INTEGRATING THE MACROSCOPIC AND MICROSCOPIC TRAFFIC SAFETY ANALYSIS USING HIERARCHICAL MODELS

by

QING CAI

B.S., Tongji University, China, 2011
M.S., Tongji University, China, 2014

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Civil, Environmental and Construction Engineering
in the College of Engineering and Computer Science
at University of Central Florida
Orlando, Florida

Summer Term
2017

Major Professor: Mohamed Abdel-Aty

ABSTRACT

Crash frequency analysis is a crucial tool to investigate traffic safety problems. With the objective of revealing hazardous factors which would affect crash occurrence, crash frequency analysis has been undertaken at the macroscopic and microscopic levels. At the macroscopic level, crashes from a spatial aggregation (such as traffic analysis zone or county) are considered to quantify the impacts of socioeconomic and demographic characteristics, transportation demand and network attributes so as to provide countermeasures from a planning perspective. On the other hand, the microscopic crashes on a segment or intersection are analyzed to identify the influence of geometric design, lighting and traffic flow characteristics with the objective of offering engineering solutions (such as installing sidewalk and bike lane, adding lighting). Although numerous traffic safety studies have been conducted, still there are critical limitations at both levels. In this dissertation, several methodologies have been proposed to alleviate several limitations in the macro- and micro-level safety research. Then, an innovative method has been suggested to analyze crashes at the two levels, simultaneously.

At the macro-level, the viability of dual-state models (i.e., zero-inflated and hurdle models) were explored for traffic analysis zone based pedestrian and bicycle crash analysis. Additionally, spatial spillover effects were explored in the models by employing exogenous variables from neighboring zones. Both conventional single-state model (i.e., negative binomial) and dual-state models such as zero-inflated negative binomial and hurdle negative binomial models with and without spatial effects were developed. The model comparison results for pedestrian and bicycle crashes revealed that the models that considered observed spatial effects perform better than the models that did not consider the observed spatial effects. Across the models with spatial spillover

effects, the dual-state models especially zero-inflated negative binomial model offered better performance compared to single-state models. Moreover, the model results clearly highlighted the importance of various traffic, roadway, and sociodemographic characteristics of the TAZ as well as neighboring TAZs on pedestrian and bicycle crash frequency.

Then, the modifiable areal unit problem for macro-level crash analysis was discussed. Macro-level traffic safety analysis has been undertaken at different spatial configurations. However, clear guidelines for the appropriate zonal system selection for safety analysis are unavailable. In this study, a comparative analysis was conducted to determine the optimal zonal system for macroscopic crash modeling considering census tracts (CTs), traffic analysis zones (TAZs), and a newly developed traffic-related zone system labeled traffic analysis districts (TADs). Poisson lognormal models for three crash types (i.e., total, severe, and non-motorized mode crashes) were developed based on the three zonal systems without and with consideration of spatial autocorrelation. The study proposed a method to compare the modeling performance of the three types of geographic units at different spatial configuration through a grid based framework. Specifically, the study region was partitioned to grids of various sizes and the model prediction accuracy of the various macro models was considered within these grids of various sizes. These model comparison results for all crash types indicated that the models based on TADs consistently offer a better performance compared to the others. Besides, the models considering spatial autocorrelation outperformed the ones that do not consider it. Finally, based on the modeling results, it is recommended to adopt TADs for transportation safety planning.

After determining the optimal traffic safety analysis zonal system, further analysis was conducted for non-motorist crashes (pedestrian and bicycle crashes). This study contributed to

the literature on pedestrian and bicyclist safety by building on the conventional count regression models to explore exogenous factors affecting pedestrian and bicyclist crashes at the macroscopic level. In the traditional count models, effects of exogenous factors on non-motorist crashes were investigated directly. However, the vulnerable road users' crashes are collisions between vehicles and non-motorists. Thus, the exogenous factors can affect the non-motorist crashes through the non-motorists and vehicle drivers. To accommodate for the potentially different impact of exogenous factors we converted the non-motorist crash counts as the product of total crash counts and proportion of non-motorist crashes and formulated a joint model of the negative binomial (NB) model and the logit model to deal with the two parts, respectively. The formulated joint model was estimated using non-motorist crash data based on the Traffic Analysis Districts (TADs) in Florida. Meanwhile, the traditional NB model was also estimated and compared with the joint model. The results indicated that the joint model provides better data fit and could identify more significant variables. Subsequently, a novel joint screening method was suggested based on the proposed model to identify hot zones for non-motorist crashes. The hot zones of non-motorist crashes were identified and divided into three types: hot zones with more dangerous driving environment only, hot zones with more hazardous walking and cycling conditions only, and hot zones with both.

At the microscopic level, crash modeling analysis was conducted for road facilities. This study, first, explored the potential macro-level effects which are always excluded or omitted in the previous studies. A Bayesian hierarchical model was proposed to analyze crashes on segments and intersection incorporating the macro-level data, which included both explanatory variables and total crashes of all segments and intersections. Besides, a joint modeling structure was adopted to consider the potentially spatial autocorrelation between segments and their connected

intersections. The proposed model was compared with three other models: a model considering micro-level factors only, one hierarchical model considering macro-level effects with random terms only, and one hierarchical model considering macro-level effects with explanatory variables. The results indicated that models considering macro-level effects outperformed the model having micro-level factors only, which supports the idea to consider macro-level effects for micro-level crash analysis. Besides, the micro-level models were even further enhanced by the proposed model. Finally, significant spatial correlation could be found between segments and their adjacent intersections, supporting the employment of the joint modeling structure to analyze crashes at various types of road facilities.

In addition to the separated analysis at either the macro- or micro-level, an integrated approach has been proposed to examine traffic safety problems at the two levels, simultaneously. If conducted in the same study area, the macro- and micro-level crash analyses should investigate the same crashes but aggregating the crashes at different levels. Hence, the crash counts at the two levels should be correlated and integrating macro- and micro-level crash frequency analyses in one modeling structure might have the ability to better explain crash occurrence by realizing the effects of both macro- and micro-level factors. This study proposed a Bayesian integrated spatial crash frequency model, which linked the crash counts of macro- and micro-levels based on the spatial interaction. In addition, the proposed model considered the spatial autocorrelation of different types of road facilities (i.e., segments and intersections) at the micro-level with a joint modeling structure. Two independent non-integrated models for macro- and micro-levels were also estimated separately and compared with the integrated model. The results indicated that the integrated model can provide better model performance for estimating macro- and micro-level crash counts, which validates the concept of integrating the models for the two levels.

Also, the integrated model provides more valuable insights about the crash occurrence at the two levels by revealing both macro- and micro-level factors. Subsequently, a novel hotspot identification method was suggested, which enables us to detect hotspots for both macro- and micro-levels with comprehensive information from the two levels. It is expected that the proposed integrated model and hotspot identification method can help practitioners implement more reasonable transportation safety plans and more effective engineering treatments to proactively enhance safety.

ACKNOWLEDGMENT

I would like to express my appreciation and gratitude to my advisor Dr. Mohamed Abdel-Aty. His support and brilliant guidance contributed greatly to the research I have accomplished. Also, his spirits of seriousness, enthusiasm of research, and confidence let me understand how to become an outstanding transportation researcher.

I would like to thank the Dissertation Committee members, Dr. Naveen Eluru, Dr. Samuil Hasan, Dr. Xin Yan, and Dr. Jaeyoung Lee, for providing valuable suggestions for my research work.

Appreciation is due all my colleagues for discussing and sharing ideas about the research. It is really a lot of fun to work with them.

Last but not the least, many thanks to my families. Especially, I would like to thank my wife. It is so nice that I can always have her support in my whole PhD study. Without her, I couldn't accomplish the work I have done.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

AIC	Akaike Information Criterion
BG	Block Group
BIC	Bayesian Information Criterion
CB	Census Block
CDC	Centers for Disease Control and Prevention
CT	Census Tract
CTTP	Census Transportation Planning Product
DIC	Deviance Information Criterion
DOT	Department of Transportation
EB	Empirical Bayes
FARS	Fatality Analysis Reporting System
FDOT	Florida Department of Transportation
FHWA	Federal Highway Administration
GIS	Geographic Information System
LR	Log-Likelihood

LRTP	Long Range Transportation Plan
MPO	Metropolitan Planning Organization
MAE	Mean Absolute Error
NB	Negative Binomial
NHTSA	National Highway Traffic Safety Administration
PDO	Property Damage Only
PLN	Poisson-lognormal Model
RMSE	Root Mean Square Error
SPF	Safety Performance Function
TAD	Traffic Analysis District
TAZ	Traffic Analysis Zone
TSP	Transportation Safety Planning
VMT	Vehicle-Miles-Traveled
ZCTA	ZIP Code Tabulation Area

CHAPTER 1: INTRODUCTION

1.1 Overview

Traffic safety is considered one of the most critical issues of the transportation system. The consistent efforts of government officials and transportation engineers have ensured that fatalities from traffic collisions have gradually declined in the recent decades in the United States. However, traffic fatalities rose in 2012 and 2015 highlighting the challenge faced by the safety community. Particularly, the nation lost 35,092 people in traffic crashes during 2015, a 7.2-percent increase from 32,744 in 2014. The increase is the largest percentage increase in nearly 50 years (NHTSA., 2016). Thus, it is necessary to devote many efforts to reduce traffic crashes and enhance road safety. The continued efforts of traffic safety analysis are required to identify hazardous factors affecting crash occurrence.

One of the most widely used approaches to investigate traffic safety is crash frequency modeling, which can quantify exogenous factors contributing to the number of traffic crashes. At the macroscopic level, crashes from a spatial aggregation (such as traffic analysis zone or county) are considered to quantify the impacts of socioeconomic and demographic characteristics, transportation demand and network attributes so as to provide countermeasures from a planning perspective. On the other hand, the microscopic crashes on a segment or intersection are analyzed to identify the influence of geometric design, lighting and traffic flow characteristics with the objective of offering engineering solutions (such as installing sidewalk and bike lane, adding lighting).

Many macroscopic and microscopic safety researches have been conducted to facilitate the implementation of traffic safety plans or roadway engineering solutions. The macro-level crash safety researches have been conducted based on different zonal systems such as traffic analysis zone, census tract, county, and state. Since these zonal systems were developed for different usages and criteria, the statistical inference and interpretation derived from the zones would be also various, referring as the modifiable areal unit problem. Hence, it is necessary to suggest clear guidelines for the appropriate zonal system selection for macro-level safety analysis.

Walking and bicycling are two active forms of transportation, which can offer an environmentally friendly and physically active alternative for short distance trips. A strong impediment to universal adoption of active forms of transportation, particularly in North America, is the inherent safety risk for active modes of transportation. Towards developing counter measures to reduce safety risks, it is essential to study the influence of exogenous factors on pedestrian and bicycle crashes at the macro-level.

At the micro-level, the effects of traffic characteristics and road features on crashes of segments and intersections have been identified. Most of crash frequency studies at the micro-level have omitted the effects of macro-level factors. It would be reasonable to claim that the road facilities which are located in the same zone should share certain zonal factors, which may affect crash occurrence through driving behaviors and transportation modes.

Previous studies have explored traffic safety at either macroscopic or microscopic level, i.e., no study has investigated the two levels. If traffic safety research is conducted for the same study area, macro- and micro-level crash analyses would investigate the same crashes but by different

aggregation levels. Hence, we can assume that the crash counts at the two levels are correlated. Particularly, the total number of crashes in each zone (macro-level) is supposed to be the same as the total number of crashes from all road entities including segments and intersections (micro-level) located in the zone of interest. Hence, it would be beneficial if the integrated traffic safety modeling analysis can be conducted for the two levels. This approach can simultaneously exam the traffic safety problems for different zones and road facilities by employing the data aggregated at the two levels. Such integrated approach is supposed to identify hazardous factors at both macro- and micro-levels. Subsequently, the incorporated countermeasures can be proposed to extensively reduce crashes. It is expected it can be easier to achieve the goal of a traffic safety plan with effective safety improvement of roadway infrastructure. Meanwhile, the engineering countermeasures can be more appropriate with the guidance of traffic safety plans.

Therefore, the objective of this study is to explore possible limitations of individual macroscopic or microscopic crash analysis, and subsequently develop integrated hierarchical models to investigate traffic safety problems at the two levels, simultaneously. Based on the integrated models, a guideline with comprehensive perspectives will be suggested to enhance traffic safety at both levels.

1.2 Research Objectives

The dissertation focuses on suggesting appropriate methodologies to explore hazardous factors affecting crash occurrence at either macroscopic or microscopic level and develop a novel methodology to integrate traffic safety analysis at the two levels. The specific objective will be achieved by the following procedures:

1. Conducting preliminary pedestrian and bicycle safety studies at the macroscopic level;
2. Determining the optimal zonal system for macroscopic traffic safety analysis;
3. Suggesting appropriate methods to analyze crashes based on the determined optimal zonal system;
4. Exploring the potential macro-level effects in micro-level crash analysis, and;
5. Integrating macroscopic and microscopic traffic safety analysis using hierarchical models.

The first objective has been achieved in Chapter 3 by the following tasks:

a) Discussing the excess-zero problems for the pedestrian and bicycle crashes based on traffic analysis zones;
b) Exploring the viability of dual-state models for pedestrian and bicycle crash analysis.

The second objective has been achieved in Chapter 4 by the following tasks:

c) Selecting different zonal systems including traffic analysis zones, census tracts, and traffic analysis districts which are transportation-related geographic units or have been widely used for macro-level crash analysis;
d) Developing multiple crash frequency models based on different zonal systems for different crash types;
e) Suggesting a grid-based method to compare modeling performance based on different zonal systems;

f) Adopting appropriate goodness-of-fit measures to compare performance of models based on different zonal systems and suggesting the most appropriate zonal system for macro-level crash analysis.

The following tasks have been implemented in Chapter 5 to achieve the third objective:

g) Analyzing pedestrian and bicycle crashes based on the optimal zonal systems suggested in Chapter 4;
h) Developing a joint model for pedestrian and bicycle crashes to recognize effects of explanatory variables on vehicle drivers and non-motorists;
i) Suggesting a joint screening method to identify hot zones of non-motorist crashes with more details.

The fourth objective has been achieved in Chapter 6 by the following tasks:

j) Developing a Bayesian hierarchical model to investigate macro-level effects for micro-level crash analysis;
k) Considering the potentially spatial correlation between segments and intersection by adopting a joint modeling structure;
l) Estimating three other models at the micro-level and comparing them with the proposed model.

The last objective has been achieved in Chapter 7 by the following tasks:

m) Aggregating crashes from the same area at the macro- and micro-levels and examining the potential correlation between macro- and micro-level crashes;
n) Suggesting a hierarchical integrated model which could simultaneously analyze crashes at the macro and micro-levels based on the spatial interactions;
o) Comparing the proposed integrated model with the non-integrated models at both levels to validate the concept of integrating the models for the levels;
p) Proposing a novel screening method which could detect hotspots for both macro- and micro-levels with comprehensive information from the two levels.

1.3 Dissertation Organization

The organization of the dissertation is as follows: Chapter 2, following this chapter, summarizes literature review about previous macroscopic and microscopic traffic safety analyses, current issues of the safety researches, and related studies. Additionally, the statistic methodology for the safety analysis has been also discussed. Chapter 3 addresses the excess-zero issue for pedestrian and bicyclist crashes based on traffic analysis zone (TAZ) by adopting two-stage models. Chapter 4 compares different zone systems for macroscopic traffic crash modeling and recommends the optimal zone system. Chapter 5 develops a joint model for non-motorist crashes to identify different impacts of exogenous variables on vehicle drivers and non-motorists. Chapter 6 suggests a Bayesian hierarchical model to investigate the macro-level effects on micro-level crashes. Chapter 7 formulates an integrated model to analyze traffic crashes at the macro- and micro-levels, simultaneously. Based on the spatial interaction between zones and

road entities, the expected crash counts at the macro- and micro-levels are linked by an adjustment factor. Additionally, the spatial autocorrelations at the macro-level and micro-level are also considered. Finally, Chapter 8 summarizes the overall dissertation and proposes a set of recommendations and follow-up studies.

CHAPTER 2: LITERATURE REVIEW

2.1 General

The review of literature is divided into three main sections: First, the previous traffic safety studies (crash frequency models) at the macroscopic level have been summarized. The different zonal systems used for the crash analysis and factors contributing to crash frequency have been discussed. Second, the past microscopic traffic safety researches have been discussed in detail. Particularly, the researches based on segments and intersections have been summarized. Finally, a review of the statistical methodology for the crash frequency analysis has been presented.

2.2 Macroscopic Safety Study

In the recent decade, there has been growing recognition to incorporate roadway safety in the long-term transportation planning process. Several planning acts have emphasized the importance of macroscopic crash analysis. Initially, the Transportation Equity Act for the 21st Century (Houston, 1998) suggested to consider safety in the transportation planning process. Later, Washington et al. (2006) discussed how to incorporate safety into transportation planning at different levels. Currently, the Moving Ahead for Progress in the 21st Century Act (MAP-21 Act) (US Congress, 2012) and Fixing America's Surface Transportation Act (FAST Act) (U.S. DOT, 2015) require the incorporation of transportation safety in the long-term transportation planning process. Generally, macroscopic safety studies are to quantify the statistical relation between characteristics and crashes at zonal levels. Also, various zonal systems have been explored for the macroscopic crash analysis. Thus, the following sub-chapters will briefly discuss about the different zonal systems and zonal characteristics in macroscopic studies.

2.2.1 Zonal Systems for Studies

Most of previous macroscopic safety studies were conducted based on single type of zonal system. The zonal systems include: block groups (Levine *et al.*, 1995), census tracts (LaScala *et al.*, 2000; Loukaitou-Sideris *et al.*, 2007; Wier *et al.*, 2009; Wang and Kockelman, 2013), ZIP code areas (Lee *et al.*, 2013; Lee *et al.*, 2015a), traffic analysis zones or TAZs (Hadayeghi *et al.*, 2003; Ladrón de Guevara *et al.*, 2004; Hadayeghi *et al.*, 2010; Abdel-Aty *et al.*, 2011b; Lee *et al.*, 2013; Dong *et al.*, 2015; Wang et al., 2016; Cai *et al.*, 2016; Wang and Huang, 2016; Yasmin and Eluru, 2016), counties (Aguero-Valverde and Jovanis, 2006; Huang *et al.*, 2010), states (Noland, 2003), and Grids (Kim *et al.*, 2006b). Most of these zonal systems were developed for different specific usages.

(1) Block Groups and Census Tracts

The block groups (BGs) and census tracts (CTs) are census based zonal systems developed for the collection and tabulation of decennial census data (CensusBureau, 1992). Both the BGs and CTs are developed based on the census blocks (CBs), which are the smallest geographic units used by United States Census Bureau. The census blocks (CBs) are very small, especially in the urban area. Besides, the detailed information is not available based on (CBs). Thus, CBs are not usually used for the macro-level safety studies.

A BG is developed by combing CBs and each BG contains 39 CBs in average. Population in a BG is between 600 and 3,000 people. A CT is a combination of BGs and relatively permanent subdivisions of a county or equivalent entity to present statistical data such as poverty rates, income levels, etc. On average, a CT has about 4,000 inhabitants. CTs are designed to be

relatively homogeneous units with respect to population characteristics, economic status, and living conditions. Several macroscopic studies were conducted based on BGs and CTs.

(2) ZIP Code

ZIP codes are a system of postal codes used by the United States Postal Service (USPS). Basically, the ZIP codes are developed for mail delivery routing. However, besides tracking of mail, the ZIP codes are also used for gathering geographical statistics. The U.S. Census Bureau calculates approximate boundaries of ZIP codes areas, which is called ZIP Code Tabulation Areas (ZCTAs). Statistical data are provided based on ZCTAs. In the crash data, the ZIP codes are included as the residence information. Thus, several studies which focus on road users have been conducted based on ZIP codes.

(3) Traffic Analysis Zones (TAZs)

Traffic Analysis Zones (TAZs) are geographic entities delineated by state or local transportation officials to tabulate traffic-related data such as journey-to-work and place-of-work statistics (FHWA, 2014). TAZs are defined by grouping together census blocks, block groups, or census tracts. A TAZ usually covers a contiguous area with a 600-minimum population and the land use within each TAZ is relatively homogeneous (Abdel-Aty *et al.*, 2013). Previously, since TAZs are the only traffic related zonal system, they have been most widely used in the macroscopic safety literature. However, considering that TAZs are not delineated for traffic crash analysis, there are possible limitations of TAZs for macroscopic safety analysis. Thus, it should be necessary to evaluate the viability of TAZs for safety study.

Besides TAZs, a new and higher-level zonal system, Traffic Analysis Districts (TADs), were developed for traffic analysis (FHWA, 2011a). TADs are built by aggregating TAZs, block groups or census tracts. In almost every case, the TADs are delineated to adhere to a 20,000 minimum population criteria (FHWA, 2014) and more likely to have mixed land use. No research has been conducted based on TADs for safety analysis. In this study, the viability of TADs as a zonal system for macro-level crash modeling will be explored and the comparison between TAZs and TADs will be also conducted.

(4) Counties and States

Compared with zonal systems discussed above, counties and states are higher-level geographic units for macroscopic analysis. Both of them are polity related zonal systems. A state is an organized community living under a single political structure and government, sovereign or constituent while a county is an administrative division of the state in which its boundary is drawn. Several researches focusing on comparison between high-level zonal systems have been conducted based on counties or states.

(5) Grids

Since the study zonal systems are developed for specific purposes, their number of units, aggregation levels and zoning configuration can vary substantially across different zonal systems. Regarding this, Kim *et al* (2006b) developed a uniform 0.1 square mile grid structure to explore the impact of socio-demographic characteristics such as land use, population size, and employment by sector on crashes. Compared with other existing geographic units, the grid

structure is uniformly sized and shaped which can eliminate the artifact effects. However, considering the availability and use of the various zonal systems for other transportation purposes creating a uniform grid structure would not be feasible from the perspective of state and regional agencies.

Recently, besides single type of zonal system, several research studies have been conducted to compare different geographic units. Abdel-Aty *et al.* (2013) conducted modeling analysis for three types of crashes (total, severe, and pedestrian crashes) with three different types of geographic units (block groups, TAZs, and census tracts). Inconsistent significant variables were observed for the same dependent variables, validating the existence of zonal variation. However, no comparison of modeling performance was conducted in this research. Lee *et al.* (2014) aggregated TAZs into traffic safety analysis zones (TSAZs) based on crash counts. Four different goodness-of-fit measures (i.e., mean absolute deviation, root mean squared errors, sum of absolute deviation, and percent mean absolute deviation) were employed to compare crash model performance based on TSAZs and TAZs. The results indicated that the model based on the new zone system can provide better performance. Instead of determining the best zone system, Xu *et al.* (2014) created different zoning schemes by aggregating TAZs with a dynamical method. Models for total/severe crashes were estimated to explore variations across zonal schemes with different aggregation levels. Meanwhile, deviance information criterion, mean absolute deviation, and mean squared predictive error were calculated to compare different models. However, the employed measures for the comparison can be largely influenced by the number of observations and the observed values. Thus, the comparison results might be limited in the two studies (Lee *et al.*, 2014; Xu *et al.*, 2014) since the measures were calculated based on zonal systems with different number of zones.

2.2.2 Characteristics for Macroscopic Crash Analysis

Various explanatory variables aggregated at zonal level have been investigated by macroscopic crash analysis. Generally, the variables can be grouped into five categories: traffic, road network, socioeconomic characteristics, commuting characteristics, and land use. The following parts will present discussion about different explanatory variables explored in macro-level safety studies.

(1) Traffic

Usually, two variables, Vehicle Miles Travelled (VMT) and proportion of heavy vehicle mileage are investigated in macro-level study. The VMT is employed as exposure of traffic and always found to have positive effect on crash frequency (Lee *et al.*, 2014; Dong *et al.*, 2015). The increased proportion of heavy vehicle mileage reflects rural area where the exposure of traffic is comparatively low. Thus, the increased proportion will result in reducing crash frequency (Cai *et al.*, 2016).

(2) Road Network

Several road networks related information are considered in macroscopic studies: roadway density, road functional classification, speed limit, number of lanes of road, lane width, pavement condition, intersection types, roundabout, sidewalk, and bike lane. Some variables (such as roadway density and proportion of different road functional classification) are found to have different impacts on different types of crashes. For example, some studies found that the roadway density has a positive relation with total crashes (Noland and Oh, 2004) and slight

injury crashes while has a negative association with fatalities (Noland and Quddus, 2004). With the increasing proportion of freeway, the total crash will decrease (Noland and Quddus, 2004) while the fatalities can increase (Li *et al.*, 2013). Meanwhile, some consistent impacts can be observed for some variables. For example, it is revealed that proportion of roadway with poor pavement condition can increase crashes (Lee *et al.*, 2015). Besides, the zones with numerous intersections would have more crashes (Amoros *et al.*, 2003; Huang *et al.*, 2010). The variables length of sidewalk and length of bike lane are usually adopted for pedestrian and bicycle crash analysis. Both of the two variables are found to have positive effects on pedestrian and bicycle crashes (Cho *et al.*, 2009; Cai *et al.*, 2016).

(3) Socioeconomic Characteristics

In terms of sociodemographic characteristics, five types of variables, population, age, gender, and land use, are usually employed for crash frequency analysis. The population density can reflect traffic exposure and is found to have positive relation with crashes (Ladron de Guevara *et al.*, 2004; Permpoonwiwat and Kotrajaras, 2012). Also, Male and young drivers are more likely to increase crashes (MacNab, 2004; Li *et al.*, 2013). On the other hand, the order people are tend to reduce total crash while they can cause more severe injury crashes (Noland, 2003). Furthermore, the impacts of land use have been investigated for different crash types (Noland and Quddus, 2004; Wier *et al.*, 2009). Especially, the land use in zones has been associated with pedestrian and bicycle crashes-with increases predicted by increasing proportion of land use for commercial, mixed use, park, retail, or community uses (Geyer *et al.*, 2006; Kim *et al.*, 2006b; Wedagama *et al.*, 2006; Loukaitou-Sideris *et al.*, 2007; Wier *et al.*, 2009).

The variables, employment and household income, are usually used as socioeconomic characteristics for the analysis. The impacts of different types of employment were explored in previous studies (Hadayeghi *et al.*, 2010; Pulugurtha *et al.*, 2013). The positive effect of employment density on crashes was reflected in the studies (Loukaitou-Sideris *et al.*, 2007; Wier *et al.*, 2009). Meanwhile, the negative impact of median household income was always observed (Siddiqui *et al.*, 2012; Xu and Huang, 2015; Huang *et al.*, 2016).

(4) Commuting Characteristics

As for the commuting characteristics, proportion of commuters by different transportation modes and commute time are explored in the previous studies. The proportions of commuters by public transportation, walking, and bike are always employed for pedestrian and bicycle crash analysis and are found to have positive effects on crashes (Graham and Glaister, 2003; Wier *et al.*, 2009; Cai *et al.*, 2016). Longer commute time is likely to increase crash frequency since it increase the exposure (Abdel-Aty *et al.*, 2013).

(5) Land Use

The impacts of different types of land use and proportion of urban area or distance to the nearest urban location are also studied for macro-level analysis. Some studies (Kim and Yamashita, 2002; Wier *et al.*, 2009; Ukkusuri *et al.*, 2012) find that areas with commercial and residential land use have a higher frequency of crashes. Besides, urban location is found to be positively associated with the crashes, especially pedestrian and bicycle crashes (Siddiqui *et al.*, 2012; Li *et al.*, 2013).

2.3 Microscopic Safety Study

As for the microscopic safety studies, wide arrays of researches have been conducted at different types of segments and intersections. The impacts of different characteristics such as traffic flow, geometry, and signal phase on crashes have been investigated in the previous studies. In the subchapter, the safety study at micro-level will be briefly presented.

