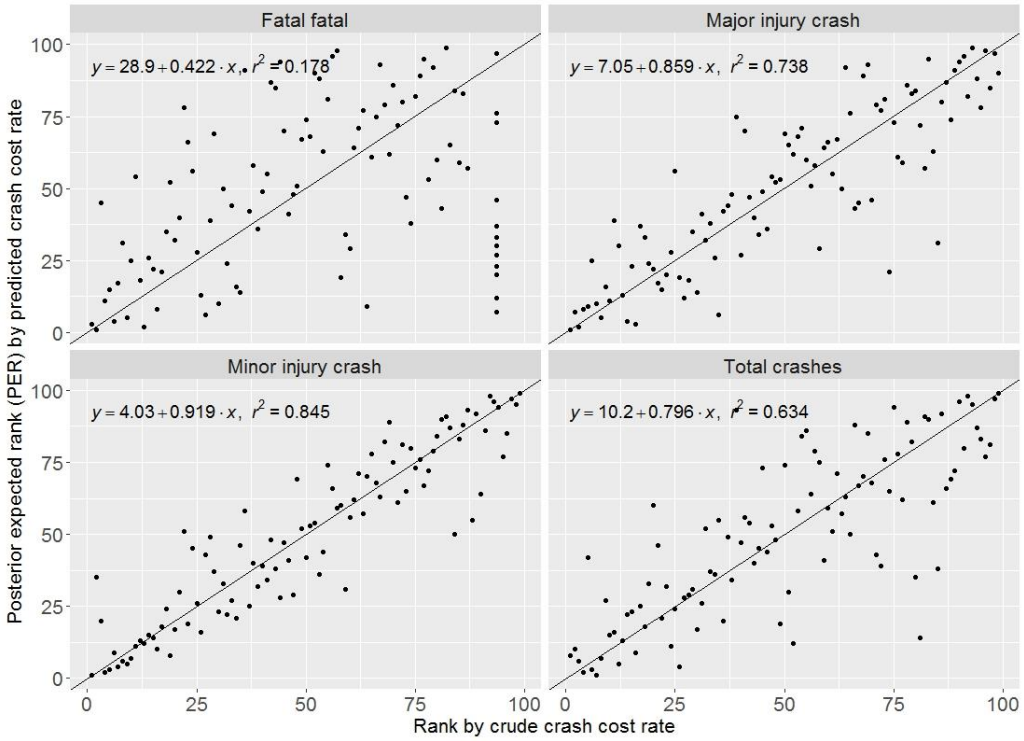

Figure 3-6 *County rank by the crude crash cost rate versus county posterior expected rank by the predicted crash cost rate in 2015*

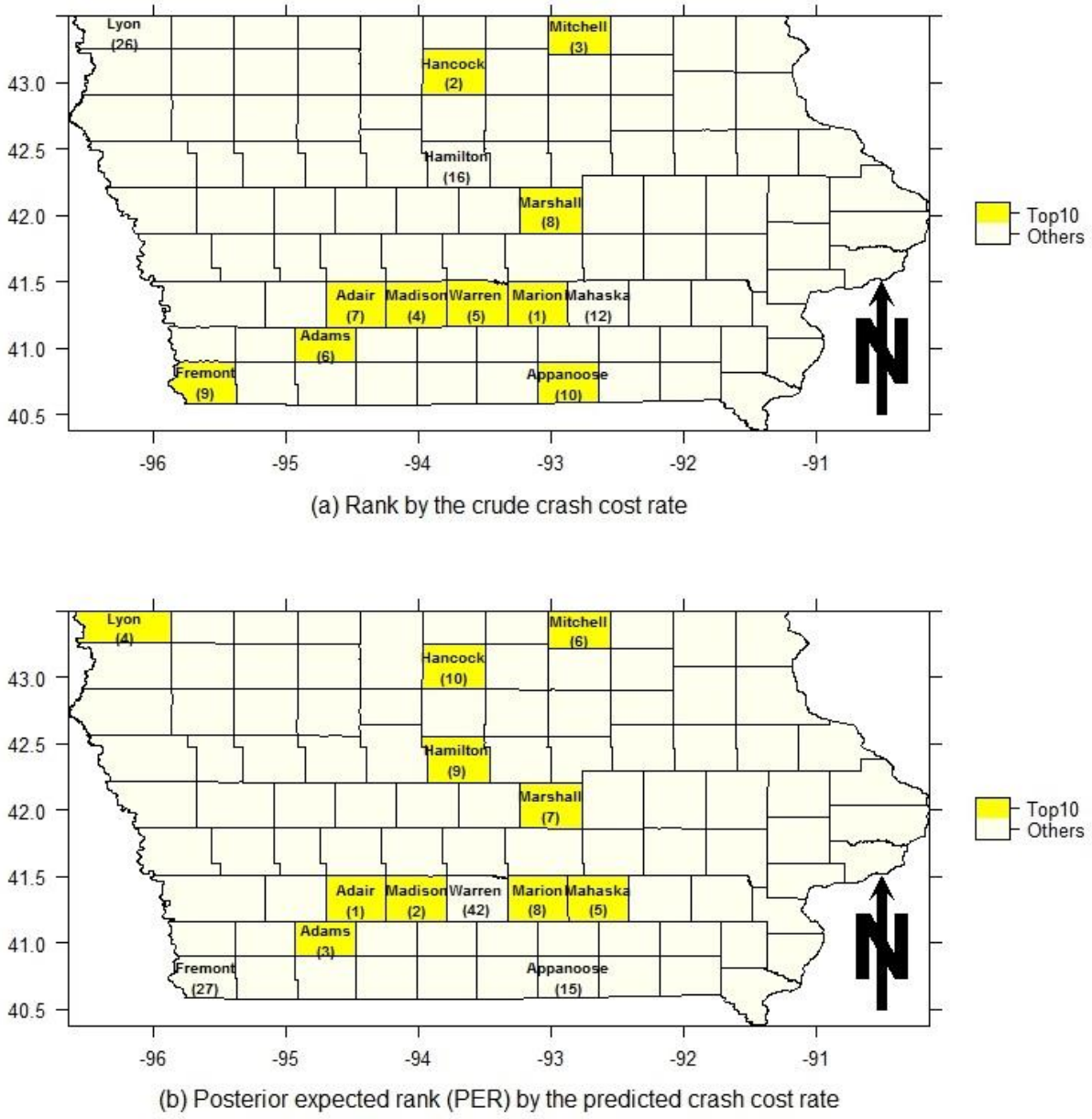

Figure 3-7 *Counties with the 10 highest crash cost rates using the two ranking methods*

3.5 Conclusions and Discussions

Unobserved heterogeneity of crashes over space and time is often a big issue in crash frequency analysis. When multiple crashes are analyzed, correlations across crash types may also produce unobserved heterogeneity, which may exist in space, time, and space–time interactions. In this study, we used the multivariate spatio-temporal Bayesian model to analyze the yearly

county-level fatal, major injury, and minor injury crash counts in Iowa from 2006 to 2015. Income, rainfall, snowfall, and temperature did not have significant influences on the frequencies of any of the three crash types, whereas unemployment rate showed significantly negative influences on major injury and minor injury crash counts, and VMT showed significantly positive influences on all three crash types.

All three crash types showed very strong spatial correlations. The counties located in northern and southwestern Iowa tended to have fewer crashes, whereas the remaining counties tended to have more crashes. All three crash types generally showed descending trends from 2006 to 2015. Both spatial and temporal effects were non-negligible, and they played nearly the same roles for all three crash types with slight differences. In addition, all three crash types showed significantly positive correlations between each other across space but not across time. The crude data and the predicted data were generally consistent for major injury and minor injury crashes but were very different for fatal crashes, the crude data of which were more stochastic due to the low counts. The predicted data from the multivariate spatio-temporal model were smoother than were the crude data. The crash cost rates were calculated based on crash rates and crash costs by injury severity and were used as ranking indicators. Two ranking methods, crude rank by the crude crash cost rate and PER by the predicted crash cost rate, were presented to identify the counties with higher risks for traffic safety. The two methods produced very different ranking results, and the latter method was thought to be able to better represent the true status quo of traffic safety. The ranking results would be helpful for transportation agencies drawing up traffic safety improvement programs in the future.

In future research, the data may be analyzed using smaller space and time scales, which would produce more targeted and practical findings. In addition, as shown in Table 3-3, the

spatial correlations of all three crashes were different in different years. That is, the spatial correlations may evolve dynamically over time. Similar situations may also appear in temporal correlations, whereby the descending rates of crashes in different counties may be different. Thus, dynamic spatio-temporal models should be considered in future studies. Meanwhile, in this study, only random effects were thought to be correlated in space and time, but regression coefficients might also be correlated in space and time. Thus, future researchers may want to consider spatio-temporal-varying coefficient models. It is suggested that the review by Mannering (2018) about temporal instability in accident analysis be consulted for more ideas. All the above-mentioned directions would need more data or more complex statistical models, so computation may be a big concern, especially when using MCMC simulation to estimate Bayesian models. Some emerging fast Bayesian estimation tools, such as integrated nested Laplace approximation (Rue et al., 2009), should be considered. As was shown in this study, care should also be taken in the selection of appropriate priors and initial values for MCMC simulations. Finally, for this study we adopted two common spatial and temporal models; however, there are many other spatial and temporal models available. Future researchers may also explore the effectiveness of other models in crash frequency analysis.

3.4 References

Aguero-Valverde, J., 2011. Direct spatial correlation in crash frequency models. 3rd International Conference on Road Safety and Simulation, Indianapolis, IN, USA.

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

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention 38 (3), 618–625.

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record 2061, 55–63.

Aguero-Valverde, J., Jovanis, P.P., 2010. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transportation Research Record 2136, 82–91.

Aguero-Valverde, J., Wu, K.F., Donnell, E.T., 2016. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accident Analysis and Prevention 87, 8–16.

Ahmed, M.M., Abdel-Aty, M., 2015. Evaluation and spatial analysis of automated red-light running enforcement cameras. Transportation Research Part C 50, 130–140.

Andrey, J., 2010. Long-term trends in weather-related crash risks. Journal of Transport Geography 18 (2), 247–258.

Anselin, L., 1988. Spatial Econometrics: Methods And Models. Springer, Netherlands.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1–15.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43 (1), 1–20.

Bivand, R., Piras, G., 2015. Comparing implementations of estimation methods for spatial econometrics. Journal of Statistical Software 63 (18), 1–36.

Blangiardo, M., Cameletti, M., Baio, G., Rue, H., 2013. Spatial and spatio-temporal models with R-INLA. Spatial and Spatio-temporal Epidemiology 4, 33–49.

Boulieri, A., Liverani, S., Hoogh, K. de, Blangiardo, M., 2017. A space–time multivariate Bayesian model to analyse road traffic accidents by severity. Journal of the Royal Statistical Society Series A 180 (1), 119–139.

Brijs, T., Karlis, D., Wets, G., 2008. Studying the effect of weather conditions on daily crash counts using a discrete time-series model. Accident Analysis and Prevention 40 (3), 1180–1190.

Eckley, D.C., Curtin, K.M., 2013. Evaluating the spatiotemporal clustering of traffic incidents. Computers, Environment and Urban Systems 37, 70–81.

El-Basyouny, K., Barua, S., Islam, M.T., 2014. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis and Prevention 73, 91–99.

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41 (4), 820–828.

Erdogan, S., 2009. Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. Journal of Safety Research 40 (5), 341–351.

Gelman, A., Meng, X.-L., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6 (4), 733–807.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7 (4), 457–511.

Guo, F., Wang, X., Abdel-Aty, M.A., 2010. Modeling signalized intersection safety with corridor-level spatial correlations. Accident Analysis and Prevention 42 (1), 84–92.

Hu, S., Ivan, J.N., Ravishanker, N., Mooradian, J., 2013. Temporal modeling of highway crash counts for senior and non-senior drivers. Accident Analysis and Prevention 50, 1003–1013.