2.3.1 Road Facilities for Study

(1) Segment

Abdel-Aty and Radwan (2000) divided a 227 km long two-lane road into 566 segments based on homogeneous characteristics in terms of traffic flow and geometry and they found that the variables degree of horizontal curvature, shoulder and median widths, rural/urban classification, lane width and number of lanes are strongly related to the accident occurrence. Also, the research concluded that to obtain a reliable accident prediction model, sections should be 0.8 km or longer.

Mayora and Rubio (2003) developed models for two different types of segments of two-lane roads, 1-km fixed length segments and network links joining two consecutive nodes with variable lengths ranging from 3 km to 25 km. The significant correlations between crashes and access density, average sight distance, average speed limit and proportion of no passing zones were revealed in the study.

Hauer *et al.* (2004) analyzed crash frequency on undivided four-lane urban roads. The effects of various characteristics, AADT, percentage of trucks, degree and length of horizontal curves,

grade of tangents and length vertical curves, lane width, shoulder width and type, roadside hazard rating, speed limit, access points, etc, were evaluated. The finding showed that significant variables were: AADT, the number of driveways, and speed limit.

Zhang and Ivan (2005) evaluated the effects of roadway geometric features on occurrence of head-on crashes on two-lane rural roads. Variables found to influence of head-on crashes significantly were speed limit, sum of absolute change rate of horizontal curvature, maximum degree of horizontal curve, and sum of absolute change rate of vertical curvature. Meanwhile, all of these variables except speed limit have positive impacts on the number of head-on crashes.

Kononov *et al.* (2008) investigated the relationship between safety, congestion, and number of lanes on urban freeways. It is suggested that adding lanes may initially result in a temporary safety improvement that disappeared as congestion increases. Meanwhile, accident will increase at a faster rate than would be expected from a freeway with fewer lanes as annual average of daily traffic increases.

Schneider IV *et al.* (2010) explored the impacts of horizontal curvature and other geometric features on the frequency of single-vehicle motorcycle crashes along segments of rural two-lane highways. The findings show that the radius and length of each horizontal curve significantly influence the frequency of motorcycle crashes, as do shoulder width, annual average daily traffic, and the location of the road segment in relation to the curve.

Haleem and Gan (2011) identified and compared the factors that contribute to injury severity on urban freeways and arterials. Both traditional (such as traffic volume, speed limit, and road

surface condition) and nontraditional (such as crash distance to the nearest ramp location, detailed vehicle types, and lighting and weather conditions) factors are explored. The results reveal that the increase of the distance of crash to the nearest ramp junction/access point will significantly increases the severity of crashes. Also, other significant factors included traffic volume, speed limit, at-fault driver's age, road surface condition, alcohol and drug involvement, and left and right shoulder widths are also observed.

Wang *et al.* (2015) analyzed traffic safety on urban arterials using variables including geometric design features, land use, traffic volume, and travel speeds. The average speed extracted from GPS data from taxi was used in the study. It is found that the higher average speeds are associated with higher crash frequencies during peak periods, but not during off-peak periods. Besides, several geometric design features including average segment length of arterial, number of lanes, presence of non-motorized lanes, number of access points, and commercial land use, are found to positively related to crash frequencies.

(2) Intersection

Poch and Mannering (1996) estimated a negative binomial model of the crash frequency at intersections. The estimation results uncover important interactions between geometric and traffic-related elements and crash frequency.

Vogt (1999) explored crashes for different types of intersections on rural roads. The research revealed the variables having significant impacts on crashes: major and minor road traffic, peak major and minor road left-turning percentage, number of driveways, channelization, median

widths, and vertical alignment. As for the signalized intersections, the presence or absence of protected left-turn phases and peak truck percentage is also found significant.

Kim and Washington (2006) investigated the endogeneity problems for left-turn lanes at intersections. The research shows that without accounting for endogeneity, left-turn lanes appear to contribute to crashes; however, when endogeneity is accounted; left-turn lanes reduce angle crash frequencies as expected by engineering judgment.

Wang *et al.* (2006) studied crash risk at intersections with the consideration of time effects. The research identified the variables having significant effects on crash risk. Intersection with heavy traffic, a larger total number of lanes, a large number of phases per cycle, and high speed limits and those in high population were correlated with high crash frequencies. The intersections with more exclusive right-turn lanes with a partial left-turn protection phase had lower crash risks.

Wang and Abdel-Aty (2008) divided left-turn crashes at signalized intersections into nine patterns based on vehicle maneuvers and then were assigned to intersections approaches. The traffic flows to which the colliding vehicles belong are identified to be significant for each pattern. However, obvious differences in the other factors that cause the occurrence of different left-turn collision patterns were observed. The width of the crossing distance is associated with more left-turn traffic colliding with opposing through traffic, but with less left-turning traffic colliding with near-side crossing through traffic.

Ye *et al.* (2009) developed a simultaneous equations model of crash frequencies by collision type at rural intersections. Based on the modeling results, the significant common unobserved factors across crash types were observed.

Schneider *et al.* (2010) conducted study for pedestrian crashes at intersections. By using negative binomial regression, the authors found that significantly more pedestrian crashes occurred at intersections with more right-turn-only lanes, more nonresidential driveways with 50ft, more commercial properties with 0.1mi, and a greater percentage of residents with 0.25mi who were younger than age 18 years. Besides, raised medians on both intersecting streets were associated with lower number of pedestrian crashes.

Haleem and Abdel-Aty (2010) conducted analysis of crash injury severity at three- and four-legged intersections in the state of Florida. Several important factors affecting crash severity were identified. These include the traffic volume on the major approach, the number of through lanes on the minor approach, among the geometric factors, the upstream and downstream distance to the nearest signalized intersection, shoulder width, number of left turn movements on the minor approach, and number of right and left turn lanes on the major approach.

Pulugurtha and Sambhara (2011) developed different models of pedestrian crashes at different signalized intersections. This study found that socio-demographic characteristics have significant effects on pedestrian crashes.

Dong *et al.* (2014a) developed multivariate regression models for crash frequencies by collision vehicle types at urban signalized intersections. The results suggest that traffic volume, truck

percentage, lighting condition, and intersection angle significantly affect intersection safety. Besides, the important differences in car, car-truck, and truck crash frequencies with respect to various factors are found to exist between models.

Agbelie and Roshandeh (2015) investigated the impacts of signal-related characteristics on crash frequency at urban signalized intersections. The study found the significant association between signal phase and crash frequency, i.e., a unit increase in the number of signal phases would increase crash frequency by 0.4.

2.3.2 Characteristics for Microscopic Crash Analysis

A wide array of variables at the microscopic level has been investigated for crash at road facilities. Generally, the variables can be grouped into four categories: traffic, geometric features, control types, and environment conditions. The following parts will present discussion about different explanatory variables explored in micro-level safety studies.

(1) Traffic Characteristics

Traffic variable can play a vital role in crash occurrence. Noland and Quddus (2004) used proximate variables to represent the different traffic flow scenarios on road segments. The results indicated that traffic flow has a high influence on increasing causalities.

Wang et al. (2017) proposed a joint model to analyze the real-time crash risk and aggregated crash count by 5 minutes on freeway. The result suggested that the vehicle count in 5 minutes,

average speed, speed standard deviation, lane occupancy standard deviation, and truck percentage can affect the crash occurrence. Among the significant variables, the average speed is negatively related to crash risk while other variables have positive effects.

At the intersections, the prior studies indicated that traffic volumes including the AADT at both the major and minor road are significant for intersection crashes and are positively correlated with crash occurrence (Lee and Abdel-Aty, 2005; Mitra and Washington. 2012; Xie et al., 2013; Wang et al., 2017; Lee et al., 2017).

(2) Geometric Features

As for geometric features, Miaou et al. (1992) developed a count model to explore the relationships between trucks accidents and key highway geometric design variables. The final model suggested that annual average daily traffic per lane, horizontal curvature, and vertical grade were significantly correlated with truck accident involvement but that shoulder width has comparably less correlation.

Wang and Abdel-Aty (2006) investigate read-end crashes at signalized intersections. The study suggested that intersection having more right and left-turn lanes on the major roadway would like to have more rear-end crashes. On the other hand, intersections with three legs, having channelized or exclusive righr-turn lanes on the minor roadway, with protected left-turn on the major roadway, with medians on the minor roadway, and having longer signal spacing might have a lower frequency of rear-end crashes.

Park et al. (2015) assessed safety effects of different geometric feature related variables on urban roadway. The authors found that paved shoulder and wider median are significantly related lower crash frequency.

(3) Control Types

Roadway and intersection control types can definetely affect crash occurrence while the appropriate control types could help improve traffic safety (Cai et al., 2014; Wang et al., 2016). Wang and Abdel-Aty (2006) analyzed the rear-end crashes at signalized intersections considering the spatial correlation. It was found that intersections having a large number of phases per cycle (indicated by the left-turn protection on the minor roadway) and high speed limits on the major roadway were prone to have more rear-end crashes.

Wang et al. (2015) adopted a before-after study of converting a stop-controlled to a signal-control intersection and installing red light running cameras. The results of the signalization show that rear-end crashes were loewer at the early phase after the signalization but gradually increased from the 9^{th} month. On the other hand, the angle crashes became higher at the early phase after adding red light running cameras but decreased adter the 9^{th} month and then became stable.

Huang et al. (2017) proposed a multivariate spatial model to jointly analyze motor vehicle, pedestrian, and bicycle crashes at intersections. The study found that the traffic signal indicator is positively associated with all the crash types while the speed limits at both major and minor roades only have significant effects on motor vehicle crashes.

(4) Environment Conditions

The environment conditions especially the weather conditions are relevant to crash occurrence. Researchers have developed several ways to the effects of weather in the crash frequency models. Caliendo et al. (2007) adopted negative multinomial regression models to analyze crashes at a four-lane median-divided motorway in Italy. The effects of rain precipitation have been considered in this study by using hourly rainfall data and transforming them into binary indicators of daily status of the pavement surface (dry or wet).

Malyshkina et al. (2009) considered multiple weather variables such as precipitation, snowfall amounts, temperature averaged over weeks. The results indicated that more crashes would like to occur with extreme temperatures (low during winter and high during summer), rain precipitations, snowfalls, and low visibility conditions.

Usman et al. (2010) investigate the relationship between crash frequency during a snow storm event with the roadway surface conditions. Weather related variables including visibility, air temperature, and total precipitation were considered in the models. It was found that visibility was found to be significant with a negative sign in the models while air temperature and precipitation became insignificant.

Yu et al. (2013) investigate mountainous freeway crash by incorporating real-time weather and traffic data. The study concluded that the weather condition variables, especially precipitation, played a key role in the crash occurrence models.

Wu et al. (2017) introduced real-time traffic and weather data to compare crash risk under fog and clear condition on freeway roads. The results indicated that crash risk would increase under fog conditions; especially the traffic volume was high and on the inner-most lane.

It should be noted that some studies also included macro-level variables for the analysis of segments and intersections. Park et al. (2015) estimated segment-level crash models to evaluate the effectiveness of bicycle facilities. The authors included block-group based data including population density and income and found they are significantly related to the crash counts at segments.

For the intersections, macroscopic variables such as population density, proportion of young population, proportion of old population, proportion of workers commuting by walking, median household income, proportion of urbanized area, and school enrollment density have been adopted for the analysis of motor vehicle, pedestrian, and bicycle crashes (Wang et al., 2017; Lee et al., 2017).

2.4 Statistical Methodology

2.4.1 Statistical Models

For both macroscopic and microscopic safety analysis, a wide array of statistical techniques has been developed. Lord and Mannering (2010) and Mannering and Bhat (2014) presented summary of the statistical methods for crash frequency analysis. In this proposal, mainly statistical models for crash counts have been addressed: Poisson model, negative binomial model, Poisson lognormal model, models dealing with spatial spillovers effects and excess zeros.

The crash aggregated at a certain level, with any given time interval, are non-negative integer events. These integer counts are examined employing count regression models. The Poisson model is the traditional starting model for crash frequency analysis (Jovanis and Chang, 1986; Joshua and Garber, 1990; Sheather and Jones, 1991; Miaou and Lum, 1993).

The Poisson model can be calculated by:

$$\mathrm{P}(y_i) = {EXP(-\lambda_i)\lambda_i^{y_i}}\Big/{y_i!} \tag{2-1}$$

where, $\mathrm{P}(y_i)$ is the probability of entity i having y_i crashes by given time period and λ_i is the Poisson parameter for the entity (zone, segment, intersection, etc) i, which is equal to entity i's expected number of crashes per year, $\mathrm{E}[\,y_i\,]$. Poisson regression models are estimated by estimating the Poisson parameter λ_i (the expected number of crashes) as a function of explanatory variables:

$$\lambda_i = EXP(\beta X_i) \tag{2-2}$$

where, X_i is a vector of explanatory variables and β is a vector of estimable parameters.

The Poisson model assumes that the mean and variance of the distribution are the same. Thus, the Poisson model cannot deal with the over-dispersion (i.e. variance exceeds the mean).

The negative binomial (NB) or Poisson-gamma is extension of the Poisson model to deal with the over-dispersion problem. The NB model relaxes the equal mean variance assumption of

Poisson model and allows for over-dispersion parameter by adding an error term,ε_i, to the mean of the Poisson model as:

$$\lambda_i = \exp(\beta x_i + \varepsilon_i) \quad (2\text{-}3)$$

Usually, $\exp(\varepsilon_i)$ is assumed to be gamma-distributed with mean 1 and variance α so that the variance of the crash frequency distribution becomes $\lambda_i(1 + \alpha\lambda_i)$ and different from the mean λ_i.

The NB model has been most widely employed in crash count analysis (Maycock and Hall, 1984; Persaud, 1994; Kumala, 1995; Karlaftis and Tarko, 1998; Abdel-Aty and Radwan, 2000; Carson and Mannering, 2001; Miaou and Lord, 2003; Alaluusua *et al.*, 2004; Ladron de Guevara *et al.*, 2004; Lord *et al.*, 2005b; Kim *et al.*, 2006a; Wang *et al.*, 2006; Graham *et al.*, 2010; Abdel-Aty *et al.*, 2011a). The NB model can generally account over-dispersion resulting from unobserved heterogeneity and temporal dependency, but may be improper for accounting for the over-dispersion caused by excess zero counts (Rose et al., 2006).

Recently, a Poisson-lognormal (PLN) model was adopted as an alternative to the NB model for crash count analysis. The model structure of Poisson-lognormal model is similar to NB model, but the error term $\exp(\theta_i)$ in the model is assumed lognormal distributed. In other words, θ_i can be assumed to have a normal distribution with mean 0 and variance σ^2. Several crash studies have been conducted using PLN models (Miaou *et al.*, 2003; Aguero-Valverde and Jovanis, 2008; Lord and Miranda-Moreno, 2008; Ma *et al.*, 2008; El-Basyouny and Sayed, 2009; Haque *et al.*, 2010; Abdel-Aty *et al.*, 2013; Lee *et al.*, 2014; Lee *et al.*, 2015).

2.4.2 Handling Spatial Spillover Effects

In macroscopic and microscopic analysis, crashes occurring in a spatial unit or site are aggregated to obtain the crash frequency. The aggregation process might introduce errors in identifying the exogenous variables for the spatial unit or site. To accommodate for such spatial unit or site induced bias, spatial correlation should be considered in the crash model estimates. The inclusion of spatial correlation has two main advantages: 1) the spatial correlation model can realize the unobserved effects from neighboring sites, thereby improving model parameter estimation (Aguro-Valverde and Jovanis, 2008); and 2) spatial correlation can be a surrogate for unobserved but relevant covariates, which can reflect unmeasured confounding factors (Dubin, 1988; Choiu et al., 2014).

Two approaches to incorporate spatial correlation are considered: (1) spatial error correlation effects (unobserved exogenous variables at one location affect dependent variable at the targeted and neighboring locations) and (2) spatial spillover effects (observed exogenous variables at one location having impacts on the dependent variable at both the targeted and neighboring locations) (Narayanamoorthy *et al.*, 2013). Several research efforts have accommodated for spatial random error or spatial spillover effects in safety literature (LaScala *et al.*, 2000; Quddus, 2008; Ha and Thill, 2011). However, the utility of such spatially lagged dependent variable models, particularly for prediction, is limited since observed crash at neighboring spatial units is needed as an independent variable in the model.

Another alternative approach to accommodate the spatial dependency of in the count model is the conditional autoregressive model (CAR) (Besag *et al.*, 1991). The Conditional

Autoregressive (CAR) model takes account of both spatial dependence and uncorrelated heterogeneity with two random variables. Thus, the CAR model seems more flexible and appropriate for analyzing crash counts. Usually, the Poisson-lognormal Conditional Autoregressive (PLN-CAR) model, which adds a second error component (φ_i) as the spatial dependence (as shown below), was adopted for modeling.

The model can be specified by:

$$\lambda_i = \exp(\beta_i x_i + \theta_i + \varphi_i) \tag{2-4}$$

φ_i is assumed as a conditional autoregressive prior with Normal ($\overline{\varphi}_\iota, \frac{\gamma^2}{\sum_{i=1}^{K} w_{ki}}$) distribution recommend by Besag *et al.* (1991). The $\overline{\varphi}_\iota$ is calculated by:

$$\overline{\varphi}_\iota = \frac{\sum_{i=1}^{K} w_{ki}\varphi_i}{\sum_{i=1}^{K} w_{ki}} \tag{2-5}$$

where w_{ki} is the adjacency indication with a value of 1 if i and k are adjacent or 0 otherwise.

Efforts including Aguero-Valverde and Jovanis (2008), Huang *et al.*(2010), Lee *et al.* (2015), Siddiqui *et al.* (2012), and Dong et al. (2016), examine the potentially spatial correlation among crash data by employing the CAR models based on traffic analysis zones. According to these studies, the spatial models can generally provide consistent results and can better fit the crash data based on traffic analysis zones. Cai et al., 2017(a) conducted global Moran's *I* test to investigate whether spatial correlations existed among crash counts of different zonal systems including traffic analysis zone, census tract, and traffic analysis district. It was revealed that traffic analysis zone and traffic analysis district based crashes have strong spatial clustering

while crashes based on census tract were weakly spatial correlated. Hence, it is expected that the spatial CAR model can drastically improve data fit for crashes based on traffic analysis zone and traffic analysis district while no significant improvement can be obtained for the crashes based on census tract.

Beside the macroscopic crash analysis, the CAR model has been also adopted for segments and intersections at microscopic level. Wang and Abdel-Aty (2006) used the generilized estimating equations with the negative binomial link function to model rear-end crash frequencies at signalized intersections with the consideration of the spatial correlation among the crash data. The modeling results showed that there are high correlations between the spatially correlated rear-end crashes. The same study was also conducted by Abdel-Aty and Wang (2006), which validated the findings.

Aguero-Valverde and Jovanis (2008) explored the effect of spatial correlation in models of road crash frequency at the segment level. Different segment neighboring structures are tested to best fit the crash data. Compared with the model including only heterogeneity (random effects), the model with spatial correlation could have better goodness-of-fit. Also, based on the change in the estimate of the AADT coefficient and other parameters, the potential of spatial correlation would reduce the bias associated with the model misspecification.

Huang et al. (2017) proposed a multivariate spatial model to simultaneously analye the motor vehicle, bicycle, and pedestrian crash frequency at urban intersections. The proposed model can account for both the correlation among different modes involved in crashes at individual intersection and spatial correlation between adjacent intersections. This study condirmed the

highly correlated heterogeneous residuals in modeling crash risk among motor vehicle, bicycle and pedestrian crashes.

Beside the solely spatial correlation considered for either intersections or segments, Zeng and Huang (2014) suggested a joint spatial model which can consider the cross-entity spatial correlations. The spatial correlations between segments and the connceted intersection were found to be more significant than those solely between segments or between intersections. This joint modeling structure was also adopted by Wang and Huang (2016) and Huang et al. (2016).

In addition to the two method presented above, recently, a new method named geographically weighted Poisson regression model (GWPR) has been adopted in the crash count studies. The GWPR model allows the parameters to vary over space to capture the spatially varying relationships in the crash data. The model has been used for traffic safety analysis at the traffic analysis zone (Hayayeghi et al. 2010; Zhang et al., 2012; Pirdavani et al., 2013; Zhang et al., 2015; Xu and Huang, 2015; Shariat-Mohaymany et al., 2015; Amoh-Gyimah) and county levels (Li et al., 2013). It was revealed that the method outperformed the traditional generalized linear model in capturing the spatially varying relationship between crash counts and predicting factors.

2.4.3 Handling Excess Zeros

One methodological challenge often faced in analyzing count variables is the presence of a large number of zeros. The classical count models (such as Poisson and NB) allocate a probability to observe zero counts, which is often insufficient to account for the preponderance of zeros in a count data distribution. In crash count variable models, the presence of excess zeros may result

from two underlying processes or states of crash frequency likelihoods: crash-free state (or zero crash state) and crash state (see (Shankar *et al.*, 1997) for more explanation). The zero crash state can be a mixture of true zeros (where the zones are inherently safe (Shankar *et al.*, 1997)) and sampling zeros (where excess zeros are results of potential underreporting of crash data (Miaou, 1994)). In presence of such dual-state, application of single-state model (Poisson and negative binomial) may result in biased and inconsistent parameter estimates.

In econometric literature, two potential relaxations of the single-state count models are proposed for addressing the issue of excess zeros. The first approach – the zero inflated (ZI) model - is typically used for accommodating the effect of both true and sampling zeros, and has been employed in several transportation safety studies (Shankar *et al.*, 1997; Chin and Quddus, 2003). The second approach - the Hurdle model - is typically used in the presence of sampling zeros and has seldom been used in transportation safety literature. The two approaches differ in the approach employed to address the excess zeros. The appropriate framework for analysis might depend on the actual empirical dataset under consideration. In the traffic safety field, the zero-inflated model has been often applied to explore the relationship between crash counts and the covariates. However, the hurdle model has rarely been adopted in traffic safety literature. Table 2-1 presents a summary of previous studies that have considered zero-inflated and hurdle models to analyze crashes. The table provides information on type and severity of crash analyzed, spatial and temporal unit of analysis and the data collection duration. From the table, it is evident that all the existing zero-inflated and hurdle studies are conducted at a micro-level such as segment and intersection except for Brijs *et al.* (2006) and Cai et al., (2016), which conducted crash analysis at macro-level by assigning crashes to the closest weather station. Second, with the exception of study (Hu *et al.*, 2011; Hosseinpour *et al.*, 2013; Hosseinpour *et al.*, 2014), the range of

observation of the study period is one year or less, that may explain the preponderance of zeros in the data (Lord *et al.*, 2005a). Third, the zero-inflated model always offers better statistical fit to crash data.

To be sure, several research studies have criticized the application of zero-inflated model for traffic safety analysis (Lord *et al.*, 2005a; Lord *et al.*, 2007; Kweon, 2011). The authors question the basic dual-state assumption for crash occurrence and have conducted extensive analysis at the micro-level and indicated that the development of models with dual-state process is inconsistent with crash data at the micro-level. While the reasoning behind the "non-applicability" is plausible for micro-level the reasoning does not necessarily carry over to the macro-level crash counts. For example, at the macro-level it is possible to visualize dual-state data generation with some macro-level units having zero pedestrian and bicyclist crashes – possibly because these spatial units have no pedestrian and bicycle demand (because of lack of walking and cycling infrastructure). In such cases the dual-state representation will allow us to identify spatial units that are likely to have zero cases as a function of exogenous variables (for example very low walking and cycling infrastructure might result in the higher probability of a zero state). Hence, we have considered the possible existence of dual-state models for pedestrian and bicycle crashes at the macro level in our research. If the data generation does support the dual-state models, ignoring the excess zeros and estimating traditional NB models will result in biased estimates.

Table 2-1 Summary of Previous Traffic Safety Studies Using Dual-State Models

Methodology	Study	Crash types	Spatial Unit	Temporal Unit	Number of Study Years
Zero-inflated	Shankar *et al.* (1997)	Total crashes	Road segment	2 years	2 years
	Miaou (1994)	Truck crashes	Road segment	1 year	5 years
	Chin and Quddus (2003)	Total/pedestrian/motorcycle crashes	Signalized intersection	1 year	1 year
	Brijs *et al.* (2006)	Total crashes	Weather station	1 hour	1 year
	Hu *et al.* (2011)	Total crashes	Railroad-grade crossing	3 years	3 years
	Carson and Mannering (2001)	Crashes in ice condition	Road segment	1 year	3 years
	Lee and Mannering (2002)	Run-off-roadway crashes	Road segment	1 month	3 years
	Mitra *et al.* (2002)	Head-to-side/head-to-rear crashes	Signalized intersection	1 year	8 years
	Kumara and Chin (2003)	Total crashes	Signalized intersection	1 year	9 years
	Shankar *et al.* (2004)	Pedestrian crashes	Road segment	1 year	1 year
	Qin *et al.* (2004)	Single-vehicle/multi-vehicle crashes	Road segment	1 year	4 years
	Huang and Chin (2010)	Total crashes	Signalized intersection	1 year	8 years
	Jang *et al.* (2010)	Total crashes	Road segment	1 year	1 year
	Dong *et al.* (2014b)	Truck/Car crashes	Intersection	1 year	5 years
	Dong *et al.* (2014c)	Crashes by severity	Intersection	1 year	5 years
Hurdle	Hosseinpour *et al.* (2013)	Pedestrian crashes	Road segment	4 years	4 years
	Hosseinpour *et al.* (2014)	Head-on crashes	Road segment	4 years	4 years
	Kweon (2011)	Total crashes	Road segment	< 1 hour	6 years

2.4.4 Handling Multilevel Effects

A variety of factors can potentially affect the likelihood of crash occurrence including human elements such as gender, age, and driver-passenger-related behaviors, vehicle characteristics such as vehicle-type and model year, safety-feature indicators, road characteristics such as median barrier presence, type indicators, shoulder and lane widths, and curves, traffic characteristics such traffic volume, traffic vehicle mix, speed-related measurements, naturalistic driving data, environmental characteristics such as time of day, weather conditions, and lighting conditions (Mannering et al., 2016). The potential factors may be from multiple levels.

Huang and Abdel-Aty (2010) proposed a five-level hierarchy (i.e., geographic region level, traffic site level, traffic crash level, driver vehicle unit level, and occupant level) to represent the general framework of multilevel data structures in crash data. This study suggested that factors affecting crash occurrence are from multiple levels and from both macroscopic and microscopic levels. The macroscopic level includes the top three levels: geographic region level, traffic site level, and traffic crash level while the microscopic level concerns the bottom three levels: traffic crash level, driver vehicle unit level, and occupant level. The hierarchical technique was suggested to account for the multilevel effects of crash frequency. The hierarchical modeling is a statistical technique which allows parameters estimates based on a multiple modeling structure (Gelman and Hill, 2007).

Shankar et al. (1998) estimated a hierarchical model by including site –specific random effects and time indicators into the negative binomial model to evaluate the effect of median crossover on the crash occurrence. It was showed that the inclusion of site and time indicator can

significantly improve performance of modeling results. Remarkably, the model was the first application in traffic safety study.

Jones and Jorgensen (2003) estimated hierarchical models for fatal and severe crashes in Norway. The benefits of using hierarchical modeling technologies to analyze crash data were discussed along with the limitations of traditional regression modeling approaches.

Haque et al. (2010) estimated different hierarchical Poisson models for the crash data accounting for the site-specific correlation at signalized intersections. It was found that the hierarchical model allowing autoregressive lag-1 dependence specification in the error term is the most suitable.

Ahmed et al. (2011) employed Bayesian hierarchical models to account for seasonal and spatial correlations at freeway segment. Such approach was also adopted by Yu et al. (2013) to investigate the real-time weather and traffic effects on the crashes of mountainous freeway in two different seasons.

Wang and Huang (2016) developed a Bayesian hierarchical joint model for both segments and intersections. The proposed model accounted for two-level effects of microscopic variables related to road facilities and traffic volume and macroscopic variables such as socioeconomic, trip generation, and network density. In addition, spatial correlation between segments and intersections were considered in the proposed model. By comparing the prosed hierarchical model with the previous joint model and a negative binomial model, it was concluded that the hierarchical model outperforms the joint model and negative binomial model in terms of the

goodness-of-fit, which suggested the reasonableness of accounting for the multilevel effects in the crash data analysis.

Lee et al. (2017) estimated multiple hierarchical models for total, severe, pedestrian, and bicycle crashes at intersections with macro-level data of several spatial units including census block, census block group, traffic analysis zone, census tract, ZIP-code tabulation area, traffic analysis district, census county division, and county. The results indicated that considering macro-level effects from ZIP-code tabulation area can provide best model performance for total, severe, and bicycle crashes, and including the census-tract-based effects can better explain the pedestrian crashes. It was also uncovered that the intersection crash models can be drastically improved by considering macro-level effects, even only including random effects for macro-level entities.

2.4.5 Handling Correlations between Crash Types

The frequency of different crash types occurred in the same zones and road facilities could be inter-related with each other. For example, all crashes with different types (head-on, rear-end, angular, collision with a stationary object, etc.) at the intersections could be affected by the signal control of the intersection and road geometry (Mannering and Bhat, 2014). In the previous literature, a variety of studies have adopted advanced multivariate for multiple crash types to recognize the correlation between the dependent variables.

For example, Song *et al.* (2006) developed Bayesian multivariate models to account for the interaction in different crash types (i.e., intersection, intersection-related, driveway access, and non-intersection) at the county level.

Ma and Kockelman (2006) adopted a multivariate Poisson model to simultaneously analyze crash counts within different severity levels by using a Bayesian technology, which could provide a systematic approach to estimate count data correlated with each other.

Park and Lord (2007) adopted the multivariate Poisson-lognormal model (MVPLN) to analyze crash frequency by severity levels. It was indicated that the MVPLN model would be able to account for the over-dispersion of the discrete crash data.

The same approach was adopted by EI-Basyouny and Sayed (2009) to jointly investigate crash frequency for different severity levels. A comparison analysis was conducted with the univariate models by using the goodness-of-fit measures and hazardous location identification. The results indicated that the MVPLN model can provide better performance compared with the univariate models.

Ye et al., (2009) also estimated multivariate Poisson models to analyze different crash types at the same time. The unobserved correlation effects have been recognized through the error covariance.

Wang and Kockelman (2013) proposed and developed a multivariate Poisson log-normal CAR model for pedestrian crashes based on census tracts and revealed the correlation across different severity levels of pedestrian crashes.

Lee *et al.* (2015a) estimated multivariate Poisson-lognormal models to analyze motor-vehicle, pedestrian, and bicycle crashes. Within unobserved sheared factors of geographic units, the dependencies across the different crash types were recognized.

Lee *et al.* (2015b) developed multivariate models to simultaneously analyze pedestrian crashes based on pedestrian crashes per crash location ZIP (ZIP code areas at which pedestrian crashes occurred) and crash-involved pedestrians per residence ZIP (ZIP code areas where the crash-involved pedestrians resided in). It was revealed that the product of 'Log of population' and 'Log of vehicle miles travelled (VMT)' which can reflect both population and traffic volume at the same time was the best exposure variable for pedestrian crashes per crash location's ZIP, whereas 'Log of population' was the best exposure variable for crash-involved pedestrian per residence ZIP. A random term was also found significant across the two models indicating the existence of correlation between the two dependent variables.

Furthermore, Nashad et al. (2016) developed a multivariate model by adopting a copula based bivariate negative binomial model for pedestrian and bicyclist crash frequency analysis. The authors found that the pedestrian crash count and the bicyclist crash count are more highly correlated with each other in a zone with more public transit commuters and higher school enrollment density.

Beside the multivariate modeling technology, Lee *et al.* (2016) proposed a framework where the impacts of exogenous variables are directly related to all count variables of interest simultaneously i.e. the framework where the observed propensities of crashes by different

transportation mode interact directly. The authors adopted a multinomial factional split model to explore the proportion of crashes (not frequency) by crash of different transportation modes.

Meanwhile, Yasmin *et al.* (2016) explored the dependencies between crashes of different severity levels with the same approach. The factional split modeling approach can explore the interaction between different crash types by providing more insights on the impact of exogenous variables on crash proportions.

In the earlier research, the interaction across different crash counts can be examined through either unobserved effects (count model) or exogenous variables (fractional split model). However, the direct interaction of crash counts still cannot be determined. For example, it is not clear the amount of total crashes in a zone can interact with the amount of pedestrian or bicycle crashes.

2.4.6 Handling Unobserved Heterogeneity

As introduced in the previous sections, a wide array of variables has been collected for the crash analysis. With commonly collected data, some of the factors which can affect crash occurrence may not available, resulting in variation in the impact of the effects of collected variables on the collision likelihood (Mannering et al., 2016). The unavailable factors would contribute to the unobservable heterogeneity in the crash modeling analysis. The effects of observable variables would be restricted to be the same across all observation if unobserved heterogeneity is ignored. Then, the model estimates could be biased and misleading. In the previous study, there are

generally two general approaches to account for unobserved heterogeneity: random parameter and latent segmentation.

The random parameters approach has been widely adopted in the previous studies at both macro- and micro-levels. The idea of the random parameters approach is that the heterogeneity from one observation to another is considering by allowing each estimated parameter to vary across all observations based on specified continuous distribution (such as normal distribution). A simple random parameter model would only allow the constant term varies across alternative, which has been adopted in an abundance of researches (Shankar et al., 1998; Miaou and Lord, 2003; Flahuat et al., 2003; Miaou et al., 2009; Wang and Abdel-Aty, 2006; Aguero-Valverde and Jovanis, 2006; Kim et al., 2007; Aguero-Valverde and Jovanis, 2008; Li et al., 2008; Guo et al., 2010; Aguero-Valverde and Jovanis, 2010; Ahmed et al., 2011; Yu et al., 2013; Yu and Abdel-Aty, 2013; Xie et al., 2014; Lee et al., 2015a, Lee et al., 2015b; Cai et al., 2017). Other studies assumed that all parameters have different distributions and a variety of distributions can be tested to determine which would provide the best statistical fit (Anastasopoulos and Maneering, 2009; EI-Basyouny and Sayed, 2009; Granowski and Maneer, 2011; Venkataraman et al., 2011; Ukkusuri et al., 2011; Mitra and Washington, 2012; Wu et al., 2013; Bullough et al., 2013; Castro et al., 2012; Naraysnamoorthy et al., 2013; Bhat et al., 2014a; Bhat et al., 2014b; Venkateraman et al., 2013; Chen and Tarko 2014; Xu and Huang, 2014; Venkataraman et al., 2014; Barua et al., 2015; Coruh et al., 2015; Barua et al., 2016; Buddhavarapu et al., 2016; Xu et al., 2017).

On the other hand, the latent segmentation approach addresses the unobserved heterogeneity by assuming finite mixtures (latent classes). This approach, instead of assuming heterogeneity vary

across all observation, seeks to identify clusters of observations with homogeneous observable variable effects in each cluster. A parametric requires a parametric model structure such as negative binomial with a logit model. Such approach has been adopted in several studies to examine heterogeneity in crash data (Park and Lord, 2009; Park et al., 2010; Peng and Lord, 2011; Zou et al., 2013; Zou et al., 2014; Yasmin et al., 2014; Yasmin et al., 2016; Buddhavarapu et al., 2016).

It should be noted that several studies criticized that the modeling results accounting for unobserved heterogeneity will not be transferable to different locations since the individual parameter vector associated with each data observation is unique to another. To admit, the random parameter modeling result may be not very easy to be transformed from one data set to other data sets. However, the unobserved heterogeneity could be presented at the individual level by the random parameter model. As for the fixed-parameters model, the transferability could also be problematic since the model estimates would be likely to be biased and the bias will be a function of unobserved heterogeneity.

2.5 Summary

Considerable studies have been conducted to analyze traffic crashes at both macroscopic and microscopic levels. At the macroscopic level, many studies have been conducted for different modes-vehicle (automobiles and motorbikes), pedestrian and bicycle and based on different zonal systems such as block groups, census tracts, or traffic analysis zones. There are several issues in the macroscopic crash analysis: 1) spatial autocorrelation, 2) modifiable areal unit problem, 3) excess zeros, 4) unidentified effects of explanatory variables. First, spatial

autocorrelation generally exists among zones in close proximity and should be considered in the crash analysis. Second, clear guidelines for the appropriate zonal system selection for safety analysis should be suggested to deal with the modifiable areal unit problems. Third, appropriate models should be suggested to deal with traffic analysis zone based pedestrian and bicycle crashes, which have excess zeros. Lastly, the different effects of explanatory variables on drivers and pedestrians or bicyclists should be thoroughly explored for pedestrian and bicycle crashes.

As for the microscopic crash analysis, the impacts of variables such as traffic, geometry, and signal control on crashes have been analyzed. However, most studies omitted the macroscopic data, which may result in biased and inconsistent parameter estimates. The hierarchical model might be appropriate to investigate the macro-level effects for the crash analysis for segments and intersections. Besides, the potentially spatial autocorrelation should be considered between segments and intersections.

Previous studies have explored traffic safety at either the macroscopic or microscopic level, i.e., to the best of the author's knowledge no study has integrated the two levels. If traffic safety research is conducted for the same study area, macro- and micro-level crash analyses would investigate the same crashes but by different aggregation levels. Hence, we can assume that the crash counts at the two levels are correlated. Therefore, an integrated crash frequency analysis would improve the model performance for both levels and can help in better understanding the crash mechanism as well.

CHAPTER 3: PEDESTRIAN AND BICYCLE CRASH ANALYSIS BASED ON TRAFFIC ANALYSIS ZONES

3.1 Introduction

As stated in the previous chapter, there are numerous studies to deal with preponderance of zeros in microscopic crash analysis. However, very limited analysis has been conducted for the excess of zero at macroscopic level. In this chapter, macroscopic analysis about non-motorized crashes is presented along two directions: (1) evaluate the viability of dual-state models for non-motorized crash analysis at macro-level; and (2) introduction of spatial independent variables accounting for spatial spillover effects on crash frequency. Towards this end, conventional single-state model (i.e., NB) and two dual-state models (i.e., zero-inflated NB (ZINB) and hurdle NB (HNB)) with and without spatial independent variables are developed for both pedestrian and bicycle crashes at a TAZ level in Florida. Overall, 6 model structures are estimated for pedestrian and bicycle crashes - NB model without/with spatial effects (aspatial/spatial NB), ZINB model without/with spatial effects (aspatial/spatial ZINB), and HNB model without/with spatial effects (aspatial/spatial HNB). The model development process considers a sample for model calibration and a hold-out sample for validation. A comparison exercise is undertaken to identify the superior model in model estimation and validation. Finally, average marginal effects are computed for the best model to assess the effect of different factors, including the spatial variables on crash occurrence.

This chapter is organized into six sections. The second section discusses the research methodology. The following section describes the data used. The forth section presents the

modeling results and the fifth section computes the marginal effects of the significant variables. Finally, the sixth section concludes this chapter.

3.2 Methodology

3.2.1 Single-state models

The Poisson model is the traditional starting model for crash frequency analysis (Lord and Mannering, 2010). The Poisson model assumes that the mean and variance of the distribution are the same. Thus, the Poisson model cannot deal with the over-dispersion (i.e. variance exceeds the mean). The NB model relaxes the equal mean variance assumption of Poisson model and allows for over-dispersion parameter by adding an error term, ε_i, to the mean of the Poisson model as:

$$\lambda_i = \exp(\beta_i x_i + \varepsilon_i) \tag{3-1}$$

where λ_i is the expected number of Poisson distribution for entity i, x_i is a set of explanatory variables, and β_i is the corresponding parameter. Usually, $\exp(\varepsilon_i)$ is assumed to be gamma-distributed with mean 1 and variance α so that the variance of the crash frequency distribution becomes $\lambda_i(1 + \alpha\lambda_i)$ and different from the mean λ_i. The NB model for the crash count y_i of entity i is given by

$$P(y_i)=\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}\left(\frac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}} \tag{3-2}$$

where y_i is the number of crashes y_i of entity i and $\Gamma(\cdot)$ refers to the gamma function. The NB model can generally account over-dispersion resulting from unobserved heterogeneity and temporal dependency, but may be improper for accounting for the over-dispersion caused by excess zero counts (Rose et al., 2006).

3.2.2 Dual-state models

(1) Zero-inflated model

The zero-inflated models assume that the data have a mixture with a degenerate distribution whose mass is concentrated at zero (Lambert, 1992). The first part of the mixture is the extra zero counts and the second part is for the usual single state model conditional on the excess zeros. The zero-inflated NB model can be regarded as an extension of the traditional NB specification as:

$$y_i \sim \begin{cases} 0, & with\ probability\ p_i \\ NB, & with\ probability\ 1-p_i \end{cases} \tag{3-3}$$

The logistic regression model is employed to estimate p_i,

$$p_i = \frac{\exp(\beta_i' x_i)}{1+\exp(\beta_i' x_i)} \tag{3-4}$$

where β_i' is the corresponding parameter.

Substituting Eq. (3-2) into Eq. (3-3) we can define ZINB model for crash counts y_i of entity i as

$$P(y_i) = \begin{cases} p_i + (1-p_i)\left(\frac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}}, & y_i = 0 \\ (1-p_i)\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\frac{(\alpha\lambda_i)^{y_i}}{(1+\alpha\lambda_i)^{(y_i+\frac{1}{\alpha})}}, & y_i > 0 \end{cases} \tag{3-5}$$

(2) Hurdle models

The Hurdle models, proposed by Mullahy (1986), can be regarded as two-part models. The first part is a binary model dealing with whether the response crosses the "hurdle", and the second part is a truncated-at-zero count model. Assume that the first hurdle part of process is governed by function f_1 and the second count process follows a truncated-at-hurdle function f_2. The Hurdle models are defined as follows:

$$P(y_i) = \begin{cases} f_1(0) = p_i, & y_i = 0 \\ (1 - f_1(0)) \dfrac{f_2(j)}{1 - f_2(0)}, & y_i > 0 \end{cases} \tag{3-6}$$

Hurdle NB model is obtained by specifying $f_2(\cdot)$ as the NB distribution. Substitution Eq. (3-2) into Eq. (6) will result in ZINB model as follows:

$$P(y_i) = \begin{cases} p_i, & y_i = 0 \\ (1 - p_i)\left(1 - \dfrac{1}{(1 + \alpha\lambda_i)^{\frac{1}{\alpha}}}\right) \dfrac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha}\right)} \dfrac{(\alpha\lambda_i)^{y_i}}{(1 + \alpha\lambda_i)^{\left(y_i + \frac{1}{\alpha}\right)}}, & y_i > 0 \end{cases} \tag{3-7}$$

As in the zero-inflated model, logistic regression will be applied for modeling p_i.

3.3 Data Preparation

Pedestrian and bicycle involved crashes that occurred in Florida in the period of 2010-2012 were compiled for the analysis. The State of Florida has 8,518 TAZs, with about 16,240 pedestrian and 15,307 bicycle crashes recorded. Among the TAZs, as shown in Figure 3-1, 46.18% of them have zero pedestrian crash while 49.86% of them didn't have any bicycle crashes. The explanatory variables considered for the analysis can be grouped into three categories: traffic

(such as VMT (Vehicle-Miles-Traveled), proportion of heavy vehicle in VMT), roadway (such as signalized intersection density, length of bike lanes and sidewalks,), and socio-demographic characteristics (such as population density, proportion of families without vehicle, etc.).

As highlighted earlier, the current analysis focuses on accommodating the impact of neighboring TAZs on the crash frequency models. Towards this end, for every TAZ, the TAZs that are adjacent are identified. Based on the identified neighbors, a new variable based on the value of the each exogenous variable from surrounding TAZs is computed. The variables thus created capture the spatial spillover effects of the neighboring TAZs on crash frequency. The descriptive statistics of the crash counts and independent variables are summarized in the following table. Specifically, the table provides the values at a TAZ level as well as for the neighboring TAZ variables.

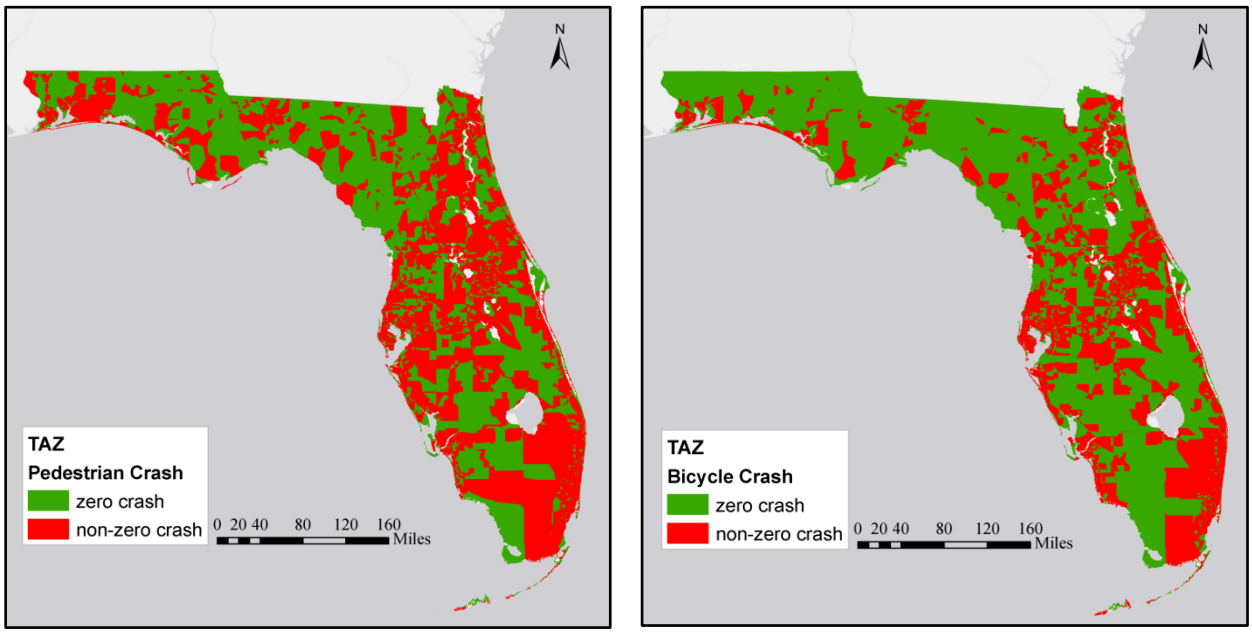


Figure 3-1 Pedestrian and bicycle crashes based on TAZs

Table 3-1 Descriptive statistics of collected data

Variables name	Targeted TAZs			Neighboring TAZs		
	Mean	**S.D.**	**Max**[a]	**Mean**	**S.D.**	**Max**[a]
Crash variables						
Pedestrian crash	1.907	3.315	39.000	-	-	-
Bicycle crash	1.797	3.309	88.000	-	-	-
Traffic & roadway variables						
VMT	31381.0	41852.3	684742.8	195519.7	169120.3	2103376.3
Proportion of heavy vehicle in VMT	0.067	0.052	0.519	0.070	0.045	0.350
Proportion of length of arterial roads	0.221	0.275	1.000	0.144	0.125	1.000
Proportion of length of collectors	0.191	0.246	1.000	0.156	0.136	1.000
Proportion of length local roads	0.572	0.329	1.000	0.680	0.200	1.000
Signalized intersection density (number of signalized intersections per mile)	0.227	0.578	8.756	0.378	5.552	495.032
Length of bike lanes	0.303	1.096	28.637	1.909	3.847	38.901
Length of sidewalks	0.993	1.750	25.683	6.304	6.745	77.720
Socio-demographic variables						
Population density	2520.3	4043.3	63069.0	2330.2	3489.7	57181.9
Proportion of families without vehicle	0.095	0.123	1.000	0.095	0.108	1.000
School enrollments density	775.02	5983.05	255147.24	684.22	2900.54	102285.73
Proportion of urban area	0.722	0.430	1.000	0.650	0.434	1.000
Distance to the nearest urban area	2.140	5.441	44.101	-	-	-
Hotels, motels, and timeshare rooms density	172.49	941.71	32609.84	121.678	528.078	11397.148
No of total employment	1140.10	1722.45	31932.15	6917.245	6725.135	76533.000
Proportion of industry employment	0.176	0.232	1.000	0.183	0.177	1.000
Proportion of commercial employment	0.299	0.235	1.000	0.305	0.177	1.000
Proportion of service employment	0.525	0.257	1.000	0.495	0.186	1.000
No of commuters by public transportation	18.813	54.273	934.000	119.582	246.299	3559.985
No of commuters by cycling	5.894	19.804	775.000	90.869	128.399	1902.135
No of commuters by walking	14.354	34.680	1288.000	37.566	74.484	1634.530

[a] The minimum values for all variables are zero.

3.4 Modeling Results and Discussion

3.4.1 Goodness of fit

In this study, from the 8518 TAZs, 80% of them were randomly selected for models calibration and 20% were used for validation of the estimated models. The overall model estimation process involved estimating six models - 3 model types (NB, ZINB, and HNB models) with and without spatial independent variables of neighboring TAZs for pedestrian and bicycle crashes. Prior to discussing the model results, we present the goodness of fit measures of the estimated models in Table 4-2. The table presents the Log-likelihood, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) - for the 6 models for estimation and validation samples. Several observations can be made from the results presented in Table 4-2.. First, across pedestrian and bicycle crash models, the models with spatial independent variables offer substantially better fit compared to models without spatial independent variables. The results validate our hypothesis that characteristics of adjacent TAZs improve our understanding of crash frequency in the target TAZ. Second, the exact ordering alters between ZINB and HNB in some cases based on log-likelihood and AIC. However, the ZINB model offers the best fit across all model structures based on the BIC. Among aspatial and spatial models, the ZINB model always has the lowest BIC value indicating strong difference between ZINB and other models. The ZINB improves data fit with only a small increase in number of parameters. Hence, in terms of our results, we can conclude that the ZINB offers the best statistical fit for pedestrian and bicycle crashes. Third, in validation exercise, it is further reinforced that ZINB offers the best data fit.

Table 3-2 Comparison of goodness-of-fits between different models

Pedestrian Crash						
	NB		ZINB		HNB	
Calibration (N=6815)	Aspatial	Spatial	Aspatial	Spatial	Aspatial	Spatial
No of parameters	15	17	20	22	24	28
Log-likelihood	-9972.4	-9926.6	-9944.3	-9890	-9964.4	-9912.5
AIC	19974.7	19887.3	19928.5	19824	19976.8	19881
BIC	20077.1	20003.3	20065.1	19974.2	20140.7	20072.2
Validation (N=1703)	Aspatial	Spatial	Aspatial	Spatial	Aspatial	Spatial
No of parameters	15	17	20	22	24	28
Log-likelihood	-2680.5	-2662.4	-2449.9	-2437.8	-2464.3	-2459.4
AIC	5391	5358.8	4939.7	4919.5	4976.5	4974.8
BIC	5472.6	5451.2	5048.5	5039.2	5107.1	5127.1
Bicycle Crash						
	NB		ZINB		HNB	
Calibration (N=6815)	Aspatial	Spatial	Aspatial	Spatial	Aspatial	Spatial
No of parameters	14	19	18	22	25	33
Log-likelihood	-9412.4	-9326.0	-9385.6	-9309.0	-9387.2	-9286.3
AIC	18852.9	18689.9	18807.2	18662.1	18824.3	18638.6
BIC	18948.5	18819.6	18930.1	18812.3	18995	18863.9
Validation (N=1703)	Aspatial	Spatial	Aspatial	Spatial	Aspatial	Spatial
No of parameters	14	19	18	22	25	33
Log-likelihood	-2771.6	-2785.9	-2393.4	-2355.6	-2396.4	-2364.8
AIC	5571.2	5609.8	4822.8	4755.2	4842.8	4795.7
BIC	5647.4	5713.2	4920.7	4874.9	4978.8	4975.2

3.4.2 Modeling Results

The results of six models (3 model types with and without spatial independent variables of neighboring TAZs) for pedestrian and bicycle crashes are displayed in Table 3-3 and Table 3-4 separately. The results for NB models only have the count frequency component. For zero-inflated and hurdle models, the modeling results consist of two components: (1) logistic model component for zero state and (2) the count frequency component. While the results for all 6

models for pedestrians and bicycle crashes are presented, the discussion focuses on the ZINB model with spatial independent variables that offers the best fit.

(1) Pedestrian crash models for TAZs

For ZINB model with spatial independent variables, twelve independent variables of targeted TAZs and four spatial independent variables are significant in the count component. The VMT variable is a measure of vehicle exposure and as expected increases the propensity for pedestrian crashes. However, with increase in heavy vehicle VMT, the likelihood of pedestrian traffic in these TAZs drops substantially thus negatively influencing crash frequency. Population density and total employment variables are surrogate measures of pedestrian exposure (Siddiqui et al., 2012). Hence, it is expected that these variables have positive impacts on crash frequency. The variables proportion of local roads by length, signalized intersection density, and length of sidewalks are reflections of pedestrian access and are likely to increase crash frequency. The number of hotels, motels and timeshare rooms reflects land use characteristics that are likely to encourage walking in the vicinity increasing pedestrian exposure. It is observed that in TAZs with higher number of commuters by walking and public transportation, the propensity for pedestrian crashes is higher. The commuters by walking and public transportation reflect zones with higher pedestrian activity resulting in increased crash risk (Abdel-Aty et al., 2013). As the distance of the TAZ centroid from the nearest urban region increases, pedestrian crash risk reduces – a sign of low pedestrian activity in the suburban regions.

Among the significant spatial spillover variables, the proportion of service employment corresponds to land use characteristics that attract pedestrians. Interestingly, the impact of

signalized intersection density of neighboring TAZs is found to be negatively associated with pedestrian crash frequency. This result is in contrast to the impact of the same variable for the targeted TAZ. A plausible explanation could be that, in TAZs with increased signalization in the neighborhood, drivers are expecting pedestrians and are likely to be alert reducing potential crashes whereas in TAZs with high signal intersection density but lower signal density in the neighborhood zones, the drivers are not expecting pedestrians thus reducing the benefit of signalization. The proportion of families without vehicles in the vicinity of TAZ represents captive individuals that are forced to use public transit and pedestrian/bicycle modes. Thus increased presence of such families is likely to increase pedestrian crash risk. Higher number of commuters by public transportation in the neighboring TAZs results in increased impact on crash frequency.

In the probabilistic component, only the length of sidewalks, number of total employment, and number of commuters by public transportation of the targeted TAZs are significant. As expected, these three variables are negatively associated with the propensity of zero pedestrian crashes. As these variables serve as surrogates for pedestrian activity, it is expected that TAZs with higher levels of these variables are unlikely to be assigned to the zero crash state. Interestingly, no spatial spillover effects are found to be significant in the probabilistic part.