Iowa Community Indicators Program, 2016. Iowa Community Indicators Program (ICIP). http://www.icip.iastate.edu/ (accessed 2.20.17).

Iowa Department of Transportation, 2016. Vehicle-miles traveled (VMT). http://www.iowadot.gov/maps/msp/vmt/vmt.html (accessed 10.20.16).

Iowa Environmental Mesonet, 2017. Iowa Environmental Mesonet (IEM). https://mesonet.agron.iastate.edu/ (accessed 2.20.17).

Jiang, X., Abdel-Aty, M., Alamili, S., 2014. Application of Poisson random effect models for highway network screening. Accident Analysis and Prevention 63, 74–82.

Kilamanua, W., Xia, J., Caulfieldb, C., 2011. Analysis of spatial and temporal distribution of single and multiple vehicle crash in Western Australia: a comparison study. 19th International Congress on Modelling and Simulation, Perth, Australia.

Leigh, J.P., Waldon, H.M., 1991. Unemployment and highway fatalities. Journal of Health Politics, Policy and Law 16 (1), 135–156.

Liu, C., Gyawali, S., Sharma, A., Smaglik, E., 2015. A methodological approach for spatial and temporal analysis of red light running citations and crashes: a case-study in Lincoln, Nebraska. Transportation Research Board 94th Annual Meeting, Washington, D.C., USA.

Liu, C., Sharma, A., 2017. Exploring spatio-temporal effects in traffic crash trend analysis. Analytic Methods in Accident Research 16, 104–116.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-Gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37 (1), 35–46.

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N., 2009. The BUGS project: Evolution, critique, and future directions. Statistics in medicine 28 (25), 3049–3067.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964–975.

Ma, X., Chen, S., Chen, F., 2017. Multivariate space-time modeling of crash frequencies by injury severity levels. Analytic Methods in Accident Research 15, 29–40.

Mannering, F.L., 2018. Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17, 1–13.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1–16.

Matkan, A., Mohaymany, A., 2013. Detecting the spatial–temporal autocorrelation among crash frequencies in urban areas. Canadian Journal of Civil Engineering 40 (3), 195–203.

Meng, X.-L., 1994. Posterior Predictive p-Values. The Annals of Statistics 22 (3), 1142–1160.

Miaou, S.-P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping a space-time modeling approach. Journal of Transportation and Statistics 6 (1), 33–57.

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

Michalaki, P., Quddus, M., Pitfield, D., Huetson, A., 2016. A time-series analysis of motorway collisions in England considering road infrastructure, socio-demographics, traffic and weather characteristics. Journal of Transport and Health 3 (1), 9–20.

MRC Biostatistics Unit, 2004. DIC: Deviance Information Criteria. https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-dic/ (accessed 8.23.17).

Pawlovich, M.D., 2007. Safety improvement candidate location (SICL) methods, Iowa Department Of Transportation, Highway Division, Engineering Bureau, Office Of Traffic And Safety. Ames, IA.

Quddus, M.A., 2008a. Time series count data models: an empirical application to traffic accidents. Accident Analysis and Prevention 40 (5), 1732–1741.

Quddus, M.A., 2008b. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis and Prevention 40 (4), 1486–1497.

R Core Team, 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rue, H., Martino, S., Chopin, N., 2009. Appoximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B 71 (2), 319–392.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43 (5), 1666–1676.

Serhiyenko, V., Ivan, J.N., Ravishanker, N., Islam, M.S., 2014. Dynamic compositional modeling of pedestrian crash counts on urban roads in Connecticut. Accident Analysis and Prevention 64, 78–85.

Shen, W., Louis, T.A., 1998. Triple-goal estimates in two-stage hierarchical models. Journal of the Royal Statistical Society Series B 60 (2), 455–471.

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97 (1), 246–273.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B 64 (4), 583–639.

Sturtz, S., Ligges, U., Gelman, A., 2005. R2OpenBUGS: a package for running OpenBUGS from R. Journal of Statistical Software 12 (3), 1–16.

Sukhai, A., Jones, A.P., Love, B.S., Haynes, R., 2011. Temporal variations in road traffic fatalities in South Africa. Accident Analysis and Prevention 43 (1), 421–428.

Truong, L.T., Kieu, L.-M., Vu, T.A., 2016. Spatiotemporal and random parameter panel data models of traffic crash fatalities in Vietnam. Accident Analysis and Prevention 94, 153–161.

U.S. Bureau of Economic Analysis, 2016. Personal income sumamry: personal income, population, per capita personal income. https://www.bea.gov/iTable/iTable.cfm?reqid=70&step=30&isuri=1&7022=20&7023=7&7024=non-industry&7033=-1&7025=4&7026=xx,19000&7027=2015,2014,2013,2012,2011,2010&7001=720&7028=3&7030=0&7031=19000&7040=-1&7083=levels&7029=20&7090=70#reqid=70&step=30&isuri=1& (accessed 2.20.17).

Wagenaar, A.C., 1983. Unemployment and motor vehicle accidents in Michigan, UMTRI-83-45. The University of Michigan, Transportation Research Institute. Ann Arbor, Michigan.

Wang, C., Quddus, M.A., Ison, S.G., 2009. Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. Accident Analysis and Prevention 41 (4), 798–808.

Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident Analysis and Prevention 43 (6), 1979–1990.

Wang, C., Quddus, M., Ison, S., 2013. A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. Transportmetrica A 9 (2), 124–148.

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis and Prevention 38 (6), 1137–1150.

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

Xie, K., Wang, X., Ozbay, K., Yang, H., 2014. Crash frequency modeling for signalized intersections in a high-density urban road network. Analytic Methods in Accident Research 2, 39–51.

Yannis, G., Antoniou, C., Papadimitriou, E., 2011. Autoregressive nonlinear time-series modeling of traffic fatalities in Europe. European Transport Research Review 3 (3), 113–127.

Zeng, Q., Huang, H., 2014. Bayesian spatial joint modeling of traffic crashes on an urban road network. Accident Analysis and Prevention 67, 105–112.

Zhao, M., Liu, C., Li, W., Sharma, A., 2017. Multivariate Poisson-lognormal model for analysis of crashes on urban signalized intersections approach. Journal of Transportation Safety & Security (In press).

CHAPTER 4. MULTIVARIATE RANDOM PARAMETERS ZERO-INFLATED NEGATIVE BINOMIAL REGRESSION FOR ANALYZING URBAN MIDBLOCK CRASHES

A paper published on the Analytic Methods in Accident Research

Abstract

Urban midblock crashes are influenced mainly by traffic operation and roadway geometric features. In this paper, 10-year crash data from 1,506 urban midblock segments in Nebraska were analyzed using the multivariate random parameters zero-inflated negative binomial model to account for unobserved heterogeneity produced by correlations across segments, correlations across crash collision types, excessive zero crashes, and over dispersion. The multivariate random parameters zero-inflated negative binomial model was superior to many common crash frequency models in terms of both goodness of fit and prediction accuracy. Compared with the multivariate fixed parameters zero-inflated negative binomial model, the multivariate random parameters zero-inflated negative binomial model identified fewer key influencing factors and revealed segment-specific effects of these factors on different crash types. It also showed that the number of lanes, annual average daily traffic per lane, and segment length might have negative effects on crash frequencies. Segments with a speed limit of 45 mph had fewer crashes than did those with lower speed limits, and there were fewer crashes on the segments in Omaha than on those in Lincoln. It was also found that neither the presence of a shoulder, on-street parking, or one-way traffic, nor lane width had significant influences on crash frequencies. These findings are informative for transportation agencies to take correct and efficient measures to accommodate diverse transportation demands without reducing traffic safety.

Keywords: unobserved heterogeneity, multivariate random parameters zero-inflated negative binomial model, crash frequency, urban midblock segments, Bayesian

4.1 Introduction

Traffic crashes can be divided into junction crashes and non-junction crashes based on where they occur (National Center for Statistics and Analysis, 2017). Non-junction crashes, also referred to as midblock crashes, are crashes that occur on roadway segments. In 2015, they accounted for 41.7% of the total number of crashes and 63.3% of fatal crashes in the United States (National Center for Statistics and Analysis, 2017). Thus, reducing midblock crashes is critical for improving traffic safety. Although midblock crashes are usually not directly influenced by junctions, they are greatly influenced by traffic operation and roadway geometric factors, which are much more complex on urban roadways than on rural roadways. On one hand, urban roadway segments usually have large traffic volumes and face diverse traffic demands, which might increase crash opportunities; for example, an increase in the number of crosswalks might increase the frequency of pedestrian crashes. On the other hand, urban development might limit or even reduce available roadway space, which might also increase crash risk; for example, vehicle lanes may be narrowed to make room for biking lanes and on-street parking. This predicament requires transportation agencies to determine what traffic operation and roadway geometric factors really influence the frequency of urban midblock crashes so that they can take effective measures to accommodate traffic demands without reducing traffic safety.

Previous studies have shown that important traffic operation and roadway geometric factors influencing midblock crashes include traffic volume (Bonneson and Mccoy, 1997; Dumbaugh, 2006; Ferreira and Couto, 2015; Greibe, 2003; Manuel et al., 2014; Zhang et al., 2012), speed limit (Dumbaugh, 2006; Greibe, 2003; Pande et al., 2010), on-street parking (Bonneson and Mccoy, 1997; Greibe, 2003), lane width (Greibe, 2003; Manuel et al., 2014), median type (Bonneson and Mccoy, 1997; Sawalha and Sayed, 2001), median width (Dumbaugh, 2006), number of lanes (Dumbaugh, 2006; Greibe, 2003; Sawalha and Sayed,

2001), land use (Bonneson and Mccoy, 1997; Greibe, 2003; Sawalha and Sayed, 2001), pavement condition (Usman et al., 2010; Xiong et al., 2014; Zeng and Huang, 2014), access points (Lee et al., 2011; Zeng and Huang, 2014), and so on. However, studies' findings have often been inconsistent, that is, some factors might have had different effects in different studies. For example, speed limit was found to be not significant for midblock crash frequencies on a 27-mile urban arterial in Florida Department of Transportation District 5 (Dumbaugh, 2006), whereas it was the most important variable for midblock crash frequencies on a 19.659-mile corridor of U.S. Route 19 in Pasco County, Florida (Pande et al., 2010). This inconsistency implies that, in practice, the effects of some factors on crashes might be location specific. Ignoring this unobserved heterogeneity might produce biased and inefficient estimated parameters, leading to erroneous inferences and predictions (Mannering et al., 2016).

One solution is to adopt random parameters count data models (Alarifi et al., 2017; Barua et al., 2016, 2015; Bhat et al., 2017; Chen and Tarko, 2014; Chen et al., 2017; Coruh et al., 2015; Lord and Mannering, 2010; Rista et al., 2017; Venkataraman et al., 2014). Compared to fixed parameters models assuming the same effects of factors on all observations, random parameters models can capture the observation-specific effects of factors on crash frequency and have also been widely applied in crash injury severity analyses (Anderson and Hernandez, 2017; Behnood and Mannering, 2017a, 2017b, 2016; Fountas and Anastasopoulos, 2017; Naik et al., 2016; Russo et al., 2014; Seraneeprakarn et al., 2017; Zhao and Khattak, 2017, 2015) and crash rate analyses (Anastasopoulos, 2016). Especially, for the data where one entity has multiple observations, such as panel data, group-specific random parameters models may be adopted to account for heterogeneity among groups (Sarwar et al., 2017; Wu et al., 2013). More details about random parameters formulations can be seen in the study by Mannering et al. (2016).