Table 3-3 Models results for pedestrian crash of TAZs

	NB				ZINB				HNB			
Count Model	**Aspatial**		**Spatial**		**Aspatial**		**Spatial**		**Aspatial**		**Spatial**	
Parameter	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
Intercept	-4.513	0.139	-4.632	0.142	-4.202	0.159	-4.323	0.162	-3.504	0.187	-3.745	0.198
TAZ independent variables												
Log (VMT)	0.145	0.009	0.142	0.009	0.155	0.009	0.154	0.009	0.112	0.011	0.103	0.011
Proportion of heavy vehicle mileage in VMT	-1.108	0.416	-1.123	0.413	-1.424	0.422	-1.522	0.416	-1.890	0.556	-1.656	0.547
Log (population density)	0.124	0.011	0.105	0.011	0.102	0.011	0.093	0.011	0.115	0.014	0.097	0.014
Log (number of total employment)	0.235	0.013	0.225	0.013	0.205	0.015	0.195	0.015	0.186	0.017	0.186	0.017
Proportion of length of local roads	0.467	0.059	0.471	0.058	0.504	0.060	0.508	0.059	0.480	0.080	0.454	0.080
Log (signalized intersection density)	0.291	0.028	0.267	0.028	0.256	0.030	0.267	0.031	0.274	0.038	0.286	0.040
Log (length of sidewalks)	0.272	0.025	0.277	0.024	0.244	0.025	0.255	0.025	0.271	0.028	0.273	0.028
Log (hotels, motels, and timeshare rooms density)	0.022	0.006	0.026	0.006	0.021	0.006	0.030	0.006	0.030	0.007	0.037	0.007
Log (number of commuters by public transportation)	0.194	0.009	0.129	0.012	0.189	0.009	0.125	0.012	0.205	0.011	0.134	0.014
Log (number of commuters by walking)	0.067	0.011	0.065	0.011	0.052	0.012	0.056	0.012	0.057	0.013	0.060	0.013
Log (number of commuters by cycling)	0.027	0.011	0.031	0.011	0.027	0.011	0.030	0.011	-	-	-	-
Log (distance to nearest urban area)	-0.027	0.006	-0.024	0.006	-0.028	0.006	-0.025	0.006	-	-	-	-
Proportion of families without vehicle	-	-	-	-	0.717	0.136	-	-	-	-	-	-
Proportion of service employment	0.314	0.062	0.221	0.068	0.296	0.062	-	-	-	-	-	-
Spatial Independent Variables												
Proportion of service employment of neighboring TAZs	-	-	0.253	0.091	-	-	0.301	0.083	-	-	0.376	0.103
Log (signalized intersection density of neighboring TAZs)	-	-	-	-	-	-	-0.291	0.063	-	-	-0.211	0.073
Proportion of families without vehicle of neighboring TAZs	-	-	-	-	-	-	1.29	0.172	-	-	-	-
Log (number of commuters by public transportation of neighboring TAZs)	-	-	0.099	0.011	-	-	0.091	0.011	-	-	0.108	0.014
Dispersion	0.445	0.020	0.423	0.020	0.393	0.022	0.367	0.021	0.419	0.028	0.386	0.026
Probabilistic Model	**Aspatial**		**Spatial**		**Aspatial**		**Spatial**		**Aspatial**		**Spatial**	
Intercept	-	-	-	-	0.070	0.413	-0.047	0.431	5.733	0.237	5.791	0.238
TAZ independent variables												
Log (VMT)	-	-	-	-	-	-	-	-	-0.188	0.015	-0.184	0.015
Log (length of sidewalks)	-	-	-	-	-2.143	0.729	-1.995	0.715	-0.500	0.064	-0.502	0.064
Log (number of total employment)	-	-	-	-	-0.240	0.070	-0.232	0.072	-0.299	0.023	-0.295	0.023
Log (number of commuters by walking)	-	-	-	-	-0.527	0.153	-0.501	0.148	-0.138	0.027	-0.136	0.027
Proportion of length of local roads	-	-	-	-	-	-	-	-	-0.510	0.104	-0.516	0.104
Log (signalized intersection density)	-	-	-	-	-	-	-	-	-0.331	0.054	-0.319	0.054
Log (population density)	-	-	-	-	-	-	-	-	-0.164	0.019	-0.155	0.019
Proportion of service employment	-	-	-	-	-	-	-	-	-0.405	0.126	-0.413	0.127
Log (number of commuters by public transportation)	-	-	-	-					-0.247	0.025	-0.192	0.030
Log (number of commuters by cycling)	-	-	-	-	-	-	-	-	-0.074	0.032	-0.074	0.032
Log (distance to nearest urban area)	-	-	-	-	-	-	-	-	0.030	0.008	0.027	0.008
Spatial Independent Variables												
Log (number of commuters by public transportation of neighboring TAZs)	-	-	-	-	-	-	-	-	-	-	-0.075	0.022

All explanatory variables are significant at 95% confidence level

(2) Bicycle crash models for TAZs

In the ZINB model with spatial variables presented in Table 3-4 eleven variables for the TAZs and five variables of neighboring TAZs affect bicycle crash frequency. The impacts of exogenous variables in the bicycle crash frequency model are very similar to the impact of these variables in the pedestrian crash frequency model. This is not surprising because, TAZs that are likely to experience high pedestrian activity are also likely to experience high bicyclist activity.

For the count component, the exogenous variables for the TAZ that increase the crash propensity are VMT, population density, total employment, proportion of local roads by length, signalized intersection density, length of sidewalks, proportion of commuters by walking as well as cycling, and proportion of service employment. The exogenous variables for the TAZ that reduce crash propensity are proportion of heavy vehicle mileage and the distance of the TAZ centroid from the nearest urban region. There are three main difference in the TAZ variable impacts between pedestrian and bicyclist crash frequency. First, the number of commuters by public transportation does not impact crash frequency as it is possible that public transportation and bicycling are not as strongly correlated as is the case with public transportation and pedestrians. Second, the density of hotel, motel and time share rooms does not impact bicycle crash frequency as tourists are unlikely to be bicyclists. Third, the number of service employment in the TAZ affects bicycle crash frequency while affecting pedestrian crash frequency as a spillover effect. While, the exact reason for the result is unclear, it could be a manifestation of differences of how land-use affects pedestrians and bicyclists.

In terms of spatial spillover effects, the significant variables vary between pedestrian and bicyclists. Specifically, the high proportion of industry employment in neighboring TAZs has a negative influence on crash propensity as these regions are unlikely to have significant bicyclist exposure. The signalized intersection density exhibits the same relationship as described for pedestrian crashes. On the other hand, from the neighboring TAZs, population density, number of commuters by public transit and cycling are likely to increase bicycle crash propensity. These variables are surrogates for bicycle exposure and are expected to increase crash risk.

In the probabilistic component, only three explanatory variables of targeted TAZs variables are significant. The length of sidewalks, population density and total employment variables, as expected, have negative influence on assigning a TAZ to a zero-crash state. The bicycle crash probabilistic component also does not have any statistically significant spatial variables.

Table 3-4 Models results for bicycle crash of TAZs

	NB				ZINB				HNB			
Count Model	**Aspatial**		**Spatial**		**Aspatial**		**Spatial**		**Aspatial**		**Spatial**	
Parameter	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
Intercept	-4.650	0.154	-4.672	0.167	-4.090	0.181	-4.673	0.190	-3.620	0.220	-4.031	0.237
TAZ independent variables												
Log (VMT)	0.190	0.009	0.162	0.010	0.186	0.010	0.164	0.010	0.168	0.013	0.148	0.013
Proportion of heavy vehicle mileage in VMT	-4.260	0.485	-3.306	0.490	-4.244	0.487	-2.787	0.496	-4.115	0.665	-2.949	0.660
Log (population density)	0.152	0.013	0.130	0.013	0.133	0.014	0.087	0.015	0.131	0.018	0.084	0.020
Log (number of total employment)	0.193	0.014	0.194	0.014	0.157	0.016	0.161	0.016	0.142	0.018	0.134	0.018
Proportion of length of local roads	0.535	0.062	0.441	0.064	0.517	0.063	0.525	0.063	0.422	0.086	0.401	0.085
Log (signalized intersection density)	0.196	0.030	0.234	0.032	0.172	0.031	0.203	0.033	0.125	0.041	0.184	0.044
Log (length of sidewalks)	0.284	0.026	0.271	0.025	0.214	0.027	0.228	0.026	0.219	0.030	0.217	0.029
Log (number of commuters by public transportation)	0.106	0.010	0.086	0.012	0.107	0.010	-	-	0.096	0.012	0.084	0.012
Log (number of commuters by walking)	0.087	0.012	0.085	0.012	0.090	0.012	0.104	0.012	0.101	0.014	0.099	0.014
Log (number of commuters by cycling)	0.109	0.011	0.070	0.012	0.110	0.011	0.088	0.012	0.108	0.012	0.071	0.013
Log (distance to nearest urban area)	-0.103	0.011	-0.098	0.011	-0.097	0.011	-0.074	0.011	-0.092	0.024	-0.065	0.023
Proportion of service employment	0.205	0.066	0.153	0.067	0.192	0.066	0.173	0.067	-	-	-	-
Spatial Independent Variables												
Proportion of industry employment of neighboring TAZs	-	-	-0.361	0.106	-	-	-0.242	0.106	-	-	-	-
Log (signalized intersection density of neighboring TAZs)	-	-	-0.319	0.075	-	-	-0.473	0.069	-	-	-0.545	0.095
Log (population density of neighboring TAZs)	-	-	-	-	-	-	0.113	0.018	-	-	0.109	0.023
Log (number of commuters by public transportation of neighboring TAZs)	-	-	0.035	0.012	-	-	0.068	0.010	-	-	-	-
Log (number of commuters by cycling of neighboring TAZs)	-	-	0.093	0.012	-	-	0.073	0.012	-	-	0.098	0.014
Proportion of length of local roads of neighboring TAZs	-	-	0.354	0.125	-	-	-	-	-	-	-	-
Dispersion	0.481	0.022	0.443	0.021	0.425	0.022	0.397	0.021	0.454	0.031	0.406	0.028
Probabilistic Model	**Aspatial**		**Spatial**		**Aspatial**		**Spatial**		**Aspatial**		**Spatial**	
Intercept	-	-	-	-	1.565	0.489	1.296	0.509	5.452	0.241	5.700	0.279
TAZ independent variables												
Log (VMT)	-	-	-	-	-	-	-	-	-0.222	0.016	-0.217	0.017
Log (length of sidewalks)	-	-	-	-	-4.455	1.272	-4.819	1.563	-0.676	0.066	-0.681	0.066
Log (population density)	-	-	-	-	-0.149	0.05	-0.135	0.053	-0.177	0.021	-0.102	0.024
Log (number of total employment)	-	-	-	-	-0.328	0.058	-0.313	0.060	-0.236	0.023	-0.216	0.024
Proportion of heavy vehicle mileage in VMT	-	-	-	-	-	-	-	-	5.347	0.836	4.258	0.861
Proportion of length of local roads	-	-	-	-	-	-	-	-	-0.709	0.109	-0.696	0.112
Log (signalized intersection density)	-	-	-	-	-	-	-	-	-0.286	0.054	-0.243	0.056
Log (number of commuters by public transportation)	-	-	-	-	-	-	-	-	-0.210	0.025	-0.147	0.031
Log (number of commuters by walking)	-	-	-	-	-	-	-	-	-0.081	0.028	-0.079	0.028
Log (number of commuters by cycling)	-	-	-	-	-	-	-	-	-0.158	0.032	-0.099	0.035
Log (distance to nearest urban area)	-	-	-	-	-	-	-	-	0.098	0.013	0.082	0.013
Spatial Independent Variables												
Proportion of length of arterial of neighboring TAZs	-	-	-	-	-	-	-	-	-	-	1.337	0.290
Log (population density of neighboring TAZs)	-	-	-	-	-	-	-	-	-	-	-0.096	0.033
Log (hotels, motels, and timeshare rooms density of neighboring TAZs)	-	-	-	-	-	-	-	-	-	-	-0.041	0.018
Log (number of commuters by public transportation of neighboring TAZs)	-	-	-	-	-	-	-	-	-	-	-0.069	0.026
Log (number of commuters by cycling of neighboring TAZs)	-	-	-	-	-	-	-	-	-	-	-0.082	0.025

All explanatory variables are significant at 95% confidence level

3.5 Marginal effects

The ZINB has two components, the probabilistic and the count component with exogenous variables possibly affecting both components. Thus, it is not straight-forward to identify the exact magnitude of the variable impact. Hence, to facilitate a quantitative comparison of variable impacts, marginal effects for the ZINB for pedestrians and bicyclists are computed. The marginal effects capture the change in the dependent variable in response to a small change in the independent variables. The results of the marginal effect calculation are presented in Table 3-5. As is expected, the sign of the marginal effects closely follow the sign from model results described in Table 10 and11.

The following observations can be made based on the results presented. First, the impact of spatial spillover effects on the crash models is significant and is comparable to the influence of other exogenous variables. Hence, it is important that analysts consider such observed spatial spillover effects in crash frequency modeling. Second, the exogenous variable impacts on pedestrian and bicycle crash models are similar for a large number of variables including VMT, population density, total employment, number of commuters by walking, proportion of local road in length, and number of public transportation commuters in neighboring TAZs. Third, the exogenous variables such as proportion of heavy vehicle VMT, proportion of service employment, number of commuters by public transportation and cycling, proportion of families without vehicles in the neighboring TAZs, service employment and industry employment in neighboring TAZs have significantly different marginal impacts across the two models. Finally, as indicated by the marginal effects of the signalized intersection density the exogenous variable

for TAZ and neighboring TAZs could exhibit distinct effects both in sign and magnitude. The allowance of such non-linear impacts accommodates for heterogeneity in the data.

Table 3-5 Average marginal effect for ZINB model with spatial independent variables

	Pedestrian		Bicycle	
Variables	**dy/dx**	**S.E**	**dy/dx**	**S.E**
TAZ independent variables				
Log (VMT)	0.292	0.018	0.291	0.018
Proportion of heavy vehicle mileage in VMT	-2.888	0.791	-4.937	0.885
Log (population density)	0.176	0.021	0.162	0.027
Log (number of total employment)	0.382	0.027	0.302	0.027
Proportion of length of local roads	0.965	0.114	0.930	0.113
Log (signalized intersection density)	0.506	0.06	0.359	0.059
Log (length of sidewalks)	0.587	0.05	0.671	0.077
Log (hotels, motels, and timeshare rooms density)	0.056	0.011	-	-
Log (number of commuters by public transportation)	0.238	0.022	-	-
Log (number of commuters by walking)	0.131	0.021	0.184	0.021
Log (number of commuters by cycling)	0.057	0.02	0.156	0.021
Log (distance to nearest urban area)	-0.047	0.011	-0.132	0.019
Proportion of service employment	-	-	0.307	0.118
Spatial Independent Variables				
Proportion of service employment of neighboring TAZs	0.572	0.158	-	-
Proportion of industry employment of neighboring TAZs	-	-	-0.428	0.189
Log (signalized intersection density of neighboring TAZs)	-0.552	0.119	-0.838	0.124
Proportion of families without vehicle of neighboring T AZs	2.447	0.329	-	-
Log (population density of neighboring TAZs)	-	-	0.200	0.033
Log (number of commuters by public transportation of neighboring TAZs)	0.173	0.021	0.120	0.019
Log (number of commuters by cycling of neighboring TAZs)	-	-	0.130	0.021

3.6 Summary and Conclusion

With growing concern of global warming and obesity concerns, active forms of transportation offer an environmentally friendly and physically active alternative for short distance trips. A strong impediment to universal adoption of active forms of transportation, particularly in North

America, is the inherent safety risk for active modes of transportation. Towards developing counter measures to reduce safety risks, it is essential to study the influence of exogenous factors on pedestrian and bicycle crashes. This study contributes to safety literature by conducting a macro-level planning analysis for pedestrian and bicycle crashes at a Traffic Analysis Zone (TAZ) level in Florida. The study considers both single state (negative binomial (NB)) and dual-state count models (zero-inflated negative binomial (ZINB) and hurdle negative binomial (HNB)) for analysis. In addition to the dual-state models, the research proposes the consideration of spatial spillover effects of exogenous variables from neighboring TAZs. The model development exercise involved estimating 6 model structures each for pedestrians and bicyclists. These include NB model with and without spatial effects, ZINB model with and without spatial effects and HNB with and without spatial effects. The estimated model performance was evaluated for the calibration sample and the validation sample using the following measures: Log-likelihood, Akaike Information Criterion and Bayesian Information Criterion.

The model comparison exercise for pedestrians and bicyclists highlighted that models with spatial spillover effects consistently outperformed the models that did not consider the spatial effects. Across the three models with spatial spillover effects, the ZINB model offered the best fit for pedestrian and bicyclists. The model results clearly highlighted the importance of several variables including traffic (such as VMT and heavy vehicle mileage), roadway (such as signalized intersection density, length of sidewalks and bike lanes, and etc.) and socio-demographic characteristics (such as population density, commuters by public transportation, walking and cycling) of the targeted and neighboring TAZs. To facilitate a quantitative comparison of variable impacts, marginal effects for the ZINB for pedestrians and bicyclists are computed. The results revealed the importance in sign and magnitude of the spatial spillover

effect relative to other exogenous variables. Further, the marginal effects computation allowed us to identify factors that substantially increase crash risk for pedestrians and bicyclists. In terms of actionable information, it is important to identify zones with high public transit, pedestrian and bicyclist commuters and undertake infrastructure improvements to improve safety.

To be sure, the study is not without limitations. While the influence of spatial spillover effects is considered, we do not consider the impact of spatial unobserved effects. Extending the current approach to accommodate for unobserved spatial terms will be useful. Also, it is possible to hypothesize that there might be common unobserved factors that affect pedestrian and bicyclists. Future research extensions might consider such unobserved effects in the model structure.

CHAPTER 4: EXPLORING ZONE SYSTEMS FOR TRAFFIC CRASH MODELING

4.1 Introduction

As shown in the literature review, previous studies have made remarkable contribution to explore MAUP effects on macro-level crash analysis. However, the employed measures for the comparison can be largely influenced by the number of observations and the observed values. Thus, the comparison results might be limited in the studies (Lee *et al.*, 2014; Xu *et al.*, 2014) since the measures were calculated based on zonal systems with different number of zones.

To address the limitation, one possible solution is to compute the measures based on a third-party zonal system so that the calculation would have the same observations. Towards this end, a grid structure that uniformly delineates the study region is suggested as a viable option. Specifically, the crash models developed for the various zonal systems will be tested on the same grid structure. To ensure that the result is not an artifact of the grid size, several grid sizes ranging from 1 to 100 square miles will be considered.

This chapter will present study to compare different geographic units for macroscopic crash modeling analysis. Towards this end, both aspatial model (i.e., Poisson lognormal (PLN) and spatial model (i.e., PLN conditional autoregressive (PLN-CAR)) are developed for three types of crashes (i.e., total, severe, and non-motorized mode crashes) based on census tracts, traffic analysis zones, and a newly developed zone system – traffic analysis districts (see the following section for detailed information). Then, a comparison method is proposed to compare the

modeling performance with the same sample sizes by using grids of different dimensions. By using different goodness-of-fit measures, superior geographic units for crash modeling are identified.

4.2 Comparison between CTs, TAZs, and TADs

In Florida, the average area of CTs, TAZs, and TADs are 15.497, 6.472, and 103.314 square miles, respectively. Across the three geographic units, which are shown in Figure 4-1, a TAD is considerably larger than a CT and TAZ while a TAZ is most likely to have the smallest size. CTs boundaries are generally delineated by visible and identifiable features, with the intention of being maintained over a long time. On the other hand, both TAZs and TADs are developed for transportation planning and are always divided by physical boundaries, mostly arterial roadways. Usually, CTs and TAZs nest within counties while TADs may cross county boundaries, but they must nest within Metropolitan Planning Organizations (MPOs) (FHWA, 2011a)

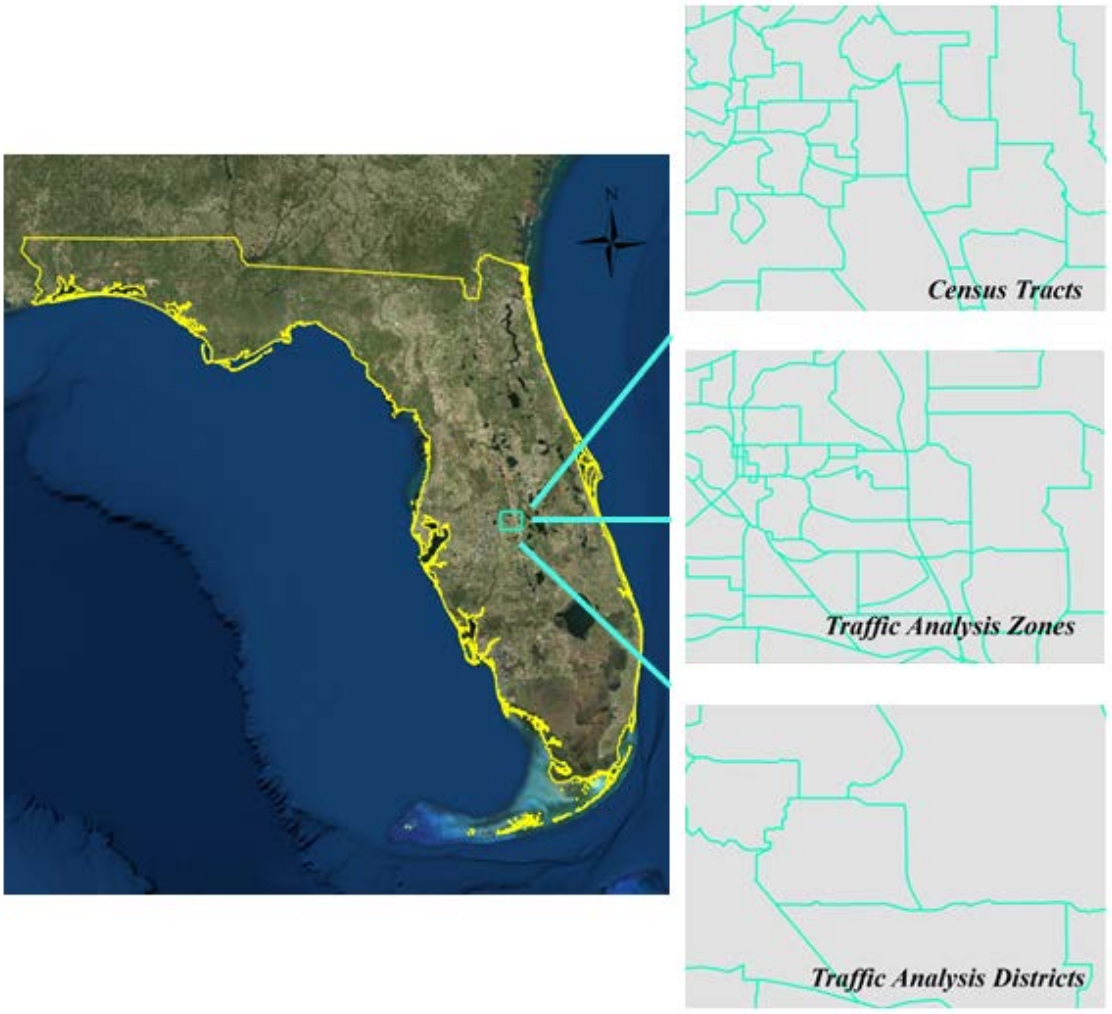


Figure 4-1 Comparison of CTs, TAZs, and TADs

4.3 Data Preparation

Multiple geographic units were obtained from the US Census Bureau and Florida Department of Transportation (FDOT). The state of Florida has 4,245 CTs, 8,518 TAZs, and 594 TADs. Crashes that occurred in Florida in 2010-2012 were collected for this study. A total of 901,235 crashes were recorded in Florida among which 50,039 (5.6%) were severe crashes and 31,547 (3.5%) were non-motorized mode crashes. In this study, severe crashes were defined as the combination of all fatal and incapacitating injury crashes while non-motorized mode crashes

were the sum of pedestrian and bicyclist involved crashes. On average, TADs have highest number of crashes since they are the largest zonal configuration. Given the large number of crashes in the Florida data, units with zero count are observed for CTs and TAZs. However, within a TAD no zero count units exist for the time period of our analysis. A host of explanatory variables are considered for the analysis and are grouped into three categories: traffic measures, roadway characteristics, and socio-demographic characteristics. For the three zonal systems, these data are collected from the Geographic information system (GIS) archived data from Florida Department of Transportation (FDOT) and U.S. Census Bureau (USCB). The traffic measures include VMT (Vehicle-Miles-Traveled), proportion of heavy vehicle in VMT. Regarding the roadway variables, roadway density (i.e., total roadway length per square mile), proportion of length roadways by functional classifications (freeways, arterials, collector, local roads, signalized intersection density (i.e., number of signalized intersection per total roadway mileage), length of bike lanes, and length of sidewalks were selected as the explanatory variables. Concerning the socio-demographic data, the distance to the nearest urban area, population density (defined as population divided by the area), proportion of population between 15 and 24 years old, proportion of population equal to or older than 65 years old, total employment density (defined as the total employment per square mile), proportion of unemployment, median household income, total commuters density (i.e., the total commuters per square mile), and proportion of commuters by various transportation modes (including car/truck/van, public transportation, cycling, and walking). It is worth mentioning that the distance to the nearest urban area is defined as the distance from the centroid of the CTs, TAZs, or TADs to the nearest urban region. So the distance will be zero if the zone is located in urban area. Also, it should be noted that the proportion of unemployment is computed by dividing the number of total unemployed people by the whole population. A summary of the crash counts and candidate explanatory variables on different zonal systems is also presented in Table 4-1.

Table 4-1. Descriptive statistics of collected data

Variables	Census tracts (N=4245)				Traffic analysis zones (N=8518)				Traffic analysis districts (N=594)			
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.
Area (square miles)	15.50	63.43	0.04	1581.94	6.47	24.80	0.00	885.32	103.31	259.86	2.62	3095.52
Crash variables												
Total crashes	212.31	234.96	0	4554.00	105.80	142.25	0	1507.00	1517.23	1603.29	188.00	15094.00
Severe crashes	11.79	11.78	0	141.00	5.87	7.94	0	111.00	84.24	60.34	4.00	534.00
Non-motorized mode crashes	7.43	7.96	0	76.00	3.70	6.08	0	121.00	53.11	60.09	1.00	562.00
Traffic & roadway variables												
VMT	91953.02	121384.56	0	1618443.43	31381.04	41852.30	0	684742.78	599646.92	428747.16	38547.00	4632468.60
Proportion of heavy vehicle in VMT	0.06	0.04	0	0.38	0.07	0.05	0	0.52	0.07	0.04	0.01	0.29
Road density	9.34	6.96	0	32.87	9.40	28.40	0	2496.05	7.61	5.31	0.07	24.56
Proportion of length of arterials	0.14	0.16	0	1.00	0.22	0.28	0	1.00	0.11	0.06	0.00	0.48
Proportion of length of collectors	0.13	0.14	0	1.00	0.19	0.25	0	1.00	0.11	0.07	0.00	0.60
Proportion of length of local roads	0.69	0.24	0	1.00	0.57	0.33	0	1.00	0.75	0.11	0.08	0.93
Signalized intersection density	4.09	227.17	0	14771.18	2.90	86.10	0	6347.67	0.12	0.13	0.00	1.36
Length of bike lanes	0.62	1.82	0	34.99	0.30	1.10	0	28.64	4.38	6.74	0.00	65.30
Length of sidewalks	1.73	2.27	0	20.84	0.99	1.75	0	25.68	12.93	11.94	0.00	87.18
Socio-demographic variables												
Distance to the nearest urban area	0.87	3.60	0	66.27	2.14	5.44	0	44.10	1.31	3.85	0.00	31.50
Population density	3255.00	3975.05	0	48304.10	2520.34	4043.35	0	63070.45	1998.61	1969.81	7.68	15341.30
Proportion of population age 15-24	0.13	0.08	0	1.00	0.13	0.08	0	1.00	0.13	0.06	0.03	0.69
Proportion of population age ≥ 65	0.18	0.14	0	0.94	0.17	0.12	0	0.94	0.17	0.09	0.03	0.66
Total employment density	2671.41	3350.12	0	45468.48	1770.29	2725.02	0	45468.48	1617.08	1609.59	6.84	13007.10
Proportion of unemployment	0.39	0.15	0	1.00	0.40	0.14	0	1.00	0.38	0.09	0.15	0.76
Median household income	59070.89	26477.95	0	215192.00	57389.53	24713.50	0	215192.00	59986.00	17747.51	21636.65	131664.42
Total commuters density	1477.99	2025.32	0	33066.11	926.73	1350.12	0	20995.26	900.67	904.09	3.60	6936.09
Proportion of commuters by vehicle	0.87	0.15	0	1.00	0.87	0.12	0	1.00	0.90	0.05	0.54	0.97
Proportion of commuters by public transportation	0.02	0.04	0	0.69	0.02	0.04	0	0.69	0.02	0.03	0.00	0.20
Proportion of commuters by cycling	0.01	0.03	0	1.00	0.01	0.03	0	1.00	0.01	0.01	0.00	0.17
Proportion of commuters by walking	0.02	0.04	0	1.00	0.02	0.04	0	0.46	0.01	0.02	0.00	0.14

4.4 Preliminary Analysis of Crash Data

The crash counts of different zonal systems were explored to investigate whether spatial correlations existed by using global Moran's I test. The absolute Moran's I value varies from 0 to 1 indicating degrees of spatial association. Higher absolute value represents higher spatial correlation while a zero value means a random spatial pattern. As shown in Table 4-2, all crash types based on different zonal systems have significant spatial correlation. TAZs and TADs based crashes have strong spatial clustering (Moran's I > 0.35) while crashes based on CTs were weakly spatial correlated (Moran's I < 0.1). It is not surprising since the TAZs and TADs were delineated based on transportation related activities. Thus, spatial dependence should be considered for modeling crashes, especially for TAZs and TADs.