Crash data usually can be divided into multiple types based on different criteria. For example, midblock crashes can be divided based on the type of collision: rear-end crashes, right-angle crashes, side-swipe (same direction) crashes, single-vehicle crashes, overturn crashes, and so on. A single factor might be expected to have different effects on different collision types, causing different outcomes. Thus, identifying the specific significant factor for each collision type is important for transportation agencies so they can take accurate countermeasures to reduce specific types of collision. When these crashes are jointly analyzed, multivariate count data models are necessary, as univariate models may produce biased and inefficient results because the unobserved heterogeneity often present across crash types is ignored (Dong et al., 2014a; Huang et al., 2008; Mannering et al., 2016). Most multivariate count data models in literature were derived from the multivariate Poisson log-normal (MVPLN) model (Aguero-Valverde and Jovanis, 2010; Barua et al., 2014; El-Basyouny and Sayed, 2009; Huang et al., 2017; Ma et al., 2008; Osama and Sayed, 2017; Serhiyenko et al., 2016; Wang et al., 2018; Zhan et al., 2015; Zhao et al., 2017), which is flexible enough to accommodate various correlations among crash types, but it does not work well for crash data with excess zeros (Dong et al., 2014a). In addition to the multivariate Poisson log-normal model, the natural extensions of the Poisson and negative binomial (NB) models to multivariate data, i.e., the multivariate Poisson (MVP) model (Johnson et al., 1997; Ma and Kockelman, 2006) and the multivariate negative binomial (MVNB) model (Anastasopoulos et al., 2012; Chen et al., 2017), also have been used in some studies. The multivariate Poisson/negative binomial models assume positive correlations across crash types, but they cannot deal with crash data with excess zeros either, as the marginal distribution per crash type is still a Poisson/negative binomial model.

The zero-inflated models are often adopted for univariate count data with excess zeros (Lambert, 1992; Lord et al., 2005). The excess zeros in crash frequency data can be explained in two ways for zero-inflated models. One explanation is that there is a two-state crash-generating process: (i) a normal count state and (ii) an accident-free state, which can be thought of as a nearly safe state, with accidents occurring extremely rarely (Malyshkina and Mannering, 2010). The other explanation is that there is a two-state crash-reporting process: (i) one in which accidents did occur, but they were not reported for some reason, such as for minor crashes, which were not necessary to report, or hit-and-run crashes, i.e., a crash-underreporting state, and (ii) one in which all accidents that occurred were reported, i.e., a normal crash reporting state. This explanation applies to many scenarios, as crash underreporting has been found to be common in practice (Elvik and Mysen, 1999; Hauer and Hakkert, 1988; Lord and Mannering, 2010; Yamamoto et al., 2008; Yannis et al., 2014). Both explanations may justify the application of zero-inflated models in our case, although it is difficult to determine what the truth is by observing the data. In cases for which crash observations at each level of classification are characterized with a significant number of zero occurrences, the zero-inflated versions of the multivariate Poisson and negative binomial models, i.e., the multivariate zero-inflated Poisson (MVZIP) model (Li et al., 1999) and the multivariate zero-inflated negative binomial (MVZINB) model, are recommended. In traffic safety studies, the multivariate zero-inflated Poisson model was first used to examine the crash frequency at signalized intersections in Tennessee, and it was found to perform better than the univariate zero-inflated Poisson (UZIP) and multivariate Poisson log-normal models in terms of goodness of fit and prediction accuracy (Dong et al., 2014b). To account for over dispersion and unobserved heterogeneity across individual sites, Dong et al. (2014a) used the multivariate random parameters zero-inflated negative binomial

(MVRPZINB) model in another crash frequency study, for which random parameters were assumed for the count part. Later, Anastasopoulos (2016) also adopted the multivariate random parameters zero-inflated negative binomial model in a crash frequency analysis, for which random parameters were assumed for both the count part and the zero-state part. Thus, the model is more flexible. In both studies, it was found that random parameter models were superior to fixed parameter models in terms of goodness of fit and prediction accuracy.

This paper presents the multivariate random parameters zero-inflated negative binomial model for analyzing urban midblock crashes by collision type. Here, midblock crashes refer to non-junction crashes that occurred on urban midblock segments bounded by signalized intersections. The objectives of this study were: (i) to identify important traffic operation and roadway geometric factors influencing urban midblock crash frequencies by collision type and (ii) to conduct a thorough review of the performance of the multivariate random parameters zero-inflated negative binomial model in accounting for unobserved heterogeneity produced by correlations across crash types, correlations across sites, excess zeros, and over dispersion. The results demonstrate the superiority of the multivariate random parameters zero-inflated negative binomial model to many common crash frequency analysis models.

4.2 Methodology

4.2.1 The Multivariate Zero-Inflated Negative Binomial Model

For an *m*-dimensional observation, $Y = (Y_1, Y_2, \dots, Y_m)$, the MVNB model is defined as (Dong et al., 2014a):

$$\begin{cases} Y_1 = Z_1 + U \\ Y_2 = Z_2 + U \\ \cdots \\ Y_m = Z_m + U \end{cases} \tag{4.1}$$

where m is dimension of Y, Z_1, Z_2,…, Z_m and U are independent NB variables with respective means λ_{10}, λ_{20}, …, λ_{m0} and λ_{00}.

An m-dimensional multivariate negative binomial model was constructed with $(m+1)$ independent negative binomial variables. The elements of Y are positively correlated with each other due to the presence of U, which is called the common negative binomial part in the following analysis. It can be proved that any marginal distribution of j variables of Y, where $j < m$, is still a j-dimensional multivariate negative binomial model.

The multivariate zero-inflated negative binomial model is an extension of the multivariate negative binomial model for multivariate zero-inflated data (Dong et al., 2014a; Li et al., 1999):

$$\begin{array}{c} (Y_1, Y_2, \dots, Y_m) \\ \sim (0,0,\dots,0) \;\; with\; probability \quad p_0 \\ \sim (NB(\lambda_1),0,\dots,0) \;\; with\; probability \quad p_1 \\ \sim (0, NB(\lambda_2),\dots,0) \;\; with\; probability \quad p_2 \\ \vdots \\ \sim (0,0,\dots, NB(\lambda_m)) \;\; with\; probability \quad p_m \\ \sim MVNB(\lambda_{10}, \lambda_{20}, \dots, \lambda_{m0}, \lambda_{00}) \;\; with\; probability \quad p_{11} \end{array} \tag{4.2}$$

where $p_0 + p_1 + p_2 + \cdots + p_{11} = 1$, $\lambda_j = \lambda_{j0} + \lambda_{00}$ for $j = 1, \dots, m$, and the MVNB model has the same definition as in Equation (1).

When Y follows the multivariate zero-inflated negative binomial distribution, the marginal distribution of Y_j is a univariate zero-inflated negative binomial model:

$$p(Y_j) = \begin{cases} \pi_j + (1-\pi_j)e^{-\lambda_j}, & Y_j = 0 \\ (1-\pi_j)\frac{\lambda_j{}^{y_j} e^{-\lambda_j}}{y_j!}, & Y_j = y_j \end{cases} \tag{4.3}$$

where $\pi_j = 1 - p_j - p_{11}$, is the probability of extra zeros, and $\lambda_j = \lambda_{j0} + \lambda_{00}$, is the mean of the NB part.

4.2.2 The Multivariate Random Parameters Zero-Inflated Negative Binomial Regression Model

A regression model was estimated to explore the influences of various factors on crash frequency. Since 10 years' of data were collected for each segment, a segment-specific random parameters model was adopted to account for possible unobserved heterogeneity across segments due to the panel structure. For the ith observation, $\lambda_i = (\lambda_{i10}, \lambda_{i20}, \dots, \lambda_{im0}, \lambda_{i00})$, the random parameters regression model is defined as:

$$\lambda_{ij0} = \exp\left(\beta_{midblock[i]j} X_i\right) * \exp(\varepsilon_{ij}) \tag{4.4}$$

$$\beta_{midblock[i]j} = \beta_j + \delta_{midblock[i]} \tag{4.5}$$

$$p_{ij} = \frac{\exp(\gamma_j * X_i)}{1 + \sum_{j=0}^{m} \exp(\gamma_j * X_i)} \tag{4.6}$$

$$\mathrm{p}_{i11} = 1 - \sum_{j=0}^{m} p_{ij} \tag{4.7}$$

where n is the number of data records; $i = 1, \dots, n$; m is the number of crash types; $j = 0,1, \dots, m$; $midblock[i] = 1, \dots, ngroup$; $ngroup$ is the number of midblock segments; K is the number of covariates, $k = 1, \dots, K$; $\beta_j (= \beta_{j0}, \beta_{j1}, \dots, \beta_{jK})$ is the coefficient vector in the count part of crash type j; $\delta_{midblock[i]} (= \delta_{midblock[i]0}, \delta_{midblock[i]1}, \dots, \delta_{midblock[i]K})$ is the random distributed error vector of regression coefficients in the count part of each segment; $\mathrm{X}_i (= 1, x_{i1}, x_{i2}, \dots, x_{iK})'$ is the covariate vector of the ith observation; $\exp(\varepsilon_{ij})$ is a gamma-distributed error term; $\gamma_j (= \gamma_{j0}, \gamma_{j1}, \dots, \gamma_{jK})$ is the regression coefficient vector of the zero-inflation part of crash type j; and $\mathrm{p}_i (= \mathrm{p}_{i0}, \mathrm{p}_{i1}, \dots, \mathrm{p}_{im}, \mathrm{p}_{i11})$ is the probability vector of the ith observation.

In this study, the parameters of the zero-inflation part are still assumed to be fixed, whereas the parameters of the zero-inflation part are still assumed to be fixed.

4.2.3 Model Estimation

The multivariate random parameters zero-inflated negative binomial model was estimated under the Bayesian framework with Markov-chain Monte Carlo (MCMC) simulation in JAGS (Just Another Gibbs Sampler) (Plummer, 2003). When conjugate priors were available, Gibbs sampling was used in JAGS. Otherwise, slicing sampling was used. R is a free programming language and software environment for statistical computing and graphics (R Core Team, 2016). JAGS was run in R using the 'runjags' package (Denwood, 2016), by which the parallel computation could be easily realized.

4.2.3.1 Prior distribution setting

Bayesian estimation requires prior distributions for the targeted unknown parameters, i.e. β_j's, $\delta_{midblock[i]}$'s, and γ_j's in this case. In this study, the priors were set as:

$$\beta_j \sim MVN(0, \Sigma_\beta^j) \tag{4.8}$$

$$\delta_{midblock[i]} \sim MVN(0, \Sigma_\delta) \tag{4.9}$$

$$\exp(\varepsilon_{ij}) \overset{iid}{\sim} Gamma(1/a_{ij}, 1/a_{ij}) \tag{4.10}$$

$$a_{ij} \sim Gamma(1000, 1000) \tag{4.11}$$

$$\gamma_j \sim MVN(0, \Sigma_\gamma^j) \tag{4.12}$$

$$\Sigma_\beta^j, \Sigma_\delta, \Sigma_\gamma^j \overset{iid}{\sim} inverse - Wishart(I_{K+1}, K+1) \tag{4.13}$$

where $\Sigma_\beta^j, \Sigma_\delta, \Sigma_\gamma^j$ are variance–covariance matrices, I is the identify matrix, and $\exp(\varepsilon_{ij})$ is set to have the same shape and rate parameter. This made the prediction easy, because the mean of $\exp(\varepsilon_{ij})$ was now one.

4.2.3.2 MCMC setting

Theoretically, the accuracy of estimated parameters would increase with the increase of sampling data, but the computing time would also increase. As a trade-off, three simulation chains were used with 35,000 iterations for each chain. The first 10,000 iterations were discarded

as warmup, and the next 25,000 iterations were used for parameter estimation with a thin interval of 5. Thus, 5,000 samples were produced for each chain. The initial values were randomly produced by JAGS. The trace plots and potential scale reduction factors of estimated parameters were checked to judge whether the posterior samples converged well. In addition, parallel computation was used to accelerate the MCMC process.

4.2.4 Model Checking and Comparison

4.2.4.1 Goodness of fit

Deviance information criteria (DIC) is a generalized version of Akaike Information Criterion (AIC) for evaluating hierarchical models (Spiegelhalter et al., 2002). Deviance is defined as $D(\theta) = -2log\big(p(y|\theta)\big)$, where y is the data, θ represents unknown parameters, and $p(y|\theta)$ is the likelihood function. DIC in JAGS was defined as (Plummer, 2002):

$$DIC = \bar{D} + pD \tag{4.14}$$

$$pD = E\left[E_{Y_{rep}|\theta^0}\left[log\left\{\frac{p(Y_{rep}|\theta^0)}{p(Y_{rep}|\theta^1)}\right\}\right]\right] \tag{4.15}$$

where $\bar{D}$ is the mean of the sampled deviances from simulations, pD is the effective number of parameters, θ^0 and θ^1 are two independent samples from the posterior distribution of θ, Y_{rep} is an independent replicate data set derived from the same data-generating mechanism as the observed data. The definition of pD in JAGS (Plummer, 2002) is slightly different from the one from Spiegelhalter et al. (2002), where $pD = \bar{D} - D(\bar{\theta})$, and $\bar{\theta}$ is the expectation of θ.

$\bar{D}$ is a measure of how well the model fits the data, whereby a smaller $\bar{D}$ value means the model fits the data better. pD shows the diffusion of posterior samples (Plummer, 2002). The larger the pD, the more diffuse the posterior samples. It is a measure of model complexity, whereby a smaller pD value means the model is less complex. Thus, DIC is a generalized penalized expected deviance of Akaike Information Criteria in Bayesian analysis. Bayesian

models with smaller DIC values are desired. Roughly, differences of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are substantial, and differences less than 5 might mean that the models are not

4.2.4.2 Prediction accuracy

Although DIC could be used for model comparison, it cannot evaluate the quality of fit of the model to the observed data. Root mean square error (RMSE) of prediction was used to evaluate the prediction accuracy of models. Similar to DIC, smaller RMSE values are desired.

$$RMSE = \sqrt{\frac{1}{n_0}\sum_{j=1}^{n_0}\left(O_j - P_j\right)^2} \tag{4.16}$$

where O_j is the jth observation value, P_j is the predicted ith value from the model, and n_0 is the number of observations.

4.3 Data Description

Yearly crash frequency data per direction for 1,506 urban midblock segments in Lincoln and Omaha, Nebraska from 2003 to 2012 were collected from the Nebraska Department of Roads. Originally, these midblock segments were selected by a technical committee from the Nebraska Department of Roads to investigate the effects of narrow lane width on urban roadway safety (Sharma et al., 2015), for which researchers focused mainly on regular vehicle crashes, and thus excluded animal crashes, alcohol-related crashes, crashes caused by road surface conditions, and heavy vehicle crashes. Sideswipe (same direction) and rear-end crashes made up 18.9% and 57.5% of the crash data, respectively, whereas most of the remaining crashes were recorded as not applicable. Thus, crashes were classified into three major types: sideswipe (same direction) crashes, rear-end crashes, and other crashes. The first two crash types were the focus of this study, but other crashes were still used in the modeling analysis, as it was believed that they might have some underlying correlations to the first two crash types, and could be utilized

to better explore the characteristics of sideswipe (same direction) and rear-end crashes in multivariate models.

In addition to crash data, many traffic operation and roadway geometric data were also collected by field study and measurements in Google Earth. A summary of collected variables is given in Table 4-1. Each midblock segment was homogenous with respect to annual average daily traffic per lane, number of through lanes, median type, left-turn treatment, and other key factors. The annual average daily traffic per lane per direction for each segment was obtained from the Nebraska Department of Roads. The lane widths of these segments included 9-ft, 10-ft, 11-ft, and 12-ft widths, and 12-ft width was used as the baseline lane width in modeling. The speed limits for these segments included 25 mph, 35 mph, 40 mph, and 45 mph, and 25 mph was used as the baseline speed limit in modeling. These segments were also classified into four groups by the National Functional Classification (NFC) system (Federal Highway Administration, 2013): NFC-14, urban principal arterial–other connecting link; NFC-15, urban principal arterial–other non-connecting link; NFC-16, urban minor arterial; and NFC-17, urban collector. NFC-17 was used as the baseline roadway class in modeling.

Variances of all three crash types were larger than their means (Table 4-1), implying over-dispersion existed for all of them. The percentages of zero values of sideswipe (same direction), rear-end, and other crashes were 81.4%, 65.2%, and 77.6%, respectively, larger than the expected probabilities of zero values (78.7%, 48.5%, and 74.1%) of Poisson distributions with the means 0.240, 0.724, and 0.300, respectively. This indicated that excess zeros existed for all three crash types, which also could be visualized in the histograms of crash data in Figure 4-1.

Table 4-1 *Descriptive statistics of collected variables*

Name	Description	Mean	Std. err.	Min.	Max.	Zero-proportion
Response variables						
Sideswipe (same direction)	Number of sideswipe (same direction) crashes per direction per segment per year	0.240	0.568	0	9	81.4%
Rear-end	Number of rear-end crashes per direction per segment per year	0.724	1.576	0	43	65.2%
Others	Number of rest crashes per direction per segment per year	0.300	0.641	0	8	77.6%
Independent variables						
Number of lanes	Number of through lanes	1.929	0.005	1	6	
Annual average daily traffic per lane	1,000 vehicles	5.668	0.019	0.100	13.975	
Segment length	Miles	0.363	0.002	0.025	2.003	
Shoulder indicator	1, shoulder exists (25.4%); 0, no shoulder (74.6%)					
Median indicator	1, median exists (79.5%); 0, no median (20.5%)					
On-street parking indicator	1, on-street parking exists (5.6%); 0, no on-street parking (94.4%)					
Central business district indicator	1, in central business district (6.1%); 0, out of central business district (93.9%)					
One-way road indicator	1, roadway is one-way (3.7%); 0, roadway is two-way (96.3%)					
Lane width	Feet (ft): 9 ft (3.1%); 10 ft (15.5%); 11 ft (29.8%); 12 ft (51.6%)					
Lane width – 9 ft indicator	1, lane width is 9 ft; 0, otherwise					
Lane width – 10 ft indicator	1, lane width is 10 ft; 0, otherwise					
Lane width – 11 ft indicator	1, lane width is 11 ft; 0, otherwise					
Speed limit	Mph: 25 (6.0%); 35 (32.6%); 40 (30.7%); 45 (30.7%)					
Speed limit – 35 mph indicator	1, speed limit is 35 mph; 0, otherwise					
Speed limit – 40 mph indicator	1, speed limit is 40 mph; 0, otherwise					
Speed limit – 45 mph indicator	1, speed limit is 45 mph; 0, otherwise					
National functional classification (NFC)	NFC-14: urban principal arterial–other connecting link, 13.9%; NFC-15: urban principal arterial–other non-connecting link, 37.6%; NFC-16: urban minor arterial, 41.2%; NFC-17: major collector, 7.3%.					
NFC-14 indicator	1, segment belongs to NFC-14; 0, otherwise					
NFC-15 indicator	1, segment belongs to NFC-15; 0, otherwise					
NFC-16 indicator	1, segment belongs to NFC-16; 0, otherwise					
City indicator	1, Omaha (68.3%); 0, Lincoln (31.7%)					

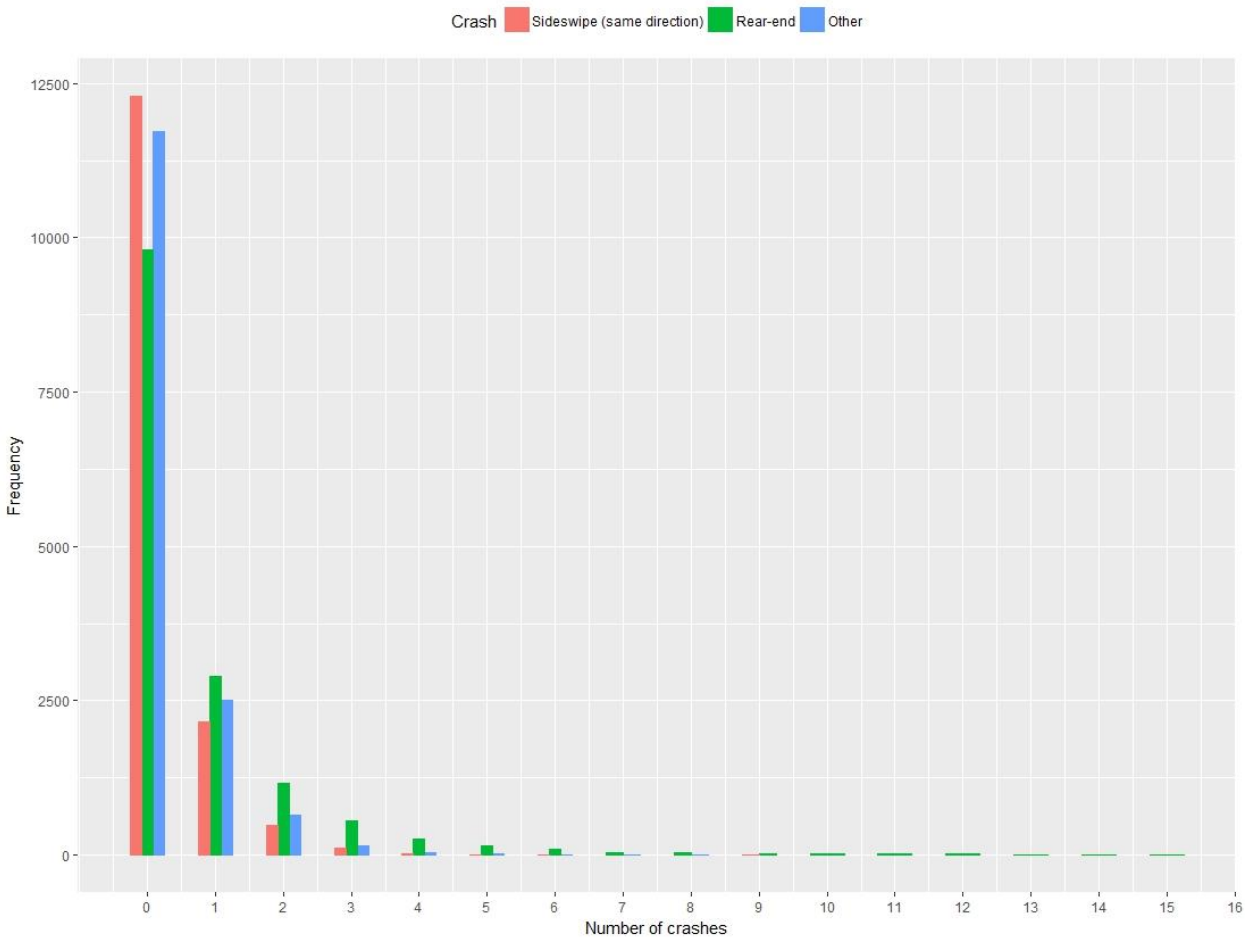

Figure 4-1 *Histogram of sideswipe (same direction), rear-end, and other crashes from 2003 to 2012*

4.4 Results and Discussions

Out of the 10 years of data, data from 2003 to 2011 were used for the model estimation, and the 2012 data were used for prediction.

4.4.1 Model Comparison

In addition to the multivariate random parameters zero-inflated negative binomial model, the multivariate Poisson log-normal model, the univariate random parameters zero-inflated Poisson model, the univariate random parameters zero-inflated negative binomial model, the

multivariate zero-inflated Poisson model, the multivariate zero-inflated negative binomial model, and the multivariate random parameters zero-inflated Poisson model were also estimated for comparison. The DIC and RMSE values of these models are shown in Table 4-2.