Table 4-2 Global Moran's *I* Statistics for Crash Data

Crash types	**Total crashes**			**Severe crashes**			**Non-motorized crashes**		
Zonal systems	**CT**	**TAZ**	**TAD**	**CT**	**TAZ**	**TAD**	**CT**	**TAZ**	**TAD**
Observed Moran's I	0.06	0.52	0.58	0.05	0.40	0.36	0.05	0.424	0.447
P-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Spatial Autocorrelation	Y	Y	Y	Y	Y	Y	Y	Y	Y

4.5 Statistical Models

Before comparison across different zonal systems, both aspatial and spatial models were employed to analyze the crash data based on each zonal system. The technology of models is briefly discussed below.

4.5.1 Aspatial Models

In the previous study about crash count analysis, the classic negative binomial (NB) model has been widely used (Lord and Mannering, 2010). The NB model assumes that the crash data follows a Poisson-gamma mixture, which can address the over-dispersion issue (i.e., variance exceeds the mean). A NB model is specified as follows:

$$y_i \sim \text{Poisson}(\lambda_i) \tag{4-1}$$

$$\lambda_i = \exp(\beta_i x_i + \theta_i) \tag{4-2}$$

where y_i is the number of crashes in entity i, λ_i is the expected number of Poisson distribution for entity i, x_i is a set of explanatory variables, β_i is the corresponding parameter, θ_i is the error term. The $exp(\theta_i)$ is a gamma distributed error term with mean 1 and variance α^2.

Recently, a Poisson-lognormal (PLN) model was adopted as an alternative to the NB model for crash count analysis (Lord and Mannering, 2010). The model structure of Poisson-lognormal model is similar to NB model, but the error term $\exp(\theta_i)$ in the model is assumed lognormal distributed. In other words, θ_i can be assumed to have a normal distribution with mean 0 and variance σ^2. In our current study, the Poisson-lognormal model consistently outperformed the NB model. Hence, for our analysis, we restrict ourselves to Poisson-lognormal model comparison across different geographical units.

4.5.2 Spatial Models

Generally, two spatial model specifications were commonly adopted for modeling spatial dependence: the spatial autoregressive model (SAR) (Anselin, 2013) and the conditional autoregressive model (CAR) (Besag et al., 1991). The SAR model considers the spatial correlation by adding an explanatory variable in the form of a spatially lagged dependent variable or adding spatially lagged error structure into a linear regression model while the Conditional Autoregressive (CAR) model takes account of both spatial dependence and uncorrelated heterogeneity with two random variables. Thus, the CAR model seems more appropriate for analyzing crash counts (Quddus, 2008; Wang & Kockelman, 2013). A Poisson-lognormal Conditional Autoregressive (PLN-CAR) model, which adds a second error component (φ_i) as the spatial dependence (as shown below), was adopted for modeling.

$$\lambda_i = \exp(\beta_i x_i + \theta_i + \varphi_i) \tag{4-3}$$

φ_i is assumed as a conditional autoregressive prior with Normal ($\overline{\varphi_i}, \frac{\gamma^2}{\sum_{i=1}^{K} w_{ki}}$) distribution recommend by Besag et al. (1991). The $\overline{\varphi_i}$ is calculated by:

$$\overline{\varphi_i} = \frac{\sum_{i=1}^{K} w_{ki}\varphi_i}{\sum_{i=1}^{K} w_{ki}} \tag{4-4}$$

where w_{ki} is the adjacency indication with a value of 1 if i and k are adjacent or 0 otherwise.

In this study, both aspatial Poisson-lognormal model (PLN) and Poisson-lognormal Conditional Autoregressive model (PLN-CAR) were estimated. Deviance Information Criterion (DIC) was computed to determine the best set of parameters for each model and to compare aspatial and

spatial models based on the same zonal system. However, it is not appropriate for comparing models across different zonal systems since they have different sample size. Instead, a new method should be proposed for the comparison.

4.6 Method for Comparing Different Zonal Systems

4.6.1 Development of Grids for Comparison

Based on the estimated models, the predicted crash counts can be obtained for the three zonal systems. One simple method to compare the models based on different geographic units is to analyze the difference directly between the observed and predicted crash counts for each geographic unit. However, this method is not really comparable across the different geographical units due to differences in sample sizes. In this study, a new method was proposed to use grid structure as surrogate geographic unit to compare the performance of models based on different zonal systems. As shown in Figure 4-2, the grid structure, unlike the CT, TAZ, or TAD, is developed for uniform length and shape across the whole state without any artifact impacts. Furthermore, the numbers of grids remain the same for all models thereby providing a common comparison platform. To implement the procedure for comparison, the first step is to count the observed crash counts in each grid by using Geographic Information System (GIS). Then, the predicted crash counts of the three zonal systems are transformed separately to the grid structure based on a method is presented in detail in the next section. For each grid, six different values of the transformed crash counts (2 model types × 3 zonal systems) can be obtained. The difference between observed and transformed crash counts for each grid structure will be analyzed. Finally, by comparing the difference of different geographic units, the superior geographic unit between CTs, TAZs, and TADs can be obliquely identified for crash modeling with the same sample size. Additionally, to avoid the impact of grid size on the comparison results, we consider several

sizes for grids. Specifically, based on the average area of the three geographic units, ten levels of grid structures with side length from 1 to 10 miles were created. Table 3 summarizes the average areas and observed crash counts of CTs, TAZs, TADs, and different grid structures. The Grid L×L means the grid structure with side length of L miles. Based on the number of zones and average crash counts, it can be concluded that the CTs, TAZs, and TADs are separately comparable with Grid 4×4, Grid 3×3, and Grid 10×10, respectively.

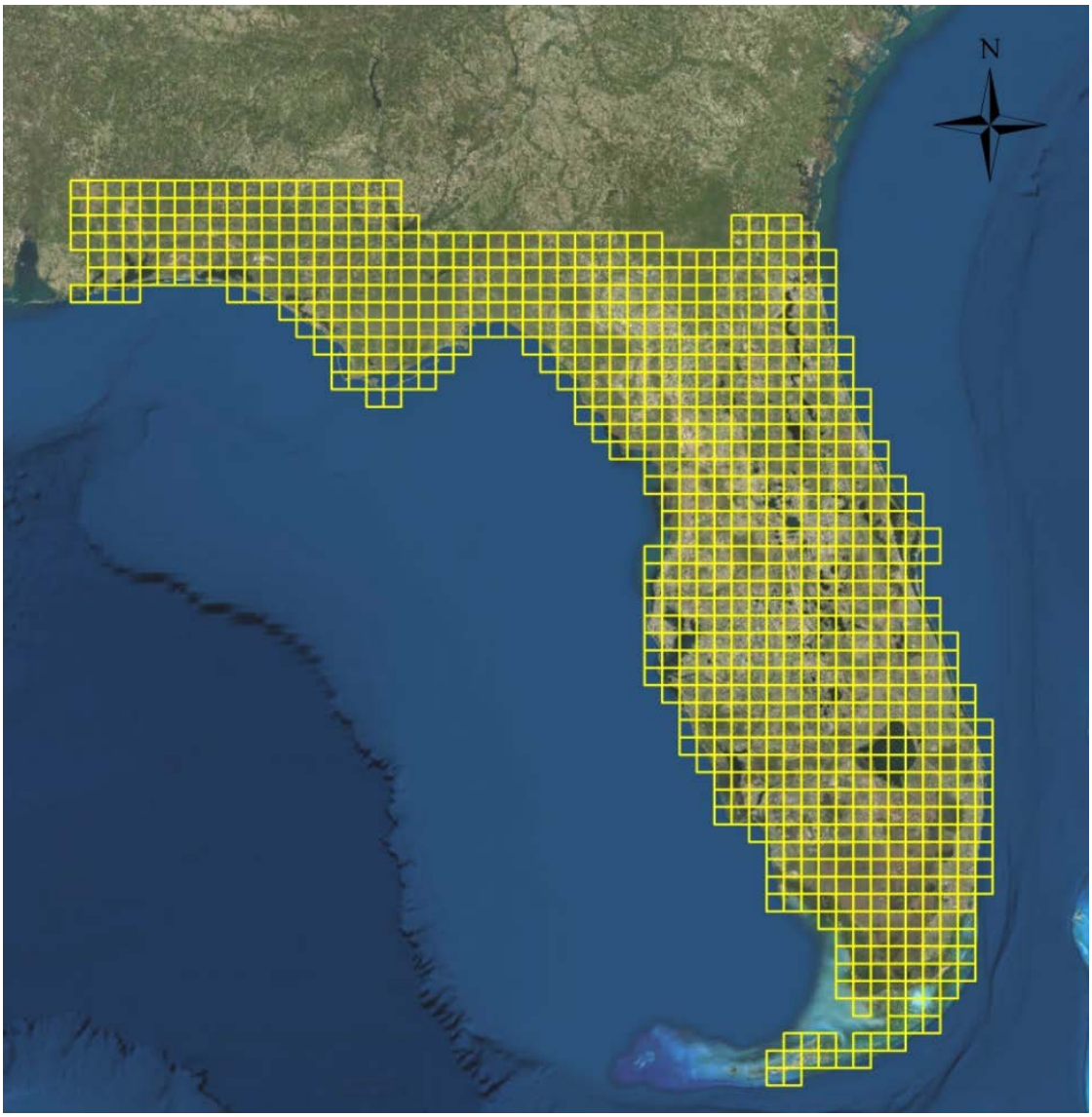


Figure 4-2. Grid structure of Florida (10×10 mile^2)

Table 4-3 Crashes of CTs, TAZs, TADs, and Grids

Geographic units	Average area ($mile^2$)	Number of zones	Total crash				Severe crash				Non-motorized mode crash			
			Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max
CT	15.497	4245	212.305	234.964	0	4554	11.788	11.775	0	141	7.432	7.964	0	76
TAZ	6.472	8518	105.804	142.253	0	1507	5.875	7.944	0	111	3.704	6.084	0	121
TAD	103.314	594	1517.230	1603.290	188	15094	84.241	60.344	4	534	53.109	60.093	1	562
Grid 1×1	1	76640	11.759	61.598	0	2609	0.653	2.614	0	90	0.412	2.484	0	182
Grid 2×2	4	19652	45.860	206.461	0	5321	2.546	8.513	0	271	1.605	7.862	0	209
Grid 3×3	9	8964	100.539	425.753	0	10531	5.582	17.295	0	448	3.519	15.634	0	310
Grid 4×4	16	5124	175.885	712.317	0	16307	9.766	28.997	0	650	6.157	26.161	0	609
Grid 5×5	25	3355	268.624	1084.990	0	25230	14.915	42.962	0	727	9.403	39.150	0	914
Grid 6×6	36	2364	381.233	1459.970	0	24617	21.167	57.821	0	749	13.345	52.004	0	842
Grid 7×7	49	1766	510.326	1889.670	0	29553	28.335	74.121	0	715	17.864	65.854	0	985
Grid 8×8	64	1362	661.700	2465.000	0	41463	36.739	95.446	0	966	23.162	84.708	0	1107
Grid 9×9	81	1094	823.798	2956.390	0	50371	45.739	114.678	0	1218	28.836	103.396	0	1352
Grid 10×10	100	907	993.644	3637.200	0	50989	55.170	141.544	0	1592	34.782	128.862	0	2185

4.6.2 Method to transform predicted crash counts

The method to obtain transformed crash counts of grids is introduced by taking TAZ and Grid 5×5 as an example. As shown in Figure 4-3, the red square is one grid (named as Grid A) which intersects with four TAZ units (named as TAZ 1, 2, 3, and 4). The four corresponding intersected entities are named as Region 1, 2, 3, and 4. It is assumed that the proportion of each region's predicted crash frequency in the TAZ is equal to the corresponding proportion of the same region's observed crash in the same TAZ. Hence, the predicted crash counts for each region can be determined by:

$$y'_{Ri} = y'_{Ti} * P'_{Ri} \tag{4-5}$$

where y'_{Ri} and y'_{Ti} are the predicted crash counts in Region i and TAZ i, P'_{Ri} is the proportion of Region i's observed crash frequency in TAZ i.

Obviously, the crashes that happened in Gird A should be equal to the sum of crashes that happed in the four intersected regions (Region 1, 2, 3, and 4). Then the predicted crash counts of the four TAZs can be transformed into Grid A by adding up the predicted crash counts of all the four intersected regions. Based on this method, the predicted crash counts of models based on CTs, TAZs, and TADs can be transformed into the same grids.

4.6.3 Comparison criteria

Two types of measures, Mean Absolute Error (MAE) and Root Mean Squared Errors (RMSE), were employed to compare the difference between observed crash counts based on grids and six corresponding transformed predicted values. The two measures can be computed by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'| \tag{4-6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} \tag{4-7}$$

where N is the number of observations, y_i and y_i' are the observed and transformed predicted values of crashes for entity i of different levels of grids. The smaller values of the two measures indicate the better performance of estimated models based on CTs, TAZs, and TADs. Also, in order to better compare the measure values across different levels of grids, the weighted MAE and RMSE are computed by dividing MAE and RMSE by the areas of grids.

Figure 4-3. Method to transform predicted crash counts

4.7 Modeling Results

In this study, overall 18 models – 2 model types (PLN and PLN-CAR models), with and without considering spatial correlation based on 3 zonal systems (CTs, TAZs and TADs), were estimated for total, severe and non-motorized crashes. The results of estimated models are displayed in Tables 4-6, separately. Significant variables related to total, severe and non-motorized mode crashes at 95% significant level were analyzed. The Deviance Information Criterion (DIC) and the Moran's I values of residual are also presented in the tables. It is observed that for each zonal system, the spatial models except for non-motorized crashes based on CTs offer substantially better fit compared to the aspatial models. The results remain consistent with the previous comparative analysis results. Also the residual of spatial models of crashes based on TAZs and TADs have weaker spatial correlation except for non-motorized crash based on TAZs, which may be due to the excess zeros. However, for the crashes based on CTs, the Moran's I values of residual have no difference between the aspatial and spatial models. It is known that models with spatially correlated residuals may lead to biased estimation of parameters, which may cause wrong interpretation and conclusion. That could explain that several significant variables in aspatial models become insignificant in the spatial models based on TAZs and TADs while parameters in the aspatial and spatial models vary based on CTs. Moreover, for different crash types, the TAZs and TADs have more significant traffic/roadway related variables compared to CTs. On the contrary, more socio-demographic variables are significant in CTs based models. These are as expected since CTs are designed for socio-demographic characteristics collection while TAZs and TADs are created according to traffic/roadway information.

In addition to the observations, the following subsections present the detailed discussion focused on the PLN-CAR model that offers better fit for total, severe, and non-motorized mode crashes.

4.7.1 Total Crash

Table 4 presents the results of model estimation for total crashes based on CTs, TAZs, and TADs. The VMT variable, as a measure of vehicular exposure, is significant in all models and as expected increases the propensity for total crashes. Besides, the models share a common significant variable length of sidewalk, which consistently has positive effect on crash frequency. The length of sidewalk can be an indication of more pedestrian activity and thus exposure. Additionally, the variable proportion of heavy vehicle in VMT is found to be negatively associated with total crashes in TAZs and TADs based models. On the other hand, the population of the old age group over 65 years old was significant in models based on CTs and TADs. Since the variable is an indication of fewer trips, it is found to have negative relation with crash frequency.

4.7.2 Severe Crash

Modeling results for severe crashes for the three geographic units are summarized in Table 5. The VMT and length of sidewalks are still significant in the three models. Higher median household income results in decreased severe crashes for TAZs and TADs. Also proportion of unemployment and proportion of commuters by public transportation are found significant in CTs and TAZs. Finally, various variables such as proportion of heavy vehicle mileage in VMT,

roadway density, proportion of length of arterials and length of bike lanes are significant solely in the TAZs based model.

4.7.3 Non-motorist Crash

The results of the non-motorized mode crashes are shown in Table 6. The models based on the three geographic units have expected variables such as VMT, proportion of heavy vehicle in VMT, length of local roads, length of sidewalks, population density, commuters by public transportation and cycling. As mentioned above, the VMT, a measure of vehicular exposure, is expected to have positive impact on non-motorized mode crashes frequency. However, the proportion of heavy vehicle VMT has a negative impact since the likelihood of non-motorists drops substantially in the zones with increase in heavy vehicle VMT. The variables proportion of local roads by length and length of sidewalks are reflections of pedestrian access and are likely to increase crash frequency (Cai et al., 2016). The population density is a surrogate measure of non-motorists exposure and is likely to increase the propensity for non-motorized mode crashes. Across the three geographic units, it is observed that the zones with higher proportion of commuters by public transportation and cycling have higher propensity for non-motorized mode crashes. The commuters by public transportation and cycling are indications of zones with higher non-motorists activity resulting in increased non-motorized mode crash risk (Abdel-Aty et al., 2013).

Table 4-4 Total crash model results by zonal systems

Zonal systems	CT				TAZ				TAD			
Variables	PLN		PLN-CAR		PLN		PLN-CAR		PLN		PLN-CAR	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Intercept	1.163	0.026	0.751	0.078	3.35	0.044	1.187	0.057	-1.554	0.023	-0.155	0.689
	(1.119, 1.207)		(0.589, 0.911)		(3.285, 3.409)		(1.066, 1.274)		(-1.591, -1.511)		(-1.674, 1.255)	
Log (VMT)	0.261	0.002	0.271	0.006	0.22	0.013	0.287	0.006	0.655	0.001	0.754	0.024
	(0.257, 0.264)		(0.261, 0.282)		(0.199, 0.240)		(0.275, 0.302)		(0.654, 0.656)		(0.713, 0.800)	
Proportion of heavy vehicle mileage in VMT	-	-	-	-	-2.189	0.29	-1.532	0.355	-2.32	0.322	-4.009	0.457
	-		-		(-2.655, -1.497)		(-2.202, -0.904)		(-2.798, -1.796)		(-4.819, -2.953)	
Log (signalized intersection density)	-	-	-	-	-	-	-	-	0.579	0.056	0.685	0.162
	-		-		-		-		(0.455, 0.682)		(0.203, 0.971)	
Log (length of sidewalks)	0.331	0.007	0.342	0.017	0.495	0.047	0.519	0.022	0.085	0.006	0.082	0.01
	(0.316, 0.345)		(0.297, 0.379)		(0.383, 0.546)		(0.475, 0.573)		(0.075, 0.095)		(0.061, 0.101)	
Log (distance to nearest urban area)	-	-	-	-	-0.513	0.023	-0.181	0.027	-	-	-	-
	-		-		(-0.560, -0.479)		(-0.274, -0.109)		-		-	
Log (population density)	-	-	-	-	-	-	-	-	0.168	0.002	0.083	0.006
	-	-	-	-	-	-	-	-	(0.163, 0.171)		(0.071, 0.097)	
Proportion of population age 15-24	-	-	0.733	0.16	-	-	-	-	-	-	-	-
	-		(0.398, 1.076)		-		-		-		-	
Proportion of population age 65 or older	-1.469	0.056	-1.07	0.087	-1.079	0.206	-0.003	0.001	-	-	-	-
	(-1.560, -1.350)		(-1.234, -0.893)		(-1.354, -0.608)		(-0.006, -0.001)		-		-	
Proportion of unemployment	-	-	-	-	-1.505	0.082	-	-	-	-	-	-
	-		-		(-1.680, -1.380)		-		-		-	
Log (Commuters density)	0.144	0.002	0.167	0.006	-	-	-	-	-	-	-	-
	(0.140, 0.148)		(0.154, 0.180)		-		-		-		-	
Proportion of commuters by public transportation	2.778	0.231	2.486	0.285	2.422	0.413	-	-	5.464	0.312	2.427	0.995
	(2.376, 3.230)		(1.834, 2.996)		(1.929, 3.257)		-		(4.975, 6.146)		(0.432, 4.378)	
Proportion of commuters by walking	1.06	0.231	-	-	-	-	-	-	-	-	-	-
	(0.698, 1.634)		-		-		-		-		-	
Log (median household income)	-	-	-	-	-0.06	0.004	-	-	-0.123	0.002	-0.301	0.063
	-		-		(-0.068, -0.054)		-		(-0.126, -0.123)		(-0.419, -0.160)	
S.D. of θ	0.695	0.003	0.339	0.064	1.033	0.006	0.378	0.04	0.388	0.001	0.136	0.01
	(0.691, 0.702)		(0.241, 0.519)		(1.024, 1.046)		(0.308, 0.467)		(0.385, 0.391)		(0.117, 0.154)	
S.D. of φ	-	-	0.213	0.028	-	-	0.393	0.083	-	-	0.14	0.011
	-		(0.166, 0.275)		-		(0.306, 0.591)		-		(0.118, 0.161)	
DIC	36898.300		36854.800		64441.000		64147.960		6446.200		6435.659	
Moran's *I* of residual*	0.053		0.006		0.460		-0.020		0.412		-0.153	

*All explanatory variables are significant at 95% confidence level; All Moran's *I* values are significant at 95% confidence level

Table 4-5 Severe crash model results by zonal systems

Zonal systems	CT				TAZ				TAD			
Variables	PLN		PLN-CAR		PLN		PLN-CAR		PLN		PLN-CAR	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Intercept	-2.493	0.094	-1.57	0.097	-1.344	0.069	-1.745	0.127	2.137	0.101	2.92	0.749
	(-2.704, -2.376)		(-1.768, -1.379)		(-1.466, -0.217)		(-2.024, -1.466)		(1.971, 2.279)		(1.375, 4.447)	
Log (VMT)	0.402	0.007	0.339	0.009	0.364	0.005	0.33	0.007	0.591	0.01	0.529	0.025
	(0.388, 0.418)		(0.322, 0.357)		(0.354, 0.371)		(0.318, 0.345)		(0.576, 0.606)		(0.476, 0.583)	
Proportion of heavy vehicle mileage in VMT	-	-	-	-	-2.383	0.277	-0.935	0.300	-1.671	0.349	-	-
	-		-		(-2.908, -1.859)		(-1.570, -0.312)		(-2.391, -1.098)		-	
Log (roadway density)	-	-	-	-	-0.024	0.011	-0.108	0.016	-	-	-	-
	-		-		(-0.050, -0.003)		(-0.140, -0.076)		-		-	
Proportion of length of arterials	-	-	-	-	-0.604	0.044	-0.591	0.045	-	-	-	-
	-		-		(-0.686, -0.518)		(-0.678, -0.502)		-		-	
Proportion of length of collectors	-	-	-0.283	0.083	-	-	-	-	-	-	-	-
	-		(-0.452, -0.123)		-		-		-		-	
Proportion of length of local roads	0.263	0.043	-	-	-	-	-	-	0.851	0.076	-	-
	(0.184, 0.352)		-		-		-		(0.701, 0.989)		-	
Log (length of bike lanes)	-	-	-	-	0.082	0.028	0.113	0.028	-	-	-	-
	-		-		(0.026, 0.134)		(0.061, 0.166)		-		-	
Log (length of sidewalks)	0.183	0.016	0.238	0.018	0.245	0.024	0.354	0.021	0.116	0.02	0.104	0.018
	(0.154, 0.214)		(0.203, 0.273)		(0.187, 0.282)		(0.313, 0.393)		(0.084, 0.151)		(0.068, 0.141)	
Log (distance to nearest urban area)	-	-	0.201	0.018	-	-	-	-	-	-	-	-
	-		(0.168, 0.238)		-		-		-		-	
Proportion of unemployment	-0.222	0.07	-0.444	0.081	-0.766	0.079	-0.152	0.089	-	-	-	-
	(-0.343, -0.063)		(-0.605, -0.278)		(-0.935, -0.614)		(-0.330, 0.032)		-		-	
Proportion of commuters by public transportation	1.423	0.268	1.554	0.269	1.724	0.256	1.015	0.33	-	-	-	-
	(0.862, 1.934)		(1.032, 2.048)		(1.244, 2.206)		(0.423, 1.670)		-		-	
Proportion of commuters by walking	0.976	0.273	-	-	-	-	-	-	-	-	-	-
	(0.450, 1.525)		-		-		-		-		-	
Log (median household income)	-	-	-	-	-0.037	0.003	-0.021	0.009	-0.589	0.007	-0.536	0.062
	-		-		(-0.043, -0.030)		(-0.039, -0.004)		(-0.604, -0.576)		(-0.659, -0.412)	
S.D. of θ	0.614	0.007	0.218	0.049	0.835	0.008	0.393	0.045	0.458	0.006	0.116	0.006
	(0.601, 0.628)		(0.166, 0.329)		(0.819, 0.852)		(0.304, 0.470)		(0.447, 0.469)		(0.107, 0.129)	
S.D. of φ	-	-	0.191	0.025	-	-	0.519	0.024	-	-	0.152	0.02
	-		(0.148, 0.247)		-		(0.278, 0.749)		-		(0.123, 0.199)	
DIC	23958.000		23835.000		38158.200		37470.090		4741.080		4696.724	
Moran's I of residual	0.065		-0.007		0.397		0.040		0.370		-0.096	

*All explanatory variables are significant at 95% confidence level; * All Moran's *I* values are significant at 95% confidence level

Table 4-6 Non-motorized mode crash model results by zonal systems

Zonal systems	CT				TAZ				TAD			
Variables	PLN		PLN-CAR		PLN		PLN-CAR		PLN		PLN-CAR	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Intercept	-2.539	0.062	-2.256	0.129	-3.612	0.157	-3.503	0.144	0.176	0.063	4.737	1.221
	(-2.664, -2.388)		(-2.510, -1.996)		(-3.812, -3.301)		(-3.800, -3.200)		(0.069, 0.285)		(2.412, 7.038)	
Log (VMT)	0.172	0.007	0.161	0.008	0.297	0.005	0.283	0.007	0.345	0.004	0.252	0.038
	(0.161, 0.186)		(0.145, 0.177)		(0.289, 0.307)		(0.268, 0.298)		(0.336, 0.352)		(0.179, 0.331)	
Proportion of heavy vehicle mileage in VMT	-1.858	0.330	-2.262	0.389	-4.389	0.432	-4.803	0.391	-3.639	0.440	-2.969	0.854
	(-2.459, -1.134)		(-3.053, -1.478)		(-5.083, -3.520)		(-5.518, -4.068)		(-4.548, -2.884)		(-4.519.-1.511)	
Log (roadway density)	-	-	-	-	0.154	0.016	0.143	0.020	-	-	-	-
	-		-		(0.128, 0.189)		(0.106, 0.182)		-		-	
Proportion of length of local roads	0.377	0.043	0.367	0.061	0.717	0.044	0.752	0.047	0.679	0.101	-	-
	(0.279, 0.453)		(0.245, 0.488)		(0.623, 0.794)		(0.661, 0.845)		(0.517, 0.838)		-	
Log (length of sidewalks)	0.48	0.017	0.488	0.019	0.506	0.022	0.558	0.022	0.283	0.015	0.306	0.027
	(0.450, 0.516)		(0.454, 0.524)		(0.458, 0.545)		(0.516, 0.602)		(0.257, 0.315)		(0.252, 0.360)	
Log (population density)	0.243	0.005	0.225	0.010	0.234	0.006	0.175	0.010	0.22	0.009	0.165	0.024
	(0.234, 0.252)		(0.206, 0.247)		(0.225, 0.246)		(0.158, 0.192)		(0.205, 0.237)		(0.125, 0.215)	
Proportion of population age 65 or older	-0.691	0.098	-0.761	0.094	-	-	-	-	-	-	-	-
	(-0.890, -0.519)		(-0.947, -0.582)		-		-		-		-	
Log (Commuters density)	-	-	-	-	-0.635	0.075	-0.398	0.099	-	-	-	-
	-		-		(-0.766, -0.450)		(-0.587, -0.199)		-		-	
Proportion of commuters by public transportation	3.532	0.260	3.565	0.292	3.467	0.258	2.949	0.282	7.525	0.606	4.802	1.286
	(3.011, 4.049)		(3.011, 4.102)		(2.919, 3.974)		(2.375, 3.457)		(6.544, 8.900)		(2.676, 7.015)	
Proportion of commuters by cycling	3.955	0.492	3.892	0.441	1.078	0.471	-	-	7.000	1.703	8.566	2.258
	(2.901, 4.918)		(3.069, 4.792)		(0.076, 1.960)		-		(4.180, 10.670)		(3.955, 12.758)	
Proportion of commuters by walking	2.476	0.329	2.595	0.306	1.877	0.280	1.757	0.294	-	-	-	-
	(1.874, 3.116)		(1.998, 3.145)		(1.321, 2.405)		(1.189, 2.325)		-		-	
Log (median household income)	-	-	-	-	-0.075	0.014	-0.047	0.01	-0.336	0.005	-0.565	0.094
	-		-		(-0.098, -0.056)		(-0.066, -0.026)		(-0.344, 0.326)		(-0.745, -0.384)	
S.D. of θ	0.605	0.009	0.361	0.090	0.790	0.011	0.518	0.144	0.456	0.008	0.222	0.023
	(0.588, 0.622)		(0.196, 0.531)		(0.769, 0.814)		(0.224, 0.715)		(0.440, 0.472)		(0.181, 0.263)	
S.D. of φ	-	-	0.053	0.008	-	-	0.037	0.058	-	-	0.198	0.028
	-		(0.042, 0.072)		-		(0.010, 0.152)		-		(0.147, 0.261)	
DIC	21032.300		21033.730		30244.700		29926.930		4317.540		4302.187	
Moran's I of residual	0.028		0.021		0.286		0.325		0.092		-0.088	

*All explanatory variables are significant at 95% confidence level; * All Moran's *I* values are significant at 95% confidence level

4.8 Comparative Analysis Results

Based on the estimated models of the three zonal systems, the predicted crash counts for each crash type of the three geographic units can be computed and then transformed into the correspondingly intersected grids. Weighted MAE and RMSE for each grid structure were calculated with the observed crash counts and transformed predicted crash counts based on different geographic units. The comparison results are as shown in Table 3-7 and several observations can be made. (1) The MAE and RMSE values consistently increase with the grid size, validating the previous discussion that the comparison measures can be influenced by the number of observations and observed values. (2) For each zonal system, the spatial (PLN-CAR) models substantially improve the performance over the aspatial (PLN) models for predicting crash counts. The results are consistent with the previous analysis results that the crash counts are spatially correlated and the model considering the spatial dependency can provide better understanding of crash frequency. Also, the improvements based on TAZs and TADs are much greater than that based on CTs which should be related to the spatial correlation levels. (3) Among aspatial and spatial models, the TADs always have the best performance indicating the advantages of TADs over the other two zonal systems. Meanwhile, CTs based on aspatial models can consistently perform better than the models based on TAZs. However, the exact ordering alters between spatial models based on CTs and TAZs according to MAE and RMSE.

The CTs are designed to be comparatively homogenous units with respect to socio-demographic statistical data. Thus, it is not surprising that CT-based models do not show the best performance. TAZs are the base zonal system of analyses for developing travel demand models and have been widely used by metropolitan planning organizations for their long-range transportation plans.

However, one of the major zoning criteria for TAZs is to minimize the number of intra-zonal trips (Meyer & Miller, 2001) which results in small area size for each TAZ. Due to the small size, a crash occurring in a TAZ might be caused by the driver from another TAZ, i.e., the characteristics of drivers who cause the crashes cannot be observed by the models based on TAZs. Also, as TAZs are often delineated by arterial roads and many crashes occur on these boundaries. The existence of boundary crashes may invalidate the assumptions of modeling only based on the characteristics of a zone where the crash is spatially located (Lee et al, 2014; Siddiqui et al., 2012). Hence, although TAZs are appropriate for transportation demand forecasting, they might be not the best option for the transportation safety planning. The TADs are another transportation-related zonal system with considerably larger size compared with TAZs. There should be more intra-zonal trips in each TAD and the drivers who cause crashes in a TAD will be more likely to come from the same TAD. Therefore, it seems reasonable that TADs are superior for macro-level crash analysis and transportation safety planning.

In summary, considering the rationale for the development of different zonal systems and the modeling results in our study, it is recommended using CTs for socio-demographic data collection, employing TAZs for transportation demand forecasting, and adopting TADs for transportation safety planning.

Table 4-7 Comparison results based on grids

	Total Crashes						Severe Crashes						Non-motorized Crashes					
	PLN			PLN_CAR			PLN			PLN_CAR			PLN			PLN_CAR		
	CT	TAZ	TAD	CT	TAZ	TAD	CT	TAZ	TAD	CT	TAZ	TAD	CT	TAZ	TAD	CT	TAZ	TAD
Weighted MAE																		
Grid 1×1	4.70	6.12	3.43	4.45	3.34	2.30	0.28	0.33	0.22	0.26	0.23	0.18	0.17	0.19	0.15	0.17	0.18	0.12
Grid 2×2	4.22	5.61	3.25	3.95	2.62	2.03	0.25	0.30	0.21	0.23	0.19	0.15	0.14	0.17	0.14	0.14	0.16	0.11
Grid 3×3	3.87	5.23	3.10	3.59	2.19	1.85	0.23	0.28	0.20	0.21	0.17	0.14	0.13	0.16	0.13	0.13	0.15	0.10
Grid 4×4	3.63	4.97	3.01	3.36	1.93	1.61	0.21	0.26	0.20	0.19	0.15	0.12	0.12	0.15	0.12	0.12	0.14	0.09
Grid 5×5	3.42	4.74	2.79	3.16	1.81	1.39	0.20	0.25	0.19	0.18	0.14	0.10	0.11	0.14	0.11	0.11	0.13	0.08
Grid 6×6	3.30	4.57	2.72	3.03	1.65	1.20	0.19	0.24	0.19	0.17	0.14	0.10	0.10	0.14	0.10	0.10	0.12	0.07
Grid 7×7	3.18	4.43	2.68	2.94	1.55	1.17	0.18	0.23	0.18	0.17	0.13	0.09	0.10	0.13	0.10	0.10	0.12	0.07
Grid 8×8	3.06	4.31	2.58	2.82	1.49	1.08	0.18	0.23	0.17	0.16	0.13	0.08	0.09	0.13	0.09	0.09	0.11	0.06
Grid 9×9	2.99	4.23	2.53	2.74	1.47	0.94	0.17	0.22	0.17	0.15	0.12	0.07	0.09	0.13	0.09	0.09	0.11	0.06
Grid 10×10	2.84	4.08	2.41	2.60	1.38	0.94	0.16	0.21	0.17	0.15	0.12	0.07	0.09	0.12	0.08	0.09	0.11	0.05
AVE	3.52	4.83	2.85	3.26	1.94	1.45	0.21	0.25	0.19	0.19	0.15	0.11	0.11	0.15	0.11	0.11	0.13	0.08
Weighted RMSE																		
Grid 1×1	31.84	39.77	27.82	29.41	20.54	19.56	1.40	1.66	1.31	1.35	1.07	1.11	1.12	1.37	1.49	1.11	1.22	1.33
Grid 2×2	25.54	32.53	22.64	23.27	12.60	14.61	1.07	1.30	1.02	1.03	0.73	0.74	0.77	0.96	1.00	0.76	0.85	0.87
Grid 3×3	22.38	28.99	18.89	20.19	9.31	11.23	0.91	1.13	0.88	0.87	0.57	0.67	0.62	0.79	0.81	0.62	0.70	0.61
Grid 4×4	20.30	26.18	16.78	18.16	7.68	7.65	0.83	1.04	0.80	0.79	0.51	0.55	0.54	0.72	0.59	0.54	0.64	0.46
Grid 5×5	19.53	25.41	16.06	17.54	6.53	7.28	0.73	0.95	0.70	0.70	0.44	0.34	0.48	0.66	0.57	0.48	0.57	0.43
Grid 6×6	18.30	23.92	15.10	16.34	5.50	5.25	0.66	0.86	0.65	0.61	0.39	0.31	0.44	0.60	0.48	0.43	0.52	0.35
Grid 7×7	17.43	22.58	14.72	15.46	4.81	5.51	0.58	0.79	0.59	0.55	0.34	0.25	0.39	0.54	0.40	0.39	0.46	0.27
Grid 8×8	17.43	22.65	14.24	15.41	4.68	4.86	0.59	0.79	0.58	0.55	0.35	0.24	0.36	0.52	0.38	0.36	0.44	0.25
Grid 9×9	16.10	21.23	12.85	14.23	4.35	3.56	0.53	0.73	0.54	0.50	0.32	0.22	0.35	0.51	0.35	0.35	0.43	0.21
Grid 10×10	15.45	21.18	12.79	13.71	3.89	4.03	0.49	0.71	0.49	0.47	0.31	0.17	0.32	0.50	0.31	0.32	0.40	0.18
AVE	20.43	26.44	17.19	18.37	7.99	8.35	0.78	0.99	0.76	0.74	0.50	0.46	0.54	0.72	0.64	0.54	0.62	0.50

4.9 Summary and Conclusion

Macro-level safety modeling is one of the important objectives in transportation safety planning. Although various geographic units have been employed for macro-level crash analysis, there has been no guidance to choose an appropriate zonal system. One of difficulties is to compare models based on different geographic units of which number of zones is not the same. This study proposes a new method for the comparison between different zonal systems by adopting grid structures of different scales. The Poisson lognormal (PLN) models without and Poisson lognormal conditional autoregressive model (PLN-CAR) with consideration of spatial correlation for total, severe, and non-motorized mode crashes were developed based on census tracts (CTs), traffic analysis zones (TAZs), and a newly developed traffic-related zone system - traffic analysis districts (TADs). Based on the estimated models, predicted crash counts for the three zonal systems were computed. Considering the average area of each geographic unit, ten sizes of grid structures with dimensions ranging from 1 mile to 100 square miles were created for the comparison of estimated models. The observed crash counts for each grid were directly obtained with GIS while the different predicted crash counts were transformed into the grids that each geographic unit intersects with. The weighted MAE and RMSE were calculated for the observed and different transformed crash counts of different grid structures. By comparing the MAE and RMSE values, the best zonal system as well as model for macroscopic crash modeling can be identified with the same sample size.

The comparison results indicated that the models based on TADs offered the best fit for all crash types. Based on the modeling results and the motivation for developing the different zonal systems, it is recommended CTs for socio-demographic data collection, TAZs for transportation

demand forecasting, and TADs for transportation safety planning. Also, the comparison results highlighted that models with the consideration of spatial effects consistently performed better than the models that did not consider the spatial effects. The modeling results based on different zonal systems had different significant variables, which demonstrated the zonal variation. Besides, the results clearly highlighted the importance of several explanatory variables such as traffic (i.e., VMT and heavy vehicle mileage), roadway (e.g., proportion of local roads in length, signalized intersection density, and length of sidewalks, etc.) and socio-demographic characteristics (e.g., population density, commuters by public transportation, walking as well as cycling, median household income, etc.).

This study focuses on the comparison of zonal systems for crash modeling and transportation safety planning. However, only three zonal systems were adopted for the validation of the proposed comparison method. Extending the current approach to compare other zonal systems (e.g., census block and counties) could be meaningful. Also, it is possible that the trip distance might be related to the size of appropriate geographic units for crash modeling. Future research extension might consider such relationship.

CHAPTER 5: JOINT APPROACH OF FREQUENCY AND PROPORTION MODELING AT MACRO-LEVEL

5.1 Introduction

With the growing concern of global warming and increasing obesity among adults and children, walking and bicycling are highly promoted by many communities. However, pedestrians and bicyclists are more vulnerable than automobile occupants and transportation safety has become a big concern for people to choose walking or bicycling. According to the National Highway Traffic Safety Administration, in 2015, totally about 6,100 pedestrians and bicyclists (i.e., non-motorists) were killed from traffic crashes which accounted for nearly 18% of all traffic fatalities in the United States (NHSTA, 2015). In order to encourage people to walk and bicycle, it is necessary to put considerable efforts to enhance road safety for pedestrians and bicyclists. An efficient approach is the application of macroscopic crash modeling, which can investigate the effects of zonal factors on non-motorist safety (Wei & Lovegrove, 2012; FMCSA, 2012) and identify hot (unsafe) zones which have safety concerns as impediments for people to adopt walking or bicycling as a preferred transportation mode to private vehicles. By understanding the impact of zonal factors on pedestrian and bicyclist safety, planning-level strategies could be proposed to proactively improve traffic safety.

This study aims to enhance pedestrians and bicyclists' safety by suggesting a joint model to examine non-motorist crashes at the macroscopic level. More specifically, this study investigates the impact of macro-level characteristics on non-motorist crashes (i.e., crashes between vehicles and non-motorists (pedestrians or bicyclists)). It was found that the crashes between vehicles and

pedestrians and crashes between vehicles and bicyclist were highly correlated and shared a vast of significant variables with the same impacts (Eluru et al., 2008; Siddiqui et al., 2012; Kaplan and Prato, 2013; Lee et al., 2015a; Zhang et al., 2015; Cai et al., 2016; Nashad et al., 2016) and it should be reasonable to combine the two types of crashes for the analysis. Several exogenous variables including traffic flow characteristics, transportation network characteristics, socio-demographic characteristics, and commuting variables are considered in the model development. The suggested model development would allow us to identify important determinants of non-motorist crashes, and also provide valuable insights on the appropriate model framework for the macro-level non-motorist crash analysis.

5.2 Statistical Methodology

5.2.1 Standard Count Model

The negative binomial (NB) model has been widely used in previous crash count studies (Lord and Mannering, 2010). The model assumes that the crash data follows a Poisson-gamma mixture which can address the over-dispersion issue (i.e., variance exceeds the mean). A NB model is specified as follows:

$$y_i^{NON} \sim Poisson\ (u_i^{NON}) \tag{5-1}$$

$$\log(u_i^{NON}) = \beta^{NON} * x^{NON} + \varepsilon_i \tag{5-2}$$

where y_i^{NON} is the number of non-motorist crashes in zone i, u_i^{NON} is the expectation of y_i^{NON}, x^{NON} is a set of explanatory variables, β^{NON} is the corresponding parameter, θ_i is the error term. The $\exp(\varepsilon_{i_i})$ is a gamma distributed error term with mean 1 and variance α^2. Based on the NB model, the variables having significant effects on the non-motorist crash counts can be identified.

However, it is not clear whether the significant variables contribute to the vehicle drivers (more total crashes) or just non-motorists (higher proportion of non-motorist crashes).

5.2.2 Joint Model

The non-motorist crashes are the result of collisions between vehicles and non-motorists. The zonal factors can affect the non-motorist crashes through either drivers, or non-motorists, or both. The zones with high non-motorist crash risk may be because of more dangerous vehicle drivers or driving environment and incautious non-motorist or hazardous walking and bicycling conditions. The total crash count can reflect drivers' behavior and the driving environment since all crashes should be vehicle related and most of crashes are among vehicles. Meanwhile, the proportion of non-motorist crashes can indicate the transportation safety level for non-motorists in each zone. Specifically, it would be dangerous to walk or bicycle instead of other transportation modes in zones with high proportions of non-motorist crashes. Thus, in the joint model we convert the non-motorist crash count into the product of the total crash count (representing vehicle drivers) multiplied by the proportion of non-motorist crashes. As for the total crash counts, a log link between the dependent and explanatory variables is specified in the modeling regression. Meanwhile, a logit transformation is applied for the proportion of non-motorist crashes to restrict the dependent variable between 0 and 1. Thus, the specific structure of the joint model for non-motorist crashes can be expressed as follows:

$$\log(u_i^{NON}) = \log(u_i^{TOT} * p_i^{NON}) + \varepsilon_i \quad (5\text{-}3)$$

$$y_i^{TOT} \sim Poisson\ (u_i^{TOT}) \quad (5\text{-}4)$$

where u_i^{TOT} is the expected total crash counts y_i^{TOT} in zone i, p_i^{NON} is the expected proportions of non-motorist crashes in zone i. To keep the same structure as the NB model, an error term with the same distribution in the NB model is also used in the equation. The expected total crash counts and proportion of non-motorist crashes can be estimated by:

$$\log(u_i^{TOT}) = \beta^{TOT} * x^{TOT} + \theta_i \quad (5\text{-}5)$$

$$logit(P_i^{NON}) = \beta^{P\_NON} * x^{P\_NON} + \varphi_i \quad (5\text{-}6)$$

where x^{TOT} and x^{P_NON} denote the explanatory variables for total crash counts and proportion of non-motorist crashes. β^{TOT} and β^{P_NON} represent the corresponding regression coefficients. θ_i and φ_i are random error terms representing normal heterogeneity of total crash count and proportion of non-motorists crashes.

5.3 Data Preparation

Data from 594 Traffic Analysis Districts (TADs) in Florida (see Figure 1) were used for the analysis. The TADs are newly developed transportation-related geographic units by combining Traffic Analysis Zones (TAZs) (FHWA, 2011). TAZs have been widely employed in many macro-level traffic safety studies. However, TAZs are often delineated by arterial roads and thus many crashes occur on these boundaries. The existence of boundary crashes may invalidate the assumptions of modeling only based on the characteristics of a zone where the crash is spatially located (Lee, 2014; Lee et al, 2014; Siddiqui et al., 2012). Also, the size of a TAZ is small and thus a driver who causes a crash in a TAZ is likely to come from other TAZs. It means the characteristics of the driver may not be considered in the TAZ-based models. In Florida, the average area of TADs (103.3 mi2) is considerably larger than that of TAZs (6.5 mi2). Therefore,

it is deduced that there should be more intra-zonal trips in each TAD and the drivers who cause crashes in a TAD would be more likely to come from the same TAD. Therefore, it is reasonable to use TADs for macro-level crash analysis (Abdel Aty et al., 2016; Cai et al., 2017). The crashes that occurred in Florida during 2010-2012 were collected from the Crash Analysis Reporting System (CARS) database of the Florida Department of Transportation. A total of 901,235 crashes were recorded in Florida among which 31,547 (3.5%) were non-motorist crashes. Given the large number of crashes in the Florida data and the sufficiently large TAD area, no zero count units exist for the time period of our analysis.

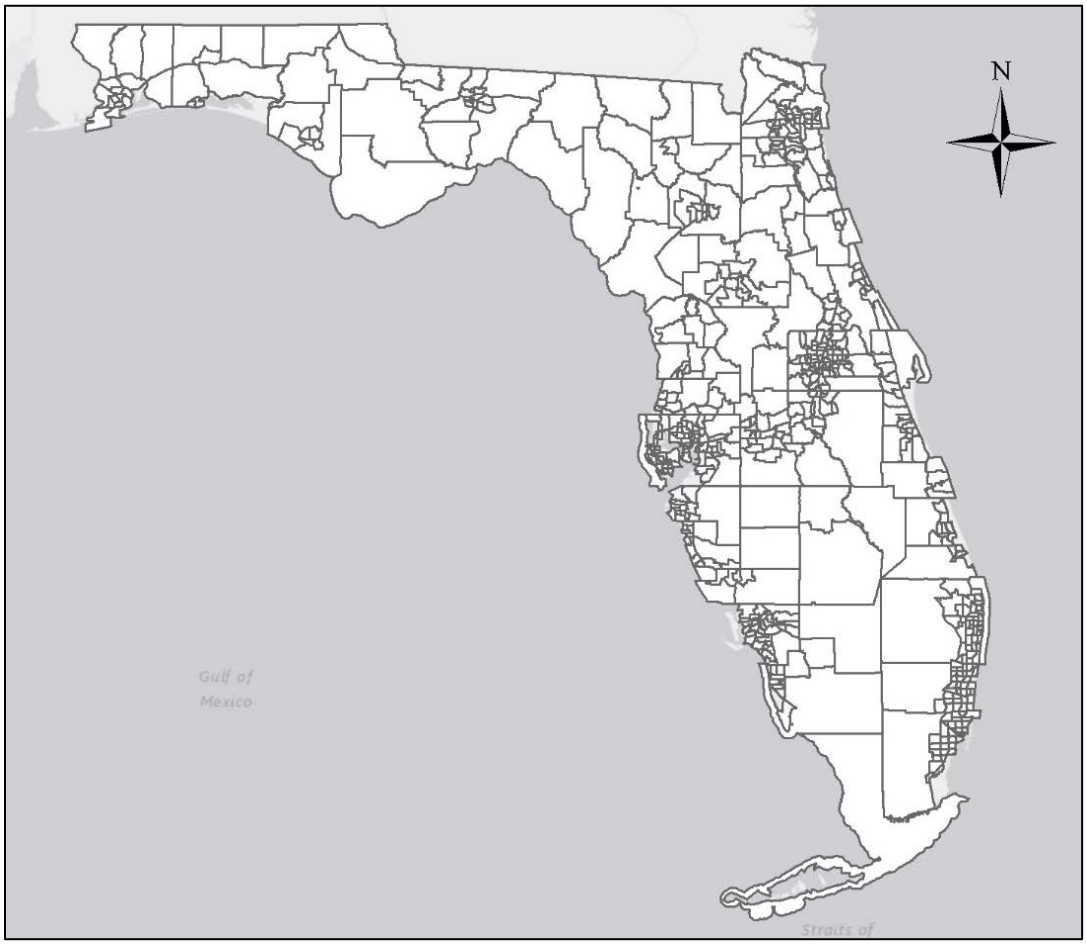


Figure 5-1. Illustration of TADs in Florida

A host of explanatory variables are considered for the analysis and are grouped into four categories: traffic exposure (i.e., Vehicle-Miles-Traveled (VMT), proportion of heavy vehicle in VMT), roadway information (e.g., proportion of length of freeway, signalized intersection density, length of bike lanes, length of sidewalks, etc.), socio-demographic characteristics (e.g., distance to nearest urban area, population density, median household income, proportion of unemployment etc.), and commuting variables (e.g., total commuters density, proportion of commuters by public transportation, etc.). All the candidate variables have been widely investigated in the previous studies (Lovegrove and Sayed, 2006; Siddiqui et al., 2012; Lee et al., 2015; Cai et al., 2017). It should be noted that the road density is defined as total roadway length per square mile which can be computed by dividing the total roadway length by the area of each TAD. The intersection density is the number of intersection divided by the length of total road length. The length of bike lanes and sidewalks is obtained from Florida Department of Transportation (FDOT) Roadway Characteristics Inventory (RCI). The bike lanes and sidewalks can be one-way or two-way. If bike lanes or sidewalks are present in both directions, the length would be added. Furthermore, the distance to the nearest urban area is defined as the distance from the centroid of the TADs to the nearest urban region. Thus, the distance would be zero if the zone is located in an urban area. The descriptive statistics of the crash counts and candidate explanatory variables are summarized in Table 1.

Table 5-1. Descriptive statistics of the collected data (N=594)

Variables	Mean	S.D.	Min.	Max.
Crash variables				
Non-motorist crash frequency	53	60	1	562
Total crash frequency	1,517	1,603	188	15,090
Proportion of non-motorist crashes	0.048	0.021	0.002	0.138
Traffic and roadway variables				
VMT (vehicle*mile)	599,647	428,747	38,547	4,632,469
Proportion of heavy vehicle in VMT	0.071	0.039	0.015	0.290
Road density (mile per mile^2)	7.613	5.311	0.074	24.560
Proportion of length of freeways	0.022	0.032	0	0.317
Proportion of length of arterials	0.111	0.060	0	0.478
Proportion of length of collectors	0.112	0.066	0	0.603
Proportion of length of local roads	0.755	0.108	0.077	0.935
Signalized intersection density (number of signalized intersection per mile)	0.121	0.126	0	1.363
Length of bike lanes (mile)	4.384	6.743	0	65.300
Length of sidewalks (mile)	12.930	11.937	0	87.180
Socio-demographic variables				
Distance to the nearest urban area (mile)	1.313	3.847	0	31.500
Population density (number of people per mile^2)	1,998.610	1,969.808	6.680	15,341.300
Proportion of population aged 15-24	0.135	0.058	0.034	0.694
Proportion of population aged 65 or over	0.167	0.089	0.032	0.660
Total employment density (number of total employment per mile^2)	1,617.080	1,609.586	6.840	13,007.100
Median household income (dollars)	59,986	17,748	21,637	131,664
Commuting variables				
Total commuters density (number of total commuters per mile^2)	900.670	904.087	3.601	6,936.093
Proportion of commuters by car	0.900	0.046	0.544	0.969
Proportion of commuters by public transportation	0.017	0.026	0	0.196
Proportion of commuters by cycling	0.061	0.010	0	0.168
Proportion of commuters by walking	0.014	0.015	0	0.142

5.4 Modeling Results

WinBUGS was used to estimate the NB model and the proposed joint model. Before the estimation of models, the correlation tests for the independent variables are conducted. To avoid the adverse impact of significant correlation, the variables with high correlation were not employed in the model at the same time. The significant independent variables were determined based on 95% certainty of Bayesian credible intervals (BCIs). Deviance information criterion (DIC) was computed to determine the best set of parameters for each model. Besides, the DIC was also employed to compare the two models. Models with smaller DIC value are preferred. Roughly, differences over 10 might indicate the model with lower DIC is significantly better (El-Basyouny and Sayed, 2009).

Tables 2 and 3 show the modeling results of the NB model and proposed joint model, respectively. It was revealed that the joint model has lower DIC value and the difference is more than 120, indicating that the proposed model offers significantly better performance over the NB model. The result of the NB model only has the count frequency component for non-motorist crashes. On the other hand, the joint model consists of two components: 1) count frequency model for total crashes; 2) logit model for the proportion of non-motorist crashes. Thus, it is as expected more different variables (e.g., signalized intersection density and proportion of population aged 65 or over) are significant in the proposed model compared with the NB model. Meanwhile, all significant variables in the NB model can also be found significant in the joint model which clearly indicates that these variables have effects on either vehicle drivers (total crash part) or non-motorists (proportion of non-motorist crash part). While the results for the two

models are presented in Tables 2 and 3, the following discussion about parameters focuses on the joint model which has better fit and more significant variables.