Table 4-2 *DIC and RMSE values of all the estimated models*

				RMSE		
Model	$\bar{D}$	pD	DIC	Sideswipe (same direction)	Rear-end	Other
MVPLN	51,939	6,480	58,419	0.500	1.345	0.582
MVP	66,342	1,870	68,212	0.507	1.336	0.583
MVNB	51,114	10,210	61,324	0.498	1.338	0.575
MVZIP	61,020	64	61,084	0.487	1.311	0.564
MVZINB	50,293	2,468	52,761	0.487	1.307	0.573
URPZIP	49,375	14,706	64,081	0.474	1.146	0.551
URPZINB	45,775	12,608	58,383	0.473	1.147	0.550
MVRPZIP	53,855	912	54,767	0.472	1.147	0.553
MVRPZINB	46,821	2,937	49,758	0.471	1.138	0.552

Note: DIC, Deviance information criteria; RMSE, root mean square error; $\bar{D}$, mean of the sampled deviances from Markov-chain Monte Carlo simulations; pD, effective number of parameters in the model; MVPLN, multivariate Poisson log-normal; MVP, multivariate Poisson; MVNB, multivariate negative binomial; MVZIP, multivariate zero-inflated Poisson; MVZINB, multivariate zero-inflated negative binomial; URPZIP, univariate random parameters zero-inflated Poisson; URPZINB, univariate random parameters zero-inflated negative binomial; MVRPZIP, multivariate random parameters zero-inflated Poisson; MVRPZINB, multivariate random parameters zero-inflated negative binomial.

From Table 4-2, the following observations can be made:

1. DIC and RMSE values of the multivariate random parameters zero-inflated negative binomial model were generally much lower than those of all the other models, showing its superiority.

2. Compared to the multivariate Poisson/negative binomial models, the multivariate zero-inflated Poisson/negative binomial models had much smaller DIC and RMSE values, respectively, which shows the superiority of multivariate zero-inflated models for analyzing the multivariate crash data with excess zeros.

3. Compared to the multivariate zero-inflated Poisson/negative binomial models, the multivariate random parameters zero-inflated Poisson/negative binomial models showed much better performance in terms of DIC and RMSE. Although the multivariate zero-inflated Poisson/negative binomial models had lower pD values, their $\overline{D}$ values were much higher. This means that the multivariate random parameters zero-inflated Poisson/negative binomial models are more complex but fit the data much better. The result is straightforward, as the random parameters models allow estimated parameters to vary across segments to account for unobserved heterogeneity. This flexibility improves the model's ability to fit the data. This finding reiterates that the unobserved heterogeneity across observations in crash analyses may not be ignored (Mannering et al., 2016).

4. The RMSE values of the univariate random parameters zero-inflated Poisson/negative binomial models and the multivariate random parameters zero-inflated Poisson/negative binomial models were similar, but the latter models had much lower DIC values. The univariate random parameters zero-inflated Poisson/negative binomial models had relatively lower $\overline{D}$ but much higher pD values, which indicated that they fit the data better but were more complex. As mentioned above, the multivariate models could account for unobserved heterogeneity across crash types. By borrowing from the strength of between-crash correlations, multivariate models could estimate parameters more accurately than univariate models. This result shows the importance of multivariate modeling in analysis of multiple crash types.

5. All the negative binomial models had much lower DIC values than did their corresponding Poisson models, but their RMSE values were very close, such as the multivariate random parameters zero-inflated Poisson model versus the multivariate random parameters zero-inflated negative binomial model. Considering that the only difference between the negative

binomial models and their Poisson counterparts was that the negative binomial models had dispersion parameters but the Poisson models did not, the results suggest that the estimated parameters of the Poisson and negative binomial models were similar except for the dispersion parameters. The negative binomial models fit the data much better, as they could account for over dispersion of crash data. The results highlight that, although random parameters and zero-inflated models can also account for over dispersion to some degree, they might not cover all of it. It may be still necessary to specifically take over dispersion into account in crash frequency analyses.

6. The most popular multivariate count data model, the multivariate Poisson log-normal model, performed worse than the multivariate zero-inflated negative binomial model did in terms of both DIC and RMSE. Because Dong et al. (2014b) showed that the multivariate zero-inflated Poisson model was superior to the multivariate Poisson log-normal model for their dataset, it was believed that the multivariate zero-inflated count data models were competitive alternatives to the multivariate Poisson log-normal model for analyzing the multivariate zero-inflated data.

In general, as shown in Table 4-2, unobserved heterogeneity stemmed from the correlations across crash types, the correlations across segments, excess zeros, and over dispersion for the studied dataset, and none of them can be ignored. The multivariate random parameters zero-inflated negative binomial model was superior to other models as it could account for various unobserved heterogeneities.

Because many independent variables were found to be not significant for the multivariate random parameters zero-inflated negative binomial model, it was re-run after removing those nonsignificant variables, and the results are discussed in the following analysis.

4.4.2 Parameter Interpretation

The means and 95% credible intervals of the estimated parameters of sideswipe (same direction), rear-end, and other crashes count parts of the multivariate random parameters zero-inflated negative binomial model and the multivariate zero-inflated negative binomial models are shown in Table 4-3 and Table 4-4, respectively. For the multivariate random parameters zero-inflated negative binomial model, if the standard deviation of parameter density function is statistically not significant, that parameter would be fixed. Probabilities of estimated parameters being negative and average marginal effects of the multivariate random parameters zero-inflated negative binomial model are shown in Table 4-5 and Table 4-6, respectively. Only significant variables are shown in these tables, and only the variables with both means and standard deviations significant were considered to be significant.

Number of lanes showed significant effects only for sideswipe (same direction) crashes. When the number of lanes increased, 89.6% of segments had more sideswipe (same direction) crashes, and on average, the number of sideswipe (same direction) crashes increased 40.9% with a one lane increase. This finding is reasonable, as with more lanes, vehicles have more opportunities to travel parallel to each other on segments. In addition, 10.4% of segments tended to have fewer sideswipe (same direction) crashes with an increase in the number of lanes. Number of lanes did not show significant effects on rear-end or other crash types. Although more lanes might bring more traffic, drivers also have more space to maneuver to avoid crashes and they may also drive more carefully. Thus, these effects might offset each other.

Table 4-3 *Posterior summary (means and 95% credible intervals) of estimated parameters of the count part of the multivariate random parameters zero-inflated negative binomial model*

Variables	Sideswipe (same direction) crashes	Rear-end crashes	Other crashes
Number of lanes	0.343 (0.063, 0.623)	-	-
SD	0.272 (0.203, 0.349)	-	-
Annual average daily traffic per lane	0.070 (0.012, 0.121)	0.210 (0.169, 0.256)	-
SD	0.147 (0.122, 0.171)	0.145 (0.120, 0.171)	-
Median indicator	–0.356 (–0.530, –0.146)	-	–0.339 (–0.518, -0.203)
SD	0.368 (0.255, 0.520)	-	0.381 (0.264, 0.536)
Speed limit – 45 mph indicator	–0.735 (–0.943, –0.506)	–0.417 (–0.573, –0.200)	–0.414 (–0.579, –0.252)
SD	0.388 (0.248, 0.511)	0.437 (0.305, 0.579)	0.382 (0.271, 0.528)
NFC-15 indicator	–0.276 (–0.464, –0.099)	–0.229 (–0.462, –0.027)	-
SD	0.348 (0.216, 0.500)	0.471 (0.285, 0.649)	-
NFC-16 indicator	–0.543 (–0.715, –0.366)	–0.269 (–0.411, -0.097)	-
SD	0.367 (0.259, 0.501)	0.341 (0.257, 0.426)	-
City indicator	–0.564 (–0.728, –0.378	–0.816 (–0.971, –0.681)	–0.632 (–0.778, –0.481)
SD	0.348 (0.249, 0.459)	0.457 (0.301, 0.579)	0.394 (0.275, 0.509)
Constant	–1.635 (–2.269, –0.889)	–1.183 (–1.783, –0.597)	–0.553 (–0.889, –0.538)
SD	0.426 (0.245, 0.600)	0.455 (0.253, 0.699)	0.403 (0.272, 0.582)
# of significant variables	8	6	4

Note: SD: standard deviation of parameter density function; NFC, national functional classification; values are the posterior means; values in parentheses show the 95% credible intervals; "-", insignificant variables at the 95% credible level; shoulder indicator, on-street parking indicator, central business district indicator, segment length, one-way indicator, lane width (9 ft, 10 ft, and 11 ft) indicators, speed limit - 35 mph indicator, speed limit - 40 mph indicator, and NFC-14 indicator were not significant variables at the 95% credible level for any crash type.

Annual average daily traffic has been widely found to have a positive effect on crash frequency when it is assumed to have a fixed effect (Bonneson and Mccoy, 1997; Dong et al., 2014a, 2014b; Ferreira and Couto, 2015; Greibe, 2003; Zhang et al., 2012); however, this may not always be true. In this study, annual average daily traffic per lane showed a significant influence on the number of sideswipe (same direction) and rear-end crashes. The estimated

parameters were normally distributed with a mean of 0.070 (standard deviation of parameter density function = 0.147) for the sideswipe (same direction) crash and a mean of 0.210 (standard deviation of parameter density function = 0.145) for the rear-end crash. On average, the numbers of sideswipe (same direction) and rear-end crashes increased by 7.3% and 23.4%, respectively, as annual average daily traffic per lane increased by 1,000 vehicles. Even though the number of sideswipe (same direction) and rear-end crashes increased for most segments with an increase of annual average daily traffic per lane, these numbers decreased for 31.7% and 7.4% of segments, respectively. Although an increase in annual average daily traffic per lane increases crash opportunities, it could also provide some underlying safety effects, such as more cautious driving, intensive traffic enforcement, and advanced traffic control devices, which could offset the increased crash risk. Thus, the increase in annual average daily traffic per lane did not necessarily increase the number of crashes. However, this does not mean that crash frequencies would not increase or even decrease with a continuing increase of annual average daily traffic per lane. A summary of annual average daily traffic per lane by crash types and signs of estimated regression coefficients is shown in Table 4-7. The segments with positive coefficients for annual average daily traffic per lane generally had much higher annual average daily traffic per lane than did those with non-positive coefficients. That is, for segments already with very high annual average daily traffic per lane, the crash frequency was more likely to increase with an increase of annual average daily traffic per lane. In addition, it should be noted that segments with non-positive coefficients of annual average daily traffic per lane for all three crash types had a mean annual average daily traffic per lane value of around 5.5, which seemed an important threshold. Mannering et al. (2016) proposed that there might be heterogeneous linear or non-linear relationships between traffic volume and accident likelihood, which is proved somewhat

by this study. Anastasopoulos (2016) also showed that, using the multivariate random parameters zero-inflated negative binomial model, annual average daily traffic had inconsistent influences on crash frequencies for roadway segments in Indiana.

Table 4-4 *Posterior summary (means and 95% credible intervals) of estimated parameters of the count part of the multivariate zero-inflated negative binomial model*

Variables	Sideswipe (same direction) crashes	Rear-end crashes	Other crashes
Number of lanes	0.732 (0.618, 0.834)	0.637 (0.568, 0.714)	–0.128 (–0.208, –0.052)
Annual average daily traffic per lane	0.128 (0.099, 0.158)	0.290 (0.270, 0.312)	0.030 (0.005, 0.056)
Shoulder indicator	-	0.131 (0.039, 0.227)	-
Median indicator	–0.208 (–0.352, –0.070)	-	–0.148 (–0.269, –0.032)
On-street parking indicator	-	-	0.491 (0.230, 0.739)
Central business district indicator	–0.519 (–0.900, –0.209)	–0.553 (–0.845, –0.345)	-
Segment length	0.985 (0.630, 1.326)	0.895 (0.672, 1.083)	1.727 (1.479, 2.009)
One-way road indicator	-	–0.581 (–0.982, –0.077)	-
Lane width – 9 ft indicator	–0.416 (–0.697, –0.150)	–0.361 (–0.562, –0.176)	-
Lane width – 11 ft indicator	–0.149 (–0.277, –0.029)	-	-
Speed limit – 40 mph indicator	–0.248 (–0.423, –0.040)	–0.468 (–0.566, –0.374)	-
Speed limit – 45 mph indicator	–0.634 (–0.825, –0.474)	–0.391 (–0.506, –0.291)	–0.628 (–0.782, –0.467)
NFC-14 indicator	–0.420 (–0.691, –0.172)	-	0.272 (0.032, 0.509)
NFC-15 indicator	–0.311 (–0.615, –0.063)	-	0.264 (0.006, 0.507)
NFC-16 indicator	–0.569 (–0.838, –0.292)	-	-
City indicator	–0.293 (–0.409, –0.180)	–0.409 (–0.494, –0.330)	–0.140 (–0.239, –0.040)
Constant	–2.527 (–2.977, –2.014)	–2.696 (–3.167, –2.302)	–1.292 (–1.567, –1.051)
# of significant variables	14	11	10

Note: NFC, national functional classification; values are the posterior means; values in parentheses show the 95% credible intervals; "-", insignificant variables at the 95% credible level; lane width – 10ft indicator and speed limit – 35 mph indicator were not significant variables at the 95% credible level for any crash type.