Table 5-2 NB model results

Variable	NB model			
	Mean	S.D.	BCI	
			2.5%	97.5%
Intercept	0.520	0.001	0.517	0.522
Traffic characteristics				
Log(VMT)	0.332	0.001	0.330	0.334
Proportion of heavy vehicle mileage in VMT	-5.036	0.001	-5.038	-5.034
Roadway characteristics				
Proportion of length of local road	0.524	0.001	0.522	0.525
Log(length of sidewalks)	0.320	0.001	0.318	0.322
Socio-demographic characteristics				
Log(population density)	0.151	0.001	0.149	0.153
Log(median household income)	-0.288	0.001	-0.29	-0.287
Commuting characteristics				
Proportion of commuters by public transportation	8.38	0.001	8.378	8.382
Proportion of commuters by bicycle	8.973	0.001	8.971	8.975
Over-dispersion parameter	3.939	0.221	3.505	4.386
DIC	4327.320			

Table 5-3 Joint model results

Variable	Joint model			
	Mean	S.D.	BCI	
			2.5%	97.5%
Count model part				
Intercept	-1.544	0.001	-1.546	-1.542
Traffic characteristics				
Log(VMT)	0.654	0.001	0.652	0.655
Proportion of heavy vehicle mileage in VMT	-2.483	0.001	-2.485	-2.481
Roadway characteristics				
Log(signalized intersection density)	0.508	0.001	0.506	0.510
Log(length of sidewalks)	0.115	0.001	0.113	0.117
Socio-demographic characteristics				
Log(population density)	0.158	0.001	0.156	0.160
Log(median household income)	-0.116	0.001	-0.117	-0.114
Commuting characteristics				
Proportion of commuters by public transportation	6.010	0.001	6.008	6.012
Proportion model part				
Intercept	1.595	0.001	1.593	1.596
Traffic characteristics				
Log(VMT)	-0.349	0.001	-0.352	-0.348
Roadway characteristics				
Proportion of length of local road	0.541	0.001	0.539	0.543
Log(signalized intersection density)	0.761	0.001	0.759	0.763
Log(length of sidewalks)	0.116	0.001	0.114	0.118
Socio-demographic characteristics				
Proportion of population aged 65 or over	0.873	0.001	0.871	0.875
Log(median household income)	-0.114	0.001	-0.116	-0.112
Commuting characteristics				
Proportion of commuters by bicycle	5.568	0.001	5.566	5.570
Over-dispersion parameter	5.291	0.554	4.292	6.425
S.D. of θ_i	7.838	0.690	6.571	9.351
S.D. of φ_i	5.048	0.510	4.157	6.133
DIC	4206.800			

5.4.1 Count Model Part

Overall seven independent variables were found to have significant effects on vehicle drivers in the count model part. The variable VMT is a measure of vehicular exposure and the number of total crashes including non-motorist crashes increases as the VMT increases. The variable proportion of heavy vehicle mileage in VMT has a negative effect. A high proportion of heavy vehicle mileage might indicate the areas where the traffic exposure is comparatively lower and drivers are likely driving more carefully. In terms of roadway characteristics, the signalized intersection density and the length of sidewalks are significant in the count model part. The increase in the two variables could increase the crash risk and indicate more conflicts. Also, improper driving decision due to the dilemma zones can lead to more crashes at the signalized intersections (Wu et al., 2014). It should be noted that the variable signalized intersection density is not significant in the NB model, which may be due to the correlation effects with other variables. The socio-demographic characteristics exhibit significant influences on crashes. Population density could be considered as a surrogate measure of traffic and thus it has a positive impact. As an indication of economic deprivation status, the higher median household income can improve the roadway condition for travelers and thus reduce the crashes. Also, it might be difficult for people from deprived areas to obtain enough information about traffic safety (Martinez and Veloz, 1996). Furthermore, the proportion of commuters by public transportation is found to have a positive effect in the count model. A possible explanation is that the area with higher proportion of commuters by public transportation should have more bus stops where vehicles may have conflicts with buses.

5.4.2 Proportion Model Part

There were seven explanatory variables which had significant impacts on pedestrians and bicyclists in the proportion model part. Although it was found that the VMT has positive effect in the frequency model part, the increased VMT would result in the decrease of non-motorists and the proportion of non-motorist crashes. Three roadway variables including proportion of length of local roads, signalized intersection density, and length of sidewalks are positively related to the proportion of non-motorist crashes. Zones with increased local roads, signalized intersections, and sidewalks may attract more pedestrians and bicyclists and are likely to increase the conflicts between vehicles and non-motorists. Also, more interaction between vehicles and non-motorists exist at intersections with signal controls and hence more crashes are prone to occur. Moreover, the variable proportion of population equal to or older than 65 years old has a positive effect on the proportion of non-motorist crashes. The result seems reasonable since older people are more likely to walk. However, it would be difficult for old pedestrians and bicyclists to across the road, increasing the probability to be hit by vehicles. The median household income is found to be negatively associated with the proportion of non-motorist crashes. It might be because the people from households with lower economic status tend to walk or ride bicycles rather than driving. Furthermore, in zones with increased proportion of commuters by bicycle, the exposure of bicycling increases and hence the proportion of non-motorists crashes increases.

5.5 Elasticity Effects

The parameters of the exogenous variables in Table 3 do not directly provide the magnitude of the effects on the macro-level non-motorists crash frequency. Thus, we compute the elasticity

effects of exogenous variables for both the standard NB model and the proposed joint model. The elasticity effects are calculated by evaluating the change in non-motorist crash frequency in response to increasing the value of each exogenous variable by 10% (see Eluru and Bhat (2007) for more details for computing elasticities). The computed elasticities are presented in Table 4 and the numbers presented in the table represent the expected percentage change in non-motorist crash frequency in response to the change in exogenous variables. For example, the elasticity effect for Vehicle Miles Travelled (VMT) based on the proposed joint model indicates that the expected crashes could increase by 3.075% with an increase in 10% of VMT.

Based on the elasticity effects of NB and joint models, several observations can be made. First, the elasticity effects of the same variables (such as VMT, proportion of heavy vehicle mileage in VMT, proportion of length of local road, etc.) retain the same signs in the two models. Second, although the signs of parameters for VMT in the count and the proportion parts are different in the proposed joint model, its elasticity effect is finally positive which supports previous studies (Lee et al., 2015b; Cai et al., 2016). Third, the elasticity effects of two additional variables signalized intersection density and proportion of population equal to or older than 65 years old can be observed in the proposed model, which further demonstrate the advantage of the joint model. Finally, the elasticity analysis could help provide a clear picture of the exogenous factors' impact on zonal non-motorist crash counts, providing an illustration on how the proposed model can be applied.

Table 5-4 Elasticity effect of independent variables

Variable	NB model	Joint model
VMT	3.215	3.075
Proportion of heavy vehicle mileage in VMT	-3.484	-1.738
Proportion of length of local road	4.036	4.006
Length of sidewalks	3.097	2.184
Population density	1.450	1.517
Median household income	-2.708	-2.129
Proportion of commuters by public transportation	1.445	1.029
proportion of commuters by bicycle	0.555	0.326
Signalized intersection density	-	12.542
Proportion of population aged 65 or over	-	1.411

5.6 Hot Zone Identification Analysis

One potential application of the model results in to allow identification of hot zones experiencing high crash risk based on the detected variables to support long term transportation planning to enhance traffic safety. Based on the joint model, we propose a joint method to identify hot zones for non-motorist crashes. The proposed joint model has two components corresponding to the two modeling targets: crash frequency and crash proportion. As for the crash frequency, the Highway Safety Manual (HSM) (AASHTO, 2010) suggests to employ Excess Predicted Average Crash Frequency (EPF) or Potential for Safety Improvement (PSI) based on Safety Performance Functions (SPFs). The measure can be calculated by the difference between the expected and predicted crash counts. The expected number of crashes is calculated by adjusting the observed number of crashes based on the estimated SPFs to eliminate the fluctuation in the observed number of crashes. Since Bayesian models are used in this study, the expected number of crashes can be computed by the estimated SPFs with random terms (Aguero-Valverde and Jovanis, 2007). Thus, the excess predicted average total crash frequency in the count part of the joint model can be calculated as:

$$EPF_i^{TOT} = N_i^{EXP-TOT} - N_i^{PRD-TOT} \quad (5\text{-}7)$$

$$N_i^{EXP-TOT} = \exp(\beta^{TOT} * x^{TOT} + \theta_i) \quad (5\text{-}8)$$

$$N_i^{PRD-TOT} = \exp(\beta^{TOT} * x^{TOT}) \quad (5\text{-}9)$$

where, EPF_i^{TOT} is the excess predicted average total crash frequency for the count part at zone i . $N_i^{EXP-TOT}$ and $N_i^{PRD-TOT}$ are the expected and predicted number of total crashes, respectively.

As for the crash proportion, Lee et al. (2016) proposed the Excess Predicted Proportion (EPP) as a macroscopic screening performance measure by subtracting the predicted proportion from the observed proportion. Similar to the excess predicted average crash frequency, if EPP exceeds zero, the zone has higher proportion of non-motorist crashes than predicted. On the other hand, the proportion of non-motorist in the zone is lower than predicted if the EPP is smaller than zero. In this study, since the proportion is estimated based on the Bayesian model, the expected proportion can be used instead of observed proportion for the EPP computation. Then, the excess predicted average proportion of non-motorist crashes in the joint model can be calculated as:

$$EPP_i = P_i^{EXP} - P_i^{PRD} \quad (5\text{-}10)$$

$$P_i^{EXP} = exp(\beta^{P\_NON} * x^{P\_NON} + \varphi_i)/(1 + exp(\beta^{P\_NON} * x^{P\_NON} + \varphi_i)) \quad (5\text{-}11)$$

$$P_i^{PRD} = exp(\beta^{P\_NON} * x^{P\_NON})/(1 + exp(\beta^{P\_NON} * x^{P\_NON}) \quad (5\text{-}12)$$

where, EPP_i is the excess predicted average crash proportion of non-motorist crashes at zone i. P_i^{EXP} and P_i^{PRD} are the corresponding expected and predicted proportion of non-motorist crashes, respectively.

According to equations (5-7)-(5-12), the excess predicted average non-motorist crash frequency based on the two parts in the joint model can be calculated as

$$EPF_i^{NON} = N_i^{EXP-TOT} * P_i^{EXP} * \exp(\varepsilon_i) - N_i^{PRD-TOT} * P_i^{PRD} \quad (5\text{-}13)$$

where, EPF_i^{NON} is the excess predicted average non-motorist crash frequency based on the joint model at zone i.

Based on the joint model, three different excess predicted average values can be obtained: EPF of the non-motorist crashes, EPF of the total crashes, and EPP of the proportion of non-motorist crashes. According to the EPF of non-motorist crashes, all TADs in this study could be classified into three categories based on each average value: hot ('H'), warm ('W'), and cold ('C'). Specifically, a TAD was classified as a hot zone if the value is among the top 10%, a warm zone if the value is between 0 and less than the top 10%, or a cold zone if the value is less than 0. The hot zones have much more non-motorist crashes than other zones with similar characteristics. The warm zones are less risky than the hot zones but still have some room for the non-motorist safety improvement. As for the cold zones, they experience less non-motorist crashes compared to other similar zones. Also, all TADs can be classified into the three categories with the same approach based on the EPF of total crashes and the EPP of the proportion of non-motorist crashes. The TADs classified as hot zones for total crashes indicated the zones were with more dangerous driving environment while the TADs classified as hot zones for proportion of non-motorist crashes should be more hazardous for walking and cycling. Since both dangerous driving environment and hazardous walking and cycling condition can contribute to hot zones of non-motorist crashes, the three target results were combined together to provide a broad spectrum perspective for hot zones for non-motorists crashes. The combined classification results

are illustrated in Table 5. The first letter represents the classification results based on the EPFs of non-motorist crashes while the second and third letters represent the classification results based on the EPFs of total crashes and EPPs of proportion of non-motorist crashes, respectively.

Table 5-5 Example of screening results based on joint model

TAD ID	Excess Predicted Average Values			Ranking Percent (%)			Classification Results			
	EPF_ NON	EPF_ TOT	EPP	EPF_ NON	EPF_ TOT	EPP	EPF_ NON	EPF_ TOT	EPP	Combined
1	0.55	58	0.00	36	32	39	W	W	W	WWW
2	6.15	87	0.01	28	28	23	W	W	W	WWW
:	:	:	:	:	:	:	:	:	:	:
11	78.30	455	0.02	3	9	4	H	H	H	HHH
:	:	:	:	:	:	:	:	:	:	:
594	-1.23	-11	0.00	41	44	57	C	C	C	CCC

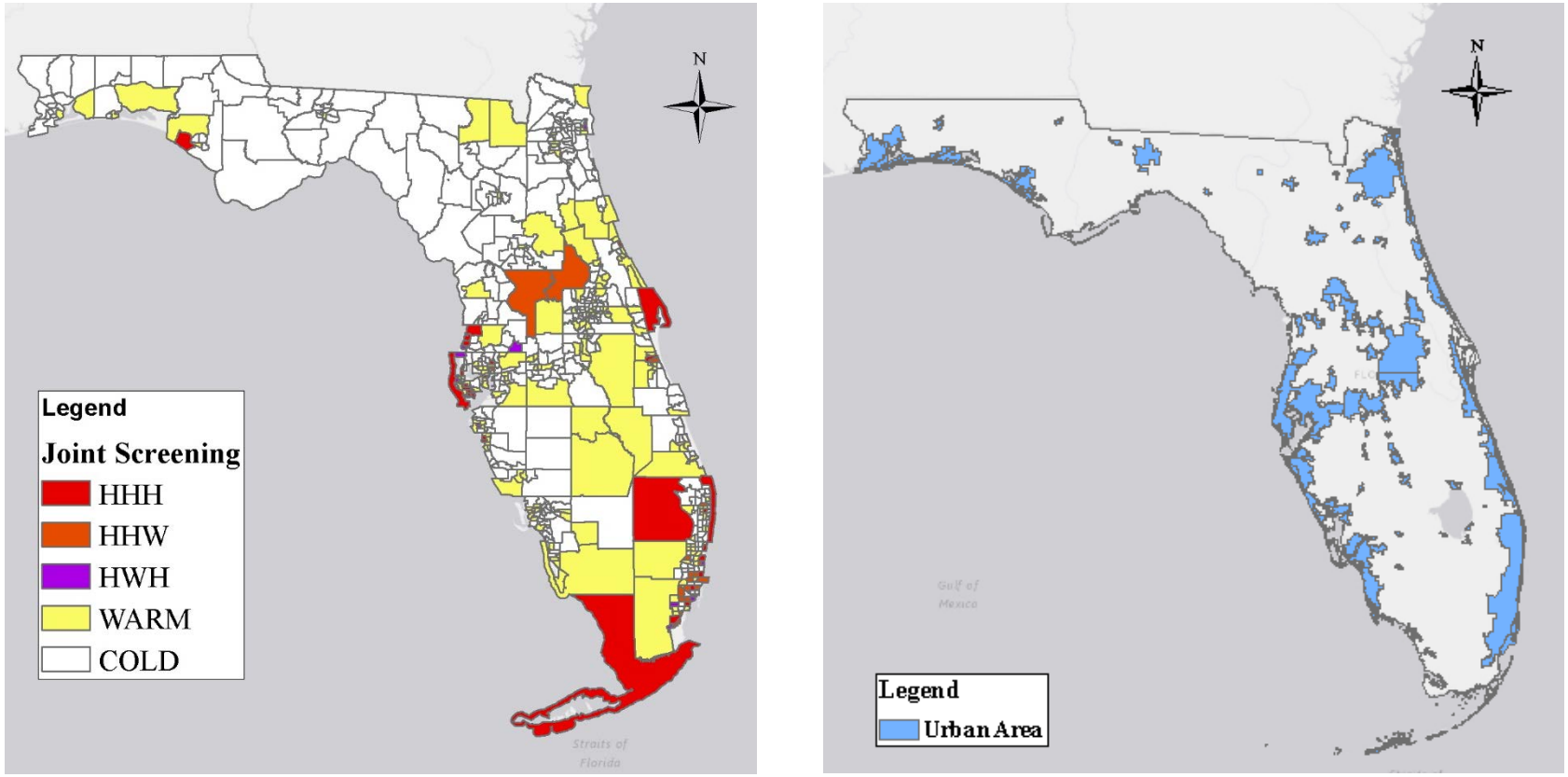
With the 3 targets and 3 traffic safety levels, totally 27 joint classifications could be obtained. However, only twelve combined classifications can be obtained for the 594 TADs as shown in Table 6. Overall, 60 (10%) hot zones of non-motorist crashes were classified, which is top priority for non-motorist safety treatment. These zones have at least dangerous driving environment or hazardous walking and cycling conditions. There are 30 (5.05%) ‘HHH’ zones identified, which require treatments for driving environment as well as walking and cycling condition. Ninteen (3.20%) ‘HHW’ zones and 11 (1.85%) ‘HWH’ zones were also identified. For these zones, the highest priority treatments should be for drivers or non-motorists only. There were 166 (27.94%) warm zones which have moderate risk of non-motorist crashes. For the warm zones of non-motorist crashes, 9 (1.52%) ‘WHW’, 1(0.17%) ‘WHC’, and 18(3.03%) ‘WWH’ zones are categorized. The ‘WHW’ and ‘WHC’ zones have dangerous driving environment, but the non-motorists are not particularly exposed to traffic crashes. On the other hand, the driving environment in the ‘WWH’ zones is moderately safe, whereas the walking and cycling conditions in these zones are dangerous. Also, there were 133(22.39%) ‘WWW’ and 5(0.84%) ‘WCW’ zones which do not have serious problems for either the driving environment

or the walking and cycling conditions. Furthermore, more than half of the zones (61.95%) were classified as cold zones. In these 368 zones, 12(2.02%) 'CWW', 16(2.69%) 'CWC', 57(9.60%) 'CCW', and 283(47.64%) 'CCC' zones were recognized. In these zones, the non-motorists are relatively safer since neither the driving environment nor the walking and cycling conditions are very dangerous.

Based on the combined classification results, the screening result is presented in Figure 2. Since the warm and cold zones of non-motorist crashes are relatively safe, they were combined as 'WARM' and 'COLD' zones, respectively. In order to better understand the spatial pattern of the classified zones, the urban areas in Florida are also presented. As shown in Figure 2, a clustered pattern of different classified zones can be clearly observed. There are several clusters containing multiple 'HHH' zones. These clusters are in the South and Center Florida areas which are mostly mixtures of residence, commerce, and tourism land use. On the other hand, most of the 'COLD' zones formed clusters in the rural areas and also some 'COLD' zones clustered in the Northeast Florida urban area. Furthermore, several clusters having 'WARM' zones can be observed in the rural areas across the whole Florida.

Table 5-6 Number of zones by hot zone classification

Hot zones of non-motorist crashes (N=60)			Warm zones of non-motorist crashes (N=166)			Cold zones for non-motorist crashes (N=368)		
Category	Number of Zones	Percentage (%)	Category	Number of Zones	Percentage (%)	Category	Number of Zones	Percentage (%)
HHH	30	5.05	WHH	0	0	CHH	0	0
HHW	19	3.20	WHW	9	1.52	CHW	0	0
HHC	0	0	WHC	1	0.17	CHC	0	0
HWH	11	1.85	WWH	18	3.03	CWH	0	0
HWW	0	0	WWW	133	22.39	CWW	12	2.02
HWC	0	0	WWC	0	0	CWC	16	2.69
HCH	0	0	WCH	0	0	CCH	0	0
HCW	0	0	WCW	5	0.84	CCW	57	9.60
HCC	0	0	WCC	0	0	CCC	283	47.64



(a) Hot zones distribution (b) Ubran areas distribution

Figure 5-2. Hot zone identification based on the joint model

5.7 Summary and Conclusion

With a growing challenge of global warming and traffic congestion, non-motorist transportation modes such as walking and cycling have been promoted as an environmentally friendly and physically active alternative for short distance trips. However, a strong impediment to universal adoption of non-motorist transportation is the inherent safety risk. Thus, it is necessary to make any effort to enhance the safety of pedestrians and bicyclists. The macro-level crash analysis allows identification of unsafe zones for non-motorists and detection of zonal factors which affect the non-motorist crash occurrences. This paper formulated and estimated models based on count and proportion models to investigate the effects of exogenous factors on pedestrian and bicycle crashes at the Traffic Analysis District (TAD) level in Florida. In order to identify potentially different impacts of exogenous variables on vehicle drivers and non-motorists, we formulated the joint model combining the negative binomial (NB) model and the logit model. More specifically, the NB model part is for the total crash counts to explore the effects on vehicle drivers while the logit model part is for the proportion of non-motorist crashes to investigate the influences on non-motorists. The model was estimated employing a comprehensive set of exogenous variables: traffic measures, roadway information, socio-demographic characteristics, and commuting variables. Also, a traditional NB model was developed and compared with the joint model.

The results of the joint model obviously highlighted the existence of different impact of exogenous factors on drivers and non-motorists for pedestrian and bicyclist crashes. The model comparison indicates that the proposed joint model can provide better performance over the NB model. In addition, more significant variables such as signalized intersection density and

proportion of population age 65 or over could be observed in the proposed model. Moreover, the result of the joint modeling emphasized that the importance of several other variables including traffic (e.g., VMT, proportion of heavy vehicle mileage, etc.), roadway (e.g., length of local road, length of sidewalk, etc.), socio-demographic characteristics (e.g., population density, median household income, etc.), and commuting variables (e.g., commuters by public transportation and those by bicycle). To provide a clear quantitative comparison of the variables' impact, elasticity effects for the NB and joint models are computed. The results revealed that the same significant variables in the two models would have the same signs of elasticity effects on the non-motorist crashes. Also, the elasticity effect calculation allows us to determine the factors that substantially increase crash risk for crashes involving pedestrians and bicyclists.

Subsequently, a novel joint method to identify hot zones of non-motorist crashes was proposed based on the joint model results. The hot zones of non-motorist crashes were classified into three categories: hot zones with dangerous driving environment only, hot zones with hazardous walking and cycling condition only, and hot zones with both dangerous driving environment and hazardous walking and cycling condition. According to the different categories, the appropriate treatments should be provided correspondingly to improve the driving environment, the walking and cycling conditions, or both.

Based on our study, it is clear that analysis of non-motorist crashes should explore the different effects of exogenous factors on both drivers and non-motorists. Despite of the contributions of this study, there are some limitations that are expected to be addressed in future research. In the proposed model, the spatial correlation among adjacent zones has not explored yet. Further study is required to accommodate for the spatial correlation as well in the proposed joint model.

Besides, the formulated model was estimated using traffic exposure, roadway information, socio-demographic characteristics, and commuting variables. The model performance can be improved if more variables reflecting the driving, walking, and cycling environment and increasing non-motorist crash occurrences could be included in the future study. The proposed model can also be adopted to explore crashes of different severity levels and other crash characteristics such as single-vehicle crashes and head-on crashes. Further, the proposed joint model can be extended to a multivariate modeling structure if researches want to simultaneously analyze the crashes of different severity levels or crash types.

CHAPTER 6: INVESTIGATING MACRO-LEVEL EFFECTS IN MICRO-LEVEL CRASH ANALYSIS

6.1 Introduction

Segments and intersections are two major parts of road network to carry traffic demands. In the previous literature, numerous traffic crash prediction models have been developed at the micro-level for the two types of road facilities. The choice of appropriate analytical models and the selection of representative variables would be the two crucial factors to obtain accurate modeling results. With various advanced statistical methodology such as spatial and temporal autocorrelation, finite mixture/latent class, zero inflation, random effects/parameters, and multilevel approaches, the effects of road features and traffic characteristics on the crashes of road facilities have been recognized and included in the micro-level crash prediction models. Beside the micro-level factors, the road facilities should share certain macro-level factors, which may affect travel behaviors, traffic modes, and further affect the crash occurrences. Although the crash studies at macro-level have suggested that the zonal factors such as socioeconomic characteristics have sustainable effects on traffic safety, only few studies have included macro-level data for micro-level safety analysis. Omission of important explanatory variables at macro-level may result in biased and inconsistent parameter estimates (Wang et al., 2017; Mannering et al., 2016).

The study of this chapter aims to investigate the potential macro-level effects on crashes at the micro-level. Toward this end, a hierarchical joint model is proposed to analyze crashes at both segments and intersections by incorporating both macro-level data. Besides, the spatial

autocorrelation between segments and intersection is also considered in the proposed model. The suggested model development would enable us to better understand crash occurrence at the micro-level by considering the macro-level effects.

The following parts of this chapter are organized into four sections. The following section introduces a new hierarchical joint model to incorporate macro-level factors in micro-level crash analysis models. The third section presents the collected data used in this chapter while the fourth section discusses the model results. Finally, the fifth section summarizes the findings of this chapter.

6.2 Methodology

The traditional Poisson and negative binomial models have been widely used to analyze the discrete, random and non-negative crash data. Nevertheless, the models assume that the observations are independent from each other and do not consider the potential correlation of the traffic crash counts, which may lead to poorly estimated results (Skinner et al., 1989; Goldstein, 1995; Lord and Mannering, 2010). Generally, two types of correlations may exist in the crash data: (1) macro-level correlation; (2) spatial correlation. First, it would be reasonable to claim that the road entities located in the same zone should share certain macro-level factors, which may affect crash occurrence through driving behaviors and transportation modes. Hence, considering the macro-level effects would enhance the crash analysis models at the micro-level (Wang et al., 2016; Lee et al., 2017). If both micro- and macro-level data are considered for the analysis, the data used naturally has a two-level hierarchy. Hence, it would be appropriate to adopt a hierarchical modeling technology, which allows multilevel data structures to be properly

estimated and specified (Gelman and Hill, 2007; Yu et al., 2013). In addition, road entities may share unobserved spatial effects if they are in close proximity (Lord and Mannering, 2010). Compared with solely spatial autocorrelation between segments or intersections, the spatial correlation effects between adjacent segments and intersections may be more significant if they are directly connected with each other. Therefore, a two-level hierarchical joint model incorporating spatial effects is proposed in this study. Specifically, the hierarchical model is composed of a micro-level model (level-1 model) and a macro-level model (level-2 model) in a Bayesian framework.

The level-1 model accounts for both micro-level factors and spatial autocorrelation in crashes between road entities. To consider the potentially spatial correlations between different types of road entities (segments and intersections), Zeng and Huang (2014) introduced a spatial joint model with a Conditional Autoregressive (CAR) prior, which was subsequently used by Wang and Huang (2016) and Huang et al. (2017). The level-1 model can be expressed as follows:

$$y_{ij}^{entity} \sim Poisson\ (\lambda_{ij}^{entity}) \tag{6-1}$$

$$\begin{aligned}\log\left(\lambda_{ij}^{entity}\right) = {} & \gamma_{ij} \times \left(\beta^{seg} x_{ij}^{seg} + log\left(length_{ij}^{seg}\right) + v_j^{seg}\right) + \left(1 - \gamma_{ij}\right) \\ & \times \left(\beta^{inter} x_{ij}^{inter} + v_j^{inter}\right) + \theta_{ij}^{entity} + \phi_{ij}^{entity}\end{aligned} \tag{6-2}$$

where, y_{ij}^{entity} is the observed crash frequency of road entity *i* in zone *j* with the underlying Poisson mean λ_{ij}^{entity}. x_{ij}^{seg} and x_{ij}^{inter} denote the set of explanatory variables of segments and intersections while β^{seg} and β^{inter} are the corresponding parameters. If road entity *ij* is a segment, $\gamma_{ij} = 1$, otherwise, $\gamma_{ij} = 0$. $\log(length_{ij}^{seg})$ is logarithm of the length of road entity *ij* if it is a

segment, otherwise it is zero. v_j^{seg} and v_j^{inter} are two intercepts, which are used to denote macro-level effects for segments and intersection, respectively. θ_i^{entity} is a random effect accounting for the unstructured over-dispersion that follows a normal distribution:

$$\theta_i^{entity} \sim N(0, \frac{1}{\tau_h}) \tag{6-3}$$

where τ_h is the precision parameter (the inverse of the variance) which follows a prior gamma (0.001, 0.001).

ϕ_i^{entity} represents the random effect which is used to deal with the spatial autocorrelation effect. If two road entities directly connect with each other the weight in the spatial proximity matrix is set to be 1, otherwise, the weight is 0. This approach in the joint model can not only capture the spatial correlation of road entities of the same type but also the two different types of road entities including segments and intersections. ϕ_{ij}^{entity} follows a normal distribution with CAR prior suggested by Besag et al. (1991):

$$\phi_{ij}^{entity} \sim \mathrm{N}(\frac{\sum w^{entity} \phi^{entity}}{\sum_{i \neq j} w^{entity}}, \frac{1}{\tau_c \sum w^{entity}}) \tag{6-4}$$

where w^{entity} is the spatial proximity weight. τ_c is the precision parameter, which follows a prior gamma (0.001,0.001).