Segment length did not show a significant influence on any crash type. Most segments studied were very short, the average segment length being 0.363 mile and 75.6% of segments

being shorter than 0.5 mile. It is thought that these segment lengths might not be different enough to show significant effects.

Table 4-5 *Probabilities of the estimated parameters being negative for the count part of the multivariate random parameters zero-inflated negative binomial model*

Variables	Sideswipe (same direction) crashes	Rear-end crashes	Other crashes
Number of lanes	0.104	-	-
Annual average daily traffic per lane	0.317	0.074	-
Median indicator	0.833	-	0.813
Speed limit – 45 mph indicator	0.971	0.830	0.861
NFC-15 indicator	0.786	0.687	-
NFC-16 indicator	0.931	0.785	-
City indicator	0.947	0.963	0.946

Note: NFC, national functional classification; "-", insignificant variables at the 95% credible level; shoulder indicator, on-street parking indicator, central business district indicator, segment length, one-way indicator, lane width (9 ft, 10 ft, and 11 ft) indicators, speed limit – 35 mph indicator, speed limit – 40 mph indicator, and NFC-14 indicator were not significant variables at the 95% credible level for any crash type.

Table 4-6 *Average marginal effects of the count part of the multivariate random parameters zero-inflated negative binomial model*

Variables	Sideswipe (same direction) crashes	Rear-end crashes	Other crashes
Number of lanes	0.409	-	-
Annual average daily traffic per lane	0.073	0.234	-
Median indicator	–0.300	-	0.288
Speed limit – 45 mph indicator	–0.520	–0.341	–0.339
NFC-15 indicator	–0.241	–0.205	-
NFC-16 indicator	–0.419	–0.236	-
City indicator	–0.431	–0.558	–0.468

Note: NFC, national functional classification; "-", nonsignificant variables at the 95% credible level shoulder indicator, on-street parking indicator, central business district indicator, segment length, one-way indicator, lane width (9 ft, 10 ft, and 11 ft) indicators, speed limit – 35 mph indicator, speed limit – 40 mph indicator, and NFC-14 indicator were not significant variables at the 95% credible level for any crash type.

Table 4-7 *Summary of annual average daily traffic per lane by crash types and signs of regression coefficients*

Crash type	Coefficient	Annual average daily traffic per lane (1,000 vehicles)		
		Min	Mean	Median
Crash type	Coefficient	AADT per lane (1,000 vehicles)		
		Min	Mean	Median
Sideswipe (same direction)	Positive	5.759	9.290	9.481
	Non-positive	0.100	5.602	5.700
Rear-end	Positive	2.781	7.724	7.625
	Non-positive	0.100	5.214	5.250
Others	Positive	-	-	-
	Non-positive	0.100	5.617	5.713

Note: "-", unavailable.

It should be noted that crash frequency is usually assumed to increase with an increase in the number of lanes, annual average daily traffic, and segment length; thus, many studies have used these variables as exposure variables (Boulieri et al., 2017; Miaou et al., 2003; Miaou and Song, 2005). As presented in Table 4-4, the estimated parameters of the multivariate zero-inflated negative binomial model are generally consistent with these beliefs, whereby number of lanes, annual average daily traffic per lane, and segment length showed positive effects on all crash types, except for number of lanes for other crash types. The inconsistent findings of the multivariate zero-inflated negative binomial model and the multivariate random parameters zero-inflated negative binomial models show the advantage of random parameter models that they can capture the segment-specific effects, which are unavailable in fixed parameter models but very important, especially when opposite segment-specific effects exist. This also suggests that researchers should be very careful in using these variables as exposure variables, as the precondition might be violated.

The presence of a shoulder had no significant influence on any crash type. For freeways or rural highways, the shoulder is very important in the event of emergency or breakdown. However, on urban arterials, these events may not interrupt traffic seriously due to lower travel

speeds, better roadway lighting, and more access points for leaving the roadway. Thus, the lack of a shoulder may not have influenced traffic safety for the studied roadways. The results are consistent with the study by Zhao et al. (2017), who found that a shoulder did not have significant effects on crash frequencies of urban signalized intersection approaches in either Lincoln or Omaha, Nebraska.

However, the presence of a median had a significant influence on the number of sideswipe (same direction) crashes with a mean of –0.356 (standard deviation of parameter density function = 0.368), and other crashes with a mean of –0.339 (standard deviation of parameter density function = 0.381). When a median was present, 83.3% and 81.3% of segments had fewer sideswipe (same direction) and other crash types, respectively. On average, the number of sideswipe (same direction) and other crash types decreased by 30.0% and 28.8%, respectively. When a road median is present, left-turn and U-turn traffic is expected to decrease, leading to fewer sideswipe (same direction) collisions. This could also reduce sideswipe (opposite direction) crashes, angle crashes, and so on, which may explain why the number of other crash types decreased for most segments. However, the number of vehicle collisions with medians may increase; thus, some segments might have more crashes.

The presence of on-street parking, being in a central business district, or one-way traffic did not show significant influence on the number of any crash type. The speed limit characteristics of segments with on-street parking, in a central business district, or with one-way traffic are shown in Table 4-8. Most of these segments had speed limits of 25 mph or 35 mph. Under such low-speed environments, these factors would not be expected to pose significant threats to traffic safety.

Table 4-8 *Speed limits of segments with on-street parking, segments in central business district, and one-way traffic*

	Speed Limit (mph)				
Segment Description	25	35	40	45	Sum
Segments with on-street parking	71.8%	27.1%	1.1%	0	100%
Segments in central business district	64.1%	35.9%	0	0	100%
One-way segments	49.1%	50.9%	0	0	100%

Narrow lanes are often needed in cities to accommodate parking, bike lanes, sidewalks, drainage, and utilities. Although it is intuitive that providing some buffer space might prevent the occurrences of crashes, past studies evaluating the impact of narrower lane width on urban roadway safety have revealed inconsistent results: negative effects (Harwood, 1990), non-linear effects (Lee et al., 2015; Park and Abdel-Aty, 2016), and no effects (Potts et al., 2007). The multivariate random parameters zero-inflated negative binomial model showed that lane width did not have a significant influence on any crash type in this dataset. Although narrow lanes might increase the opportunities for some collision types, such as sideswipe (same direction) crashes, they might also have lower speed limits, less traffic, and less aggressive driving. The net effect of these opposite forces determines the impact of narrow lanes on crash frequency. The findings of this study suggest that, for the studied roadways, safety might not be a concern if lanes need to be made narrower to accommodate other street elements.

Compared with a 25-mph speed limit, 35-mph and 40-mph speed limits did not show significant influences on midblock crash frequencies, but a 45-mph speed limit did show significant effects. For the 45-mph speed limit, the estimated normally distributed parameters had a mean of –0.735 (standard deviation of parameter density function = 0.388) for sideswipe (same direction) crashes, a mean of –0.417 (standard deviation of parameter density function = 0.437) for rear-end crashes, and a mean of –0.414 (standard deviation of parameter density

function = 0.382) for other crash types. That is, 97.1%, 83.0%, and 86.1% of the segments tended to have fewer sideswipe (same direction), rear-end, and other crashes, respectively, with a 45-mph speed limit than with lower speed limits. Simultaneously, only 2.9%, 17.0%, and 13.9% of segments tended to have more sideswipe (same direction), rear-end, and other crash types, respectively. Intuitively, it would seem that higher speed limits would increase the probability of crashes occurring, but roadways with high speed limits usually have fewer access roads and better designed facilities. Thus, it appears that the advantages of high speed limits outweighed the disadvantages for most segments. On average, sideswipe (same direction), rear-end, and other crash types decreased by 52.0%, 34.1%, and 33.9%, respectively, on segments with a 45-mph speed limit compared with those with lower speed limits. This study's findings suggest that 45 mph is an important threshold in determining speed limits for urban arterials. For the multivariate zero-inflated negative binomial model, the speed limit was also found to have negative effects on all crashes. However, although an increased speed limit might reduce the number of crashes and increase capacity for most segments, it might also increase the severity of crash damage and injuries (Malyshkina and Mannering, 2008; Renski et al., 1999), as the outcomes of high-speed object collisions are more serious. Thus, speed limit increases should be carefully studied before implementation.

Compared with major collectors (NFC-17), urban principal arterial–other non-connecting link (NFC-15) and urban minor arterial (NFC-16) showed significant influences on the number of sideswipe (same direction) and rear-end crashes. Compared to NFC-17 segments, 78.6% and 86.7% of NFC-15 segments had fewer sideswipe (same direction) and rear-end crashes, respectively, and 93.1% and 78.5% of NFC-16 segments had fewer sideswipe (same direction) and rear-end crashes, respectively. Speed limit compositions, as well as mean and median annual

average daily traffic values of segments by functional classification, are shown in Table 4-9. The mean speed limits of NFC-15 and NFC-16 segments were higher than those of NFC-17 segments, and it was proved above that the number of crashes tended to decrease with higher speed limits. Thus, this might explain why most NFC-15 and NFC-16 segments tended to have fewer crashes. However, urban principal arterial–other connecting link (NFC-14) segments did not show significant influences on any crash type, although they had the highest speed limits. A possible explanation is that the NFC-14 segments also had very large annual average daily traffic values, which probably led to the occurrence of more crashes. These factors might play different roles for segments by functional classification, leading to different results. In addition, speed limit and annual average daily traffic reflect the mobility function of roadways, whereas accessibility is another function in determining NFC levels for roadways (Federal Highway Administration, 2013). Although accessibility information was unavailable in this dataset, it may also influence crash frequencies.

Table 4-9 *Speed limit compositions, and mean and median annual average daily traffic values of segments by national function classification*

National Functional Classification	Speed Limit (mph)				Annual average daily traffic (1,000 vehicles)	
	25	35	40	45	Mean	Median
NFC-14	0	22.9%	30.5%	46.7%	15.4	14.7
NFC-15	0.7%	30.2%	25.0%	44.1%	12.6	12.2
NFC-16	3.9%	38.2%	41.0%	16.9%	8.91	8.43
NFC-17	56.9%	32.1%	1.8%	9.2%	4.11	3.63

Note: NFC, national functional classification.

The city indicator (Lincoln vs. Omaha) showed a significant influence on all crash types. The estimated normally distributed parameters had a mean of –0.564 (standard deviation of parameter density function = 0.348) for sideswipe (same direction) crashes, a mean of –0.816 (standard deviation of parameter density function = 0.457) for rear-end crashes, and a mean of –

0.632 (standard deviation of parameter density function = 0.394) for other crash types. The probabilities of the city variable being negative were 86.3%, 95.1%, and 90.7% for the three crash types, respectively. The number of sideswipe (same direction), rear-end, and other crashes for segments in Omaha were lower than those in Lincoln by an average of 25.4%, 33.6%, and 13.1%, respectively, when other characteristics were same. It should be noted that signalized intersection approaches in Omaha were also found to have fewer crashes than did those in Lincoln (Zhao et al., 2017). Considering that Lincoln and Omaha are only 45-min driving time apart, driving behaviors in the two cities are expected to be similar. Thus, some other features, such as traffic enforcement, land use, and terrain, might be responsible for this difference. Further studies are needed to investigate the true reasons, which would be very helpful for transportation agencies in formulating accurate countermeasures to improve traffic safety in Lincoln.

The estimated parameters of the zero-inflation part of the multivariate random parameters zero-inflated negative binomial model and the multivariate zero-inflated negative binomial models are shown in Table 4-10 and Table 4-11, respectively. Although both models adopted fixed parameters for the zero-inflation part, their significant variables were very different due to different count part models. This indicates that, for zero-inflated models, the count parts and zero-inflated parts were highly correlated and that a modeling framework change in one part would greatly influence the result of the other part. For both models, the number of lanes, annual average daily traffic per lane, and segment length showed significantly negative effects on the number of some crash types, which means that with the increase of the values of these covariates, these crash types were less likely to have zero values. That is, the expected crash frequencies would increase. It is reasonable to infer, as has been proved, that for most segments,

these covariates have positive effects on crash frequencies. In addition, a 45-mph speed limit showed significant positive effect on the number of sideswipe (same direction) crashes for the multivariate random parameters zero-inflated negative binomial model, which means that zero crashes were more likely to appear under the 45-mph speed limit. This result is also consistent with the results of the count part of the multivariate random parameters zero-inflated negative binomial model.

Table 4-10 *Posterior summary (means and 95% credible intervals) of estimated parameters of the zero-inflation part of the multivariate random parameters zero-inflated negative binomial model*

Variables	Sideswipe (same direction) crashes	Rear–end crashes	Other crashes
Number of lanes	-	-	–5.823 (-8.643, –3.684)
Annual average daily traffic per lane	–1.192 (–1.664, –0.719)	-	–0.331 (–0.571, –0.091)
Median indicator	-	1.990 (0.039, 5.466)	-
Central business district indicator	-	–7.451 (–13.851, -1.680)	-
Segment length	–13.512 (–18.672, –8.768)	–17.910 (–26.774, –10.887)	-
Speed limit – 45 mph indicator	2.647 (1.007, 4.272)	-	-
NFC-15 indicator	-	-	–2.727 (–5.703, –0.047)
NFC-16 indicator	5.419 (1.154, 12.108)	2.730 (0.121, 6.440)	–4.615 (–7.680, –2.202)
City indicator	–5.874 (–10.602, –1.451)	-	-
Constant	-	-	7.696 (5.119, 10.595)
# of significant variables	6	4	5

Note: NFC, national functional classification; values shown are posterior means; values in parentheses show the 95% credible intervals; "-", nonsignificant variables at the 95% credible level. Shoulder indicator, on-street parking indicator, one-way indicator, lane width (9ft, 10ft, and 11ft) indicators, speed limit – 35mph indicator, speed limit – 40mph indicator, and NFC-14 indicator were not significant variables at the 95% credible level for any crash type.

Table 4-11 *Posterior summary (means and 95% credible intervals) of estimated parameters of the zero-inflation part of the multivariate zero-inflated negative binomial model*

Variables	Sideswipe (same direction) crashes	Rear–end crashes	Other crashes
Number of lanes	-	-	-
Annual average daily traffic per lane	-0.846 (-2.997, -0.099)	-0.811 (-3.176, 0.407)	-1.264 (-1.651, -0.835)
Shoulder indicator	-	-	1.641 (0.838, 2.499)
Median indicator	2.858 (1.103, 4.912)	-	-
Segment length	-7.934 (-11.899, -0.064)	-	-
Lane width – 10 ft indicator	-	-	1.152 (0.128, 2.179)
NFC - 16 indicator	-	-	-1.744 (-2.933, -0.550)
# of significant variables	4	1	4
Variables	Sideswipe (same direction) crashes	Rear–end crashes	Other crashes
Number of lanes	-	-	-
Annual average daily traffic per lane	-0.846 (-2.997, -0.099)	-0.811 (-3.176, 0.407)	-1.264 (-1.651, -0.835)
Shoulder indicator	-	-	1.641 (0.838, 2.499)
Median indicator	2.858 (1.103, 4.912)	-	-
Segment length	-7.934 (-11.899, -0.064)	-	-
Lane width – 10 ft indicator	-	-	1.152 (0.128, 2.179)
NFC - 16 indicator	-	-	-1.744 (-2.933, -0.550)
# of significant variables	4	1	4

Note: NFC, national functional classification; values shown are posterior means; values in parentheses show the 95% credible intervals; "-", nonsignificant variables at the 95% credible level. On-street parking indicator, central business district indicator, one-way indicator, lane width – 9ft indicator, lane width – 11ft indicator, speed limit (35mph, 40mph, and 45mph) indicators, NFC-14 indicator, and NFC-15 indicator, city indicator, and number of lanes were not significant variables at the 95% credible level for any crash type.

Out of total 18 covariates, 9 and 16 covariates were found to be significant (at least in either the count part or the zero-inflation part) for the multivariate random parameters zero-inflated negative binomial model and the multivariate zero-inflated negative binomial model, respectively. That is, the multivariate random parameters zero-inflated negative binomial model

showed a much better performance with fewer variables than did the multivariate zero-inflated negative binomial model in this case, which was helpful for identifying critical crash-influencing factors. However, this may not be true in other cases, as in the study by Dong et al. (2014a), the multivariate random parameters zero-inflated negative binomial model identified more significant factors than did the multivariate zero-inflated negative binomial model.

4.5 Conclusions

In this study, we analyzed sideswipe (same direction), rear-end, and other crash types over 10 years (2003–2012) on 1,506 urban midblock segments in Lincoln and Omaha, Nebraska. Traffic operation and roadway geometry characteristics were investigated to identify significant influencing factors. Due to the concern of unobserved heterogeneity produced by correlations across crash types and segments, excess zeros, and over dispersion in crash data, the multivariate random parameters zero-inflated negative binomial model was used to simultaneously analyze these crashes. Compared to the multivariate Poisson log-normal, univariate random parameters zero-inflated Poisson, univariate random parameters zero-inflated negative binomial, multivariate zero-inflated Poisson, multivariate zero-inflated negative binomial and multivariate random parameters zero-inflated Poisson models, the multivariate random parameters zero-inflated negative binomial model provided a better fit in terms of both DIC and RMSE values for all three crash types. The model comparison showed that none of the four types of unobserved heterogeneities was negligible. The results proved the necessity and importance of using the multivariate random parameters zero-inflated negative binomial model to analyze multivariate panel crash data with excess zeros.

The multivariate random parameters zero-inflated negative binomial model revealed 9 out of 18 covariates as significantly influencing crash frequency for the studied midblock segments. The multivariate random parameters zero-inflated negative binomial model showed

that number of lanes, annual average daily traffic per lane, and segment length might have non-positive effects on crash frequencies for some segments. Thus, in future studies, care should be taken in using them as exposure variables. Segments with a speed limit of 45 mph tended to have fewer crashes than did those with lower speed limits, and the segments in Omaha tended to have fewer crashes than did those in Lincoln. It was also found that the presence of a shoulder, central business district, on-street parking, and one-way traffic, as well as lane width, did not have significant influences on crash frequencies. The multivariate random parameters zero-inflated negative binomial model also made it possible to explore influencing factors for individual segments. These findings are informative for transportation agencies as they seek to take correct and efficient measures to improve traffic safety. By contrast, the multivariate zero-inflated negative binomial model produced results consistent with intuition, but the results may be insufficient to provide actionable recommendations. The multivariate random parameters zero-inflated negative binomial model found fewer significant factors than did the multivariate zero-inflated negative binomial model, which was helpful for identifying key factors.

Several aspects of this study could be further improved in future studies. First, the multivariate random parameters zero-inflated negative binomial model were estimated using MCMC, which was time consuming and required a large capacity to store MCMC samples. With an increase in the amount and dimensions of data, MCMC would become even more cumbersome. Thus, Bayesian approximation methods, such as Integrated Nested Laplace Approximation and Variational Bayes, should be explored to improve computing efficiency. Second, the complexity of the multivariate random parameters zero-inflated negative binomial model makes the results less interpretable. For example, the rear-end crash frequencies marginally followed the zero-inflated negative binomial distribution in the multivariate zero-

inflated negative binomial model. However, it would be very difficult to calculate the marginal effects of annual average daily traffic per lane on rear-end crash frequencies in the multivariate zero-inflated negative binomial model, as this involved both the count part and the zero-inflation part. It would be even more difficult for the random parameters models. Sensitivity analysis and easy-to-understand visualization tools might be good solutions for showing the intricate correlations between covariates and response variables. Third, as an alternative to traditional zero-inflated models, the zero-state Markov switching count data model could distinguish zero-accident state and normal-count state in a straightforward manner and, as well, could capture the state change over time (Malyshkina and Mannering, 2010; Malyshkina et al., 2009); however, it has never been used in multivariate or random parameters scenarios. Future studies may explore the performance of the multivariate random parameters zero-state Markov switching count data model in analyzing similar crash data. Finally, crash frequency data are aggregated over time and space. Thus, they may have some spatial and temporal correlations (Boulieri et al., 2017; Liu et al., 2015; Liu and Sharma, 2018, 2017; Ma et al., 2017), and the effects of explanatory variables may also be instable over space and time (Mannering, 2018), which should be considered in future studies. In addition, the dataset did not include information about pavement conditions and access points, which have been proved to be very important for segment crash frequencies in many studies (Lee et al., 2011; Usman et al., 2010; Xiong et al., 2014; Zeng and Huang, 2014). Future studies should collect these data to produce more accurate results.

4.6 References

Aguero-Valverde, J., Jovanis, P.P., 2010. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transportation Research Record 2136, 82–91.

Alarifi, S.A., Abdel-Aty, M.A., Lee, J., Park, J., 2017. Crash modeling for intersections and segments along corridors: a Bayesian multilevel joint model with random parameters. Analytic Methods in Accident Research 16, 48–59.

Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Analytic Methods in Accident Research 11, 17–32.

Anastasopoulos, P.C., Shankar, V.N., Haddock, J.E., Mannering, F.L., 2012. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention 45, 110–119.

Anderson, J., Hernandez, S., 2017. Roadway classifications and the accident injury severities of heavy-vehicle drivers. Analytic Methods in Accident Research 15, 17–28.

Barua, S., El-Basyouny, K., Islam, M.T., 2014. A full Bayesian multivariate count data model of collision severity with spatial correlation. Analytic Methods in Accident Research 3–4, 28–43.

Barua, S., El-Basyouny, K., Islam, M.T., 2015. Effects of spatial correlation in random parameters collision count-data models. Analytic Methods in Accident Research 5–6, 28–42.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1–15.

Behnood, A., Mannering, F., 2017a. The effect of passengers on driver-injury severities in single-vehicle crashes: a random parameters heterogeneity-in-means approach. Analytic Methods in Accident Research 14, 41–53.

Behnood, A., Mannering, F., 2017b. Determinants of bicyclist injury severities in bicycle-vehicle crashes: a random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 16, 35–47.

Behnood, A., Mannering, F.L., 2016. An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. Analytic Methods in Accident Research 12, 1–17.

Bhat, C.R., Astroza, S., Lavieri, P.S., 2017. A new spatial and flexible multivariate random-coefficients model for the analysis of pedestrian injury counts by severity level. Analytic Methods in Accident Research 16, 1–22.

Bonneson, J., Mccoy, P., 1997. Effect of median treatment on urban arterial safety an accident prediction model. Transportation Research Record 1581, 27–36.

Boulieri, A., Liverani, S., Hoogh, K. de, Blangiardo, M., 2017. A space–time multivariate Bayesian model to analyse road traffic accidents by severity. Journal of the Royal Statistical Society Series A 180 (1), 119–139.

Chen, E., Tarko, A.P., 2014. Modeling safety of highway work zones with random parameters and random effects models. Analytic Methods in Accident Research 1, 86–95.

Chen, S., Saeed, T.U., Labi, S., 2017. Impact of road-surface condition on rural highway safety: a multivariate random parameters negative binomial approach. Analytic Methods in Accident Research 16, 75–89.

Coruh, E., Bilgic, A., Tortum, A., 2015. Accident analysis with the random parameters negative binomial panel count data model. Analytic Methods in Accident Research 7, 37–49.

Denwood, M.J., 2016. Runjags: an R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. Journal of Statistical Software 71 (9), 1–25.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014a. Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections. Accident Analysis and Prevention 70, 320–329.

Dong, C., Richards, S.H., Clarke, D.B., Zhou, X., Ma, Z., 2014b. Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression. Safety Science 70, 63–69.

Dumbaugh, E., 2006. Design of safe urban roadsides: an empirical analysis. Transportation Research Record 1961, 74–82.

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41 (4), 820–828.

Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. Transportation Research Record 1665, 133–140.

Federal Highway Administration, 2013. Highway functional classification concepts, criteria and procedures, FHWA-PL-12-026. Washington DC.

Ferreira, S., Couto, A., 2015. A probabilistic approach towards a crash risk assessment of urban segments. Transportation Research Part C 50, 97–105.

Fountas, G., Anastasopoulos, P.C., 2017. A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. Analytic Methods in Accident Research 15, 1–16.

Greibe, P., 2003. Accident prediction models for urban roads. Accident Analysis and Prevention 35 (2), 273–285.

Harwood, D.W., 1990. Effective utilization of street width on urban arterials, NCHRP Report 330. Transportation Research Board of National Research Council, Washington DC.

Hauer, E., Hakkert, A.S., 1988. Extent and some implications of incomplete accident reporting. Transportation Research Record 1185, 1–10.

Huang, H., Chin, H.C., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40 (1), 45–54.

Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. Analytic Methods in Accident Research 14, 10–21.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1997. Discrete Multivariate Distributions. John Wiley & Sons, Inc, New York, NY.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1–14.

Lee, C., Abdel-Aty, M., Park, J., Wang, J.H., 2015. Development of crash modification factors for changing lane width on roadway segments using generalized nonlinear models. Accident Analysis and Prevention 76, 83–91.

Lee, C., Xu, X., Nguyen, V., 2011. Analysis of midblock crashes in an urban divided arterial road. Journal of Transportation Safety & Security 4 (1), 1–18.

Li, C., Lu, J., Park, J., Kim, K., Brinkley, P.A., Peterson, J.P., 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics 41 (1), 29–38.

Liu, C., Gyawali, S., Sharma, A., Smaglik, E., 2015. A methodological approach for spatial and temporal analysis of red light running citations and crashes: a case-study in Lincoln, Nebraska. Transportation Research Board 94th Annual Meeting, Washington, D.C., USA.

Liu, C., Sharma, A., 2017. Exploring spatio-temporal effects in traffic crash trend analysis. Analytic Methods in Accident Research 16, 104–116.

Liu, C., Sharma, A., 2018. Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. Analytic Methods in Accident Research 17, 14–31.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-Gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37 (1), 35–46.

Ma, J., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models of injury count, by severity. Transportation Research Record 1950, 24–34.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964–975.

Ma, X., Chen, S., Chen, F., 2017. Multivariate space-time modeling of crash frequencies by injury severity levels. Analytic Methods in Accident Research 15, 29–40.

Malyshkina, N. V., Mannering, F., 2008. Effect of increases in speed limits on severities of injuries in accidents. Transportation Research Record 2083, 122–127.

Malyshkina, N. V., Mannering, F.L., 2010. Zero-state Markov switching count-data models: an empirical assessment. Accident Analysis and Prevention 42 (1), 122–130.

Malyshkina, N. V, Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accident Analysis and Prevention 41 (2), 217–226.

Mannering, F.L., 2018. Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17, 1–13.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1–16.

Manuel, A., El-Basyouny, K., Islam, M.T., 2014. Investigating the safety effects of road width on urban collector roadways. Safety Science 62, 305–311.

Miaou, S.-P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping a space-time modeling approach. Journal of Transportation and Statistics 6 (1), 33–57.

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

MRC Biostatistics Unit, 2004. DIC: Deviance Information Criteria. https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-dic/ (accessed 8.23.17).

Naik, B., Tung, L.W., Zhao, S., Khattak, A.J., 2016. Weather impacts on single-vehicle truck crash injury severity. Journal of Safety Research 58, 57–65.

National Center for Statistics and Analysis, 2017. Traffic safety facts 2015, DOT HS 812 384. U.S. Department of Transportation, National Highwway Traffic Safety Administration, Washington D.C.

Osama, A., Sayed, T., 2017. Investigating the effect of spatial and mode correlations on active transportation safety modeling. Analytic Methods in Accident Research 16, 60–74.

Pande, A., Abdel-Aty, M., Das, A., 2010. A classification tree based modeling approach for segment related crashes on multilane highways. Journal of Safety Research 41 (5), 391–397.

Park, J., Abdel-Aty, M., 2016. Evaluation of safety effectiveness of multiple cross sectional features on urban arterials. Accident Analysis and Prevention 92, 245–255.

Plummer, M., 2002. Discussion of the paper by Spiegelhalter et al. Journal of the Royal Statistical Society Series B 64, 620.

Plummer, M., 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Technische Universitat Wien, Vienna, Austria.

Potts, I., Harwood, D., Richard, K.R., 2007. Relationship of lane width to safety on urban and suburban arterials. Transportation Research Record 2023, 63–82.

R Core Team, 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Renski, H., Khattak, A.J., Council, F.M., 1999. Effect of speed limit increases on crash injury severity: analysis of single-vehicle crashes on North Carolina interstate highways. Transportation Research Record 1665, 100–108.

Rista, E., Goswamy, A., Wang, B., Barrette, T., Hamzeie, R., Russo, B., Bou-Saab, G., Savolainen, P.T., 2017. Examining the safety impacts of narrow lane widths on urban/suburban arterials: estimation of a panel data random parameters negative binomial model. Journal of Transportation Safety & Security (In press), 1–16.

Russo, B.J., Savolainen, P.T., Schneider IV, W.H., Anastasopoulos, P.C., 2014. Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. Analytic Methods in Accident Research 2, 21–29.

Sarwar, M.T., Anastasopoulos, P.C., Golshani, N., Hulme, K.F., 2017. Grouped random parameters bivariate probit analysis of perceived and observed aggressive driving behavior: a driving simulation study. Analytic Methods in Accident Research 13, 52–64.

Sawalha, Z., Sayed, T., 2001. Evaluating safety of urban arterial roadways. Journal of Transportation Engineering 127 (2), 151–158.

Seraneeprakarn, P., Huang, S., Shankar, V., Mannering, F., Venkataraman, N., Milton, J., 2017. Occupant injury severities in hybrid-vehicle involved crashes: a random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 15, 41–55.

Serhiyenko, V., Mamun, S.A., Ivan, J.N., Ravishanker, N., 2016. Fast Bayesian inference for modeling multivariate crash counts. Analytic Methods in Accident Research 9, 44–53.

Sharma, A., Li, W., Zhao, M., Rilett, L.R., 2015. Safety and operational analysis of lane widths in mid-block segments and intersection approaches in the urban environment in Nebraska, Report No. SPR-P1(13) M327. Nebraska Department of Roads, Lincoln, NE.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B 64 (4), 583–639.

Usman, T., Fu, L., Miranda-Moreno, L.F., 2010. Quantifying safety benefit of winter road maintenance: accident frequency modeling. Accident Analysis and Prevention 42 (6), 1878–1887.

Venkataraman, N., Shankar, V., Ulfarsson, G.F., Deptuch, D., 2014. A heterogeneity-in-means count model for evaluating the effects of interchange type on heterogeneous influences of interstate geometrics on crash frequencies. Analytic Methods in Accident Research 2, 12–20.

Wang, K., Zhao, S., Jackson, E., 2018. Multivariate Poisson lognormal modeling of weather related crashes on freeways. Transportation Research Record (Accepted).

Wu, Z., Sharma, A., Mannering, F.L., Wang, S., 2013. Safety impacts of signal-warning flashers and speed control at high-speed signalized intersections. Accident Analysis and Prevention 54, 90–98.

Xiong, Y., Tobias, J.L., Mannering, F.L., 2014. The analysis of vehicle crash injury-severity data: a Markov switching approach with road-segment heterogeneity. Transportation Research Part B 67, 109–128.

Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis and Prevention 40 (4), 1320–1329.

Yannis, G., Papadimitriou, E., Chaziris, A., Broughton, J., 2014. Modeling road accident injury under-reporting in Europe. European Transport Research Review 6 (4), 425–438.

Zeng, Q., Huang, H., 2014. Bayesian spatial joint modeling of traffic crashes on an urban road network. Accident Analysis and Prevention 67, 105–112.

Zhan, X., Abdul Aziz, H.M., Ukkusuri, S. V., 2015. An efficient parallel sampling technique for multivariate Poisson-lognormal model: analysis with two crash count datasets. Analytic Methods in Accident Research 8, 45–60.

Zhang, Y., Xie, Y., Li, L., 2012. Crash frequency analysis of different types of urban roadway segments using generalized additive model. Journal of Safety Research 43 (2), 107–114.

Zhao, M., Liu, C., Li, W., Sharma, A., 2017. Multivariate Poisson-lognormal model for analysis of crashes on urban signalized intersections approach. Journal of Transportation Safety & Security (In press), 1–15.

Zhao, S., Khattak, A.J., 2015. Motor vehicle drivers' injuries in train–motor vehicle crashes. Accident Analysis and Prevention 74, 162–168.

Zhao, S., Khattak, A.J., 2017. Injury severity in crashes reported in proximity of rail crossings: the role of driver inattention. Journal of Transportation Safety & Security (In press), 1–18.

CHAPTER 5. GENERAL CONCLUSIONS

This dissertation consists of three studies that focus on crash frequency analysis at the macro and micro levels respectively. The first study shows the necessity and importance of including spatial and temporal effects in crash frequency analysis, the second study extends the spatio-temporal analysis into multivariate cases, and the third study explores the heterogeneous effects of various factors on crash frequency by crash types. These three studies show how to use appropriate statistical models to deal with the common issues of crash frequency data, i.e. over dispersion, zero inflation, spatial correlations, temporal correlations, crash-between correlations, and unobserved heterogeneity.

While this study has made important contributions to the literature, future research may continue in many aspects of both the methodology and the studied objects. Firstly, as is shown in Table 3-3, the spatial correlations of crashes might evolve over time. Similarly, it is expected that the temporal correlations of crashes might evolve over space, which is partly proved by the superiority of the linear temporal component in Chapter 2. Thus, it implies that crashes might have dynamic spatio-temporal correlations, while this study assumes these correlations are static. Future studies may further explore the dynamic spatio-temporal analysis of crashes.

Secondly, the existing studies of spatial analysis of crashes mainly focus on utilizing the areal spatial statistics models as crash frequency data are usually collected over jurisdictions. However, the aggregation of crash data over space would inevitably lose important location information of each crash, which is critical for transportation agencies to identify the clustering trends of crashes in each area, as there is no reason to believe crashes would occur equally in each area. When individual crash geographical information is available, the spatial point process analysis is a good choice for crash analysis. Common spatial point data, such as crime and wide

file, might occur over the whole studied areas, i.e. the polygon spatial point data, while crashes actually only occur on the roadway network, i.e. the line spatial point data. This difference brings the new challenge that most existing spatial point pattern analysis, which is developed for polygon spatial point data, might not work well for crashes. Thus, researchers may focus on developing new spatial point process models for the line spatial point data like crashes.

Thirdly, besides traffic safety, traffic operation is another cornerstone of transportation research. Most studies often analyze them separately, however, it might produce more beneficial findings if they are analyzed at the same time. For example, if the travel speed data on segments are also available in Chapter 4, speed and crash may be analyzed simultaneously. Thus, we can get a full picture of the effects of roadway geometric characteristics on the transportation system.