The level-2 model accounts for the macro-level effects shared by road entities. In the previous study, the macro-level effects were quantified by using a random term (Ahmed et al., 2011; Usman et al., 2012; Yu et al., 2013; Yu and Abdel-Aty, 2013a; Yu and Abdel-Aty, 2013b), or a

set of macro-level explanatory variables (Wang et al., 2016; Huang et al., 2016; Lee et al., 2017). Given that road entities in a zone share not only macro-level explanatory variables but also total crashes occurring at all road entities in the same zone, it may be better to quantify the macro-level effects by considering the total crash frequency of the specific zones. Then, the level-2 model can be specified as follows:

$$y_j^{seg} \sim Poisson\ (\lambda_j^{seg}) \tag{6-5}$$

$$y_j^{inter} \sim Poisson\ (\lambda_j^{inter}) \tag{6-6}$$

$$v_j^{seg} = \delta^{seg} x_j^{seg} + \theta_j^{seg} + \phi_j^{seg} \tag{6-7}$$

$$v_j^{inter} = \delta^{inter} x_j^{inter} + \rho * \theta_j^{seg} + \theta_j^{inter} + \phi_j^{inter} \tag{6-8}$$

$$log(\lambda_j^{seg}) = h^{seg} v_j^{seg} \tag{6-9}$$

$$log(\lambda_j^{inter}) = h^{inter} v_j^{inter} \tag{6-10}$$

where y_j^{seg} and y_j^{inter} are the total crashes in all segments or intersections in the same zone *j* with the underlying Poisson means λ_j^{seg} and λ_j^{inter}. x_j^{seg} and x_j^{inter} are the macro-level explanatory variables for the segments and intersections, respectively. δ^{seg} and δ^{inter} are the corresponding parameters. θ_j^{seg} and θ_j^{inter} are random effects accounting for the unstructured over-dispersion. θ_j^{seg} with coefficient ρ is used to realize the potential correlation between macro-levels effects on segments and intersections. In addition to the equivalence relation presented in Equations (6-7) and (6-8), the macro-level effects on segments and intersection are also linked to the total expected crashes at all segments and intersections in the specific zones with an adjustment factor h^{seg} and h^{inter}. Notably, although the expected crash counts of all segments and intersection in each zone are used for the model estimation, they are not included

in the final prediction model for road entities. Instead, they serve as additional constraints, which can help better recognize the macro-level effects.

In order to validate the performance of the proposed model, two other hierarchical models were estimated: one has random terms only and one has macro-level explanatory variables but does not consider the total expected crashes of all segments and intersections in the same zone. In addition, a base model only having micro-level explanatory variables was also estimated.

All models were run considering a non-informative normal $(0,10^6)$ prior for all coefficients. To avoid the adverse impact of significant correlation, the variables with high correlation were not employed in the model at the same time. The significant explanatory variables were determined based on 95% certainty of Bayesian credible intervals (BCIs). The optimal set of parameters for each model was determined based on DIC (deviance information criterion). The DIC was also used to compare models' performance. Roughly, differences of more than ten might indicate that the model with lower DIC performs better (El-Basyouny and Sayed, 2009). Besides DIC, two other measures were employed to for the comparison: MAE (mean absolute error) and RMSE (root mean squared error). The formulae for the two measures are as follows:

$$MAE = \frac{1}{N}\sum_{ij=1}^{N}|y_{ij} - y'_{ij}| \tag{6-11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{ij=1}^{N}(y_{ij} - y'_{ij})^2} \tag{6-12}$$

where N is the number of observations, y_{ij} and y'_{ij} are the observed and predicted number of crashes of road facility ij.

6.3 Data Preparation

In this study, totally 3,316 road facilities including 2,434 segments and 882 intersections in Orlando, Florida, were selected for the empirically analysis of the proposed models (Figure 6-1(a)). Seventy-eight traffic analysis districts (TADs), zones where the road entities were located, were also selected for the analysis (Figure 6-1(b)). The TADs are newly developed transportation-related zones by combing existing traffic analysis zones (TAZs) (FHWA, 2011). In the earlier studies, the TAZs have been widely adopted for crash analysis since they are easier to be adopted to integrate traffic safety with the transportation planning process. However, many road entities are near boundaries of TAZs since one of the zoning criteria of TAZs is to recognize physical boundaries such as arterial (Lee et al., 2014; Cai et al., 2017a). Hence, it might be difficult to recognize the zonal effects of TAZs since the excess road entities are near the boundaries. In Orlando, the area of TADs (on average 36.59 mile^2) is considerably larger than that of TAZs. Therefore, it is deduced that most of road entities could be located inside of TADs (Lee et al., 2017). For the road entities on the boundaries of two or more TADs, a geospatial method was applied in this study to assign them into TADs. Specifically, each intersection was allocated into a TAD if the intersection is located within the digital boundary of the TAD. Meanwhile, each segment was assigned into a TAD if most part of the segment is in the corresponding TAD. Hence, each road facility has one corresponding TAD with the one-to-one spatial relation between road entities and TADs. In this study, four types of data including

traffic crash data, traffic characteristics, road features, and zonal factors were collected for the analysis.

Crash data in a three-year period (2010-2012) were obtained from the Florida Department of Transportation (FDOT) Crash Analysis Reporting System (CARS) and Signal Four Analytics (S4A). In the crash database, crashes were defined as "crashes at intersection" or "crashes influenced by intersection" if they occurred within 250 feet away from the intersection. Based on this principle, a 250 feet buffer around each intersection were created and crashes inside the buffers were defined as intersection-related crashes while others were categorized as segment-related crashes. A total of 60,144 crashes were collected among which 14,873 (24.7%) were intersection-related crashes and 45,271 (75.3%) were segment-related crashes. The crashes were also aggregated based on TADs by summing up the crash count of all road facilities in the corresponding TAD according to the spatial relations.

Ten segment variables and six intersection variables were collected from the FDOT Roadway Characters Inventory (RCI). Average Annual Daily Traffic (AADT), as an indicator of traffic exposure, was collected for both segments and intersections. For road features, segment variables considered in this study are functional class of roads, number of lanes, segment length, presence of median, and location of segments while intersection variables include presence of traffic signal, number of legs, and location of intersections.

The segment and intersection variables were also aggregated into TADs in a similar way as crashes. It should be noted that the intersection density is the number of intersections divided by the length of total road length. The distance to the nearest urban are is defined as the distance

from the centroid of the TADs to the nearest urban region. Beside traffic and road characteristics, the socio-demographic data were attained by aggregating census-tract-based data from the U.S. Census Bureau. These census-tract-based data could be aggregated into TADs as a TAD is a combination of multiple census tracts (Cai et al., 2017a). Table 6-1 provides descriptive statistics of collected data based on road facilities and TADs.

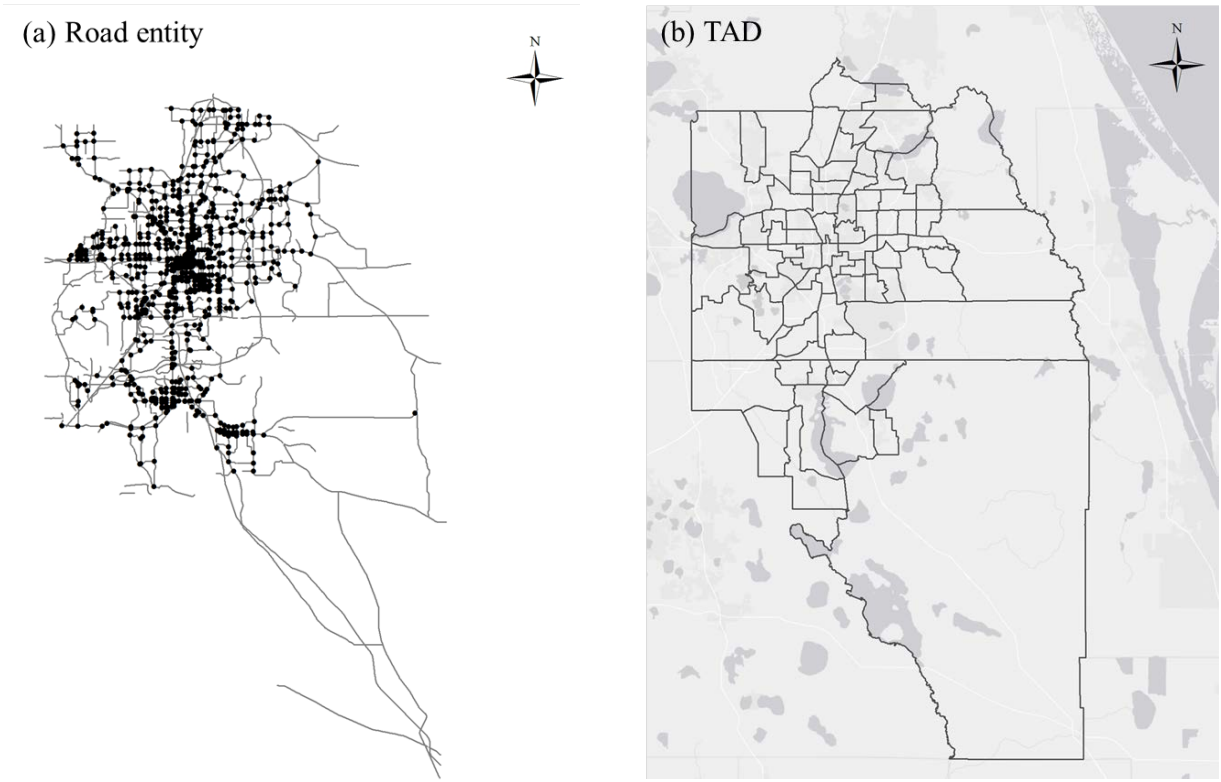


Figure 6-1 Road entities and TAD in Orlando, Florida

Table 6-1 Descriptive statistics of collected data

Variables	Definition	Mean	S.D.	Min.	Max.
Segment variables					
CRASH	Three-year crash count for each segment	6.20	12.59	0	132
AADT	Average annual daily traffic (in thousand)	20.19	25.51	0.20	195.77
LENGTH	Segment length (mile)	0.75	1.35	0.10	30.91
FREEWAY	Freeway indicator: 1 if freeway, 0 otherwise	0.11	0.31	0	1
ARTERIAL	Arterial indicator: 1 if arterial, 0 otherwise	0.39	0.49	0	1
COLLECTOR	Collector indicator: 1 if collector, 0 otherwise	0.49	0.50	0	1
LOCALROAD	Local road indicator: 1 if local road, 0 otherwise	0.01	0.11	0	1
MEDIAN	Median barrier indicator: 1 if present, 0 otherwise	0.63	0.48	0	1
LANE1_2	1 or 2 lanes indicator: 1 if yes, 0 otherwise	0.56	0.50	0	1
LANE3_4	3 or 4 lanes indicator: 1 if yes, 0 otherwise	0.30	0.46	0	1
URBAN	Urban indicator: 1 if in urban area; 0 otherwise	0.93	0.26	0	1
Intersection variables					
CRASH	Three-year crash count for each intersection	16.86	20.34	0	135
MAJ_AADT	AADT on major approach (in thousand)	23.72	15.76	0.60	81.50
MIN_AADT	AADT on minor approach (in thousand)	8.22	7.64	0.20	52.50
SIGNAL	Traffic signal indicator: 1 if present, 0 otherwise	0.76	0.43	0	1
LEG3	3-Leg intersection indicator: 1 if yes, 0 otherwise	0.31	0.46	0	1
LEG4	4-Leg intersection indicator: 1 if yes, 0 otherwise	0.69	0.46	0	1
URBAN	Urban indicator: 1 if in urban area; 0 otherwise	0.99	0.10	0	1
TAD related variables					
CRASH	Three-year crash count for each TAD	257.03	213.17	18	1038
DVMT	Daily vehicle-miles traveled (in thousand)	494.53	440.19	23.30	2210.21
P_HVMT	Proportion of heavy vehicle in DVMT	0.08	0.03	0.04	0.19
ROAD_LENGTH	Total road length in each TAD (mi)	23.60	29.72	1.53	248.65
P_FREEWAY	Proportion of segment length of freeway	0.14	0.17	0	0.71
P_ARTERIAL	Proportion of segment length of arterial	0.40	0.21	0	0.74
P_COLLECTOR	Proportion of segment length of collector	0.46	0.22	0	1
P_LOCALROAD	Proportion of segment length of local road	0.01	0.03	0	0.23
P_LANE1_2	Proportion of segment length with 1 or 2 lanes	0	0	0	0.03
P_LANE3_4	Proportion of segment length with3 or 4 lanes	0.39	0.22	0	0.87
P_LANE5MORE	Proportion of segment length with 5 lanes or over	0.16	0.17	0	0.74
INTER_DENS	Number of intersections per mile (/mile)	1.70	0.57	1	4.33
P_SINGAL	Proportion of signalized intersections	0.78	0.24	0	1
P_LEG3	Proportion of intersections with 3 legs	0.32	0.17	0	0.73
P_LEG4	Proportion of intersections with 4 legs	0.67	0.18	0	1
POP_DENS	Population density (in thousand)	2.38	1.49	0.02	6.56
P_AGE1524	Proportion of population aged 15-24	0.16	0.05	0.09	0.38
P_AGE65MORE	Proportion of population aged 65 or over	0.10	0.03	0.04	0.18
MEDIAN_INC	Median household income (in thousand)	63.40	19.47	33.99	122.77
DIS_URBAN	Distance to the nearest urban area (mi)	1.40	1.71	1	14.12

6.4 Model Results

6.4.1 Model Performance

As discussed in the previous section, totally four models were estimated in this study as follows:

- Base model: crash prediction model only having micro-level explanatory variables;
- Hierarchical model (1): crash prediction model having micro-level explanatory variables and considering macro-level effects with random terms;
- Hierarchical model (2): crash prediction model having micro-level explanatory variables and considering macro-level effects with explanatory variables;
- Hierarchical model (3): crash prediction model having micro-level explanatory variables and considering macro-level effects with both explanatory variables and total crashes of segments and intersections.

Prior to discussing the model results, the model performance was summarized and presented in Table 6-2. Several observations can be made from the results. First, it was found that the three hierarchical models consistently outperform the base model without considering the macro-level effects on the micro-level crashes. The differences of DIC between the base model and hierarchical models are at least 15, which indicates a substantial improvement by considering the macro-level effects. The results validate our hypothesis that the road entities share macro-level factors which can affect the crash occurrence in segments and intersections. Second, the exact ordering alters among three hierarchical models based on DIC, MAE, and RMSE. The hierarchical model (3) can provide significantly smaller DIC compared with other two hierarchical models (El-Basyouny and Sayed, 2009). The goodness-of-fit for the third

hierarchical model is also improved by at least 14.51% and 10.45% based on the values of MAE and RMSE. Third, although hierarchical model (2) can provides slightly better model performance compared with hierarchical model (1), the differences are not significant. Hence, in terms of the results, we can conclude that the proposed hierarchical model, which not only considers macro-level explanatory variables but also uses total crash of zones as priors in the model, offers the best statistical fit for micro-level crashes. The findings are somewhat not surprising since the hierarchical model (3) analyzes the crash frequency for road entities with the prior information that how many total segment- or intersection- crashes occur in the zones. Such prior information serves as a constraint which can help better realize the macro-level effects.

Table 6-2 Comparison results of model performance

Category	DIC	MAE	RMSE
Base model	17524.30	10.16	24.43
Hierarchical model (1)	17509.50	7.92	18.29
Hierarchical model (2)	17501.00	7.79	17.90
Hierarchical model (3)	17472.00	6.66	16.03

6.4.2 Modeling Result

The results of four models (i.e., one base model and three hierarchical models) for crashes of segments and intersections are displayed in Table 6-3. The results of the base model and hierarchical model (1) only present the micro-level variables with significant effects and random terms while the hierarchical models (2) and (3) results are composed of variables from both micro- and macro-levels. Same significant micro-level variables can be found in the four models with consistent signs of parameter. Meanwhile, more macro-level variables are found significant in hierarchical model (3). Furthermore, the variance of the macro-level random effect in the

hierarchical model (1) is statistically significant, which confirms the existence of within-zone homogeneities. While the results summarized in Table 6-3, the following discussions about the parameters estimates focuses on the hierarchical model (3) which has best fit and more significant variables.

(1) Level-1 (Micro-Level) Variables

As shown in Table 6-3, totally 8 micro-level variables are statistically significant for crashes of segments or intersections with 95% BCIs. The variables related to traffic volumes (i.e., AADT of segments, MAJ_AADT and MIN_AADT of intersections) are measures of vehicle exposure and as expected have positive effects on the propensities of crashes for both segments and intersections.

Three other variables are found to significantly affect crash occurrence on segments: functional class of roadway is arterial (ARTERIAL), number of lanes is 1 or 2 (LANE1_2), presence of median barrier (MEDIAN). Compared with other road types, arterials have partially limited accesses with comparatively higher traffic volumes. Hence, the arterial would have more traffic interactions and conflicts within the same road length. A road segment will have fewer crashes if it only has one or two lanes since interactions among vehicles are generally increased on roads with more lanes. As consistent with the previous studies (Anastasopoulos et al., 2012), the presence of median barriers will increase crash counts on the road segments.

Concerning intersections, two additional critical variables are found to be significant, i.e., presence of traffic signal (SIGNAL), number of legs is 3 (LEG3). The signal control is usually

installed at intersections with higher traffic volumes which lead more traffic interaction (Wang et al., 2016). Also, the existence of dilemma zones due to the signalized control can lead to more crashes (Wu et al., 2015). As suggested in the previous studies (Wang and Huang, 2016; Huang et al., 2016), more crashes tend to occur at intersections with more legs. Therefore, the 3-leg intersection indicator is negatively associated with the crash frequency of intersections.

(2) Level-2 (Macro-Level) Variables

The result suggests a significantly positive association between macro-level effects on road facilities and the total crashes of specific zones for both segments and intersections. The finding is expected since crashes should be more likely to occur at the road facility which is located in the zone with more crashes.

Both segments and intersections have five significant macro-level explanatory variables. Among these variables, three common variables are found for segments and intersections: daily vehicle miles travelled (DVMT), distance of TAD centroid to the nearest urban area (DIS_URBAN), and median household income (MEDIAN_INC). The DVMT can increase the likelihood of crash occurrences at both segments and intersections. It can be reasoned that increased DVMT are correlated with increases in the traffic volume of a road entity and the interactions with the connected segments or intersections. As the distance of TAD centroid to the nearest urban region increases, the traffic crash risk at segments and intersections is reduced- a sign of low traffic exposure in the suburban regions. Besides, the distance might be correlated with intensity of land use, which may be an underlying factor for some of the observed effects (Pulugurtha et al., 2013; Wang and Huang, 2016). Segments and intersections, which are located in the TAD with higher

median household income, would experience less traffic crashes. Several previous studies (Huang et al., 2010; Xu et al., 2014; Cai et al., 2017) focused on macro-level crash analysis found the similar effects and argued that individuals from relatively affluent area are more likely to be better educated and seek for safer driving behavior. Besides, drivers and passengers with higher income seem more willing to use seatbelts (Lerner et al., 2001) and their vehicles tend to be more advanced (Girasek and Taylor, 2010).

For segments, two more macro-level variables are significant. The variable proportion of heavy vehicle in DVMT (P_HVMT) is negatively related to crash occurrence at segments. The variable could be a reflection of industry area with less traffic exposure (Lee et al., 2016). Besides, compared with passenger car drivers, heavy vehicle drivers should be more professional to avoid collisions at segments (Carrigan et al., 2014). A segment would have more crashes if it is located in a TAD with high proportion of arterial (P_ARTERIAL), which is understandable since crash risk is relatively higher in arterials according to the previous study (Huang et al., 2010; Jiang et al., 2016). As discussed in the micro-level, traffic might be more complicated in arterials with partially limited access and high traffic volume. Hence, a segment would experience increased traffic interaction and conflicts if connected with arterials.

For intersections, two additional variables intersection density (INTER_DENS) and proportion of population between age 15 and 24 (P_AGE1524). High intersection density can increase the likelihood of crash occurrences (Wang et al., 2014; Xu et al., 2014). A possible reason is that higher intersection density is correlated with more vehicles turning and lane changing maneuvers, which results in increased traffic collisions (Wang et al., 2016). The finding about the young drivers is consistent with the well-known fact that young drivers prone to be involved in crashes

due to the lack of driving experience (Huang et al., 2010). Also, the young drivers are more likely to engage in aggressive driving acts, including speeding and red light running (Simons-Morton et al., 2005; Yan et al., 2005).

(3) Random Effects

In the level-1 model, the variance of spatial correlation is statistically significant in all models. This result confirms the existence of the intrinsic spatial autocorrelation between intersections and their connected segments, which is consistent with the previous researches (Zeng and Huang, 2014; Wang and Huang, 2016; Huang et al., 2016). Besides, all hierarchical models can provide smaller variance due to unobserved factors and spatial correlation compared with the base model. This indicates that the macro-level variables can be used to explain parts of the unexplained variation. In addition, the hierarchical model (3) provides the smallest variance of random effects, which further suggested the proposed model can provide better analysis results for the micro-level.

At the level-2 model, the parameter ρ is significant, which implies that there exist common factors between the macro-level effects on segments and intersections in each TAD although they are unobserved. Furthermore, the variances of spatial effects for macro-level effects were found to be significant at the 5% level. It suggests that both macro-level effects on segments and intersections are spatially correlated among adjacent zones.

Table 6-3 Modeling Result

	Base model			Hierarchical model (1)			Hierarchical model (2)			Hierarchical model (3)		
Variable	Mean	95% BCI		Mean	95% BCI		Mean	95% BCI		Mean	95% BCI	
Level-1 (Micro-Level)												
Segment												
Intercept	-3.34	-3.47	-3.21	-3.42	0.01	-3.41	-2.34	-2.50	-2.16	-5.27	-5.44	-5.11
AADT	0.55	0.54	0.56	0.57	0.00	0.57	0.54	0.53	0.55	0.56	0.55	0.57
ARTERIAL	0.27	0.22	0.34	0.39	0.01	0.39	0.36	0.28	0.44	0.37	0.31	0.45
LANE1_2	-0.41	-0.47	-0.34	-0.16	0.01	-0.16	-0.20	-0.30	-0.11	-0.16	-0.26	-0.06
MEDIAN	0.11	0.05	0.16	0.18	0.01	0.18	0.24	0.13	0.34	0.19	0.10	0.29
Intersection												
Intercept	-8.18	-8.35	-7.99	-8.31	0.02	-8.29	-7.65	-7.79	-7.37	-8.65	-8.97	-8.42
MAJ_AADT	0.75	0.74	0.76	0.80	0.00	0.80	0.83	0.81	0.85	0.80	0.77	0.81
MIN_AADT	0.29	0.27	0.31	0.27	0.00	0.27	0.27	0.25	0.29	0.24	0.22	0.28
SIGNAL	0.45	0.38	0.53	0.34	0.01	0.35	0.48	0.27	0.67	0.43	0.29	0.54
LEG3	-0.51	-0.59	-0.42	-0.54	0.01	-0.54	-0.50	-0.65	-0.36	-0.50	-0.63	-0.39
Level-2 (Macro-Level)												
h^{seg}	-	-	-	-	-	-	-	-	-	3.33	3.45	3.23
h^{inter}	-	-	-	-	-	-	-	-	-	12.50	16.67	9.09
Segment												
Fixed effect	-	-	-	-	-	-	0.25	0.02	0.47	-0.57	-0.59	-0.55
DVMT	-	-	-	-	-	-	-	-	-	0.29	0.28	0.29
P_HVMT	-	-	-	-	-	-	-0.74	-1.26	-0.14	-1.32	-1.72	-0.71
P_ARTERIAL	-	-	-	-	-	-	0.40	0.06	0.65	0.16	0.07	0.21
DIS_URBAN	-	-	-	-	-	-	-	-	-	-0.11	-0.19	-0.06
MEDIAN_INC	-	-	-	-	-	-	-0.11	-0.13	-0.10	-0.11	-0.11	-0.11
Intersection												
Fixed effect	-	-	-	-	-	-	-0.39	-0.77	-0.09	0.18	0.15	0.21
DMVT	-	-	-	-	-	-	-	-	-	0.05	0.05	0.05
INTER_DENS	-	-	-	-	-	-	-	-	-	0.14	0.12	0.15
P_AGE1524	-	-	-	-	-	-	-	-	-	0.15	0.05	0.31
DIS_URBAN	-	-	-	-	-	-	-	-	-	-0.04	-0.09	-0.01
MEDIAN_INC	-	-	-	-	-	-	-0.07	-0.08	-0.07	-0.05	-0.05	-0.05
Random effects												
Micro-level												
SD[θ_i^{entity}]	2.73	2.40	3.07	0.56	0.53	0.59	0.59	0.56	0.61	0.61	0.58	0.64
SD[ϕ_i^{entity}]	3.90	3.22	4.83	2.22	1.73	2.70	2.27	1.80	2.65	1.55	1.01	2.27
Macro-level												
ρ	-	-	-	-	-	-	2.35	-0.53	9.18	2.35	0.84	4.49
SD[θ_j^{seg}]	-	-	-	0.26	0.21	0.31	0.42	0.10	0.73	0.21	0.12	0.34
SD[θ_j^{inter}]	-	-	-	0.13	0.05	0.21	0.32	0.02	1.28	0.36	0.12	0.54
SD[ϕ_j^{seg}]	-	-	-	-	-	-	1.51	1.02	2.08	0.30	0.20	0.37
SD[ϕ_j^{inter}]	-	-	-	-	-	-	0.22	0.05	0.53	0.13	0.02	0.31

6.5 Summary and Conclusion

The study in this chapter sought to examine the effects of macro-level factors on crashes at the micro-level. For this purpose, this study formulated a Bayesian hierarchical model for both segments and intersections accounting for both macro- and micro-level data. As for the macro-level data, not only macro-level explanatory variables such as socio-economic characteristics but also the total crashes aggregated at macro-level were considered in the proposed model. Meanwhile, the road features and traffic characteristics at the micro-level were included in the proposed model. In addition, the study suggested considering the potentially spatial autocorrelation between segments and intersections by a joint modeling structure. Three models were also estimated for comparison: a base model only having micro-level explanatory variables, a hierarchical model having micro-level explanatory variables and considering macro-level effects with random terms only, and hierarchical model having micro-level explanatory variables and considering macro-level effects with both macro-level explanatory variables and total crashes. The crashes that occurred at both segments and intersections in Orlando, Florida during 2010-2012 were collected for the analysis. The selected crashes were aggregated at both macro- and micro-levels and a comprehensive set of exogenous variables from the two levels were selected for the model estimation. The estimated model performance was evaluated based on the following measures: deviance information criterion, mean absolute error, and root mean squared error.

The results clearly suggested that considering macro-level effects can improve the model performance for micro-level crash analysis. The model comparison exercise indicated that the all hierarchical models considering macro-level effects outperformed the base model. Among the

three hierarchical models, the proposed model considering both macro-level explanatory and total crashes of zones offered the best fit for the crash prediction for the micro-level. Besides, significant spatial autocorrelation can be observed between segments and intersections. Furthermore, the proposed hierarchical joint model results clearly highlighted the importance of several micro-level variables including segment-based variables (e.g., AADT, arterial indicator, 1 or 2 lanes indicator), intersection-based variables (e.g., AADT on major and minor approaches, traffic signal control indicator). Finally, the results further indicated that macro-level, such as proportion of segment length of arterial, intersection density, proportion of population aged 15-24, and median household income, have significant effects on crashes at segments and intersections.

CHAPTER 7: INTEGRATING MACRO AND MICRO LEVEL SAFETY ANALYSES

7.1 Introduction

In the last few decades, there has been a growing recognition of the importance of safety in transportation research. Initially, the Transportation Equity Act for the 21st Century (Houston, 1998) suggested to consider safety in the transportation planning process. Later, Washington et al. (2006) discussed how to incorporate safety into transportation planning at different levels. Currently, the Moving Ahead for Progress in the 21st Century Act (MAP-21 Act) (US Congress, 2012) and Fixing America's Surface Transportation Act (FAST Act) (U.S. DOT, 2015) require the incorporation of transportation safety in the long-term transportation planning process.

One of the most widely used approaches to investigate traffic safety is crash frequency modeling, which can quantify exogenous factors contributing to the number of traffic crashes. Traditionally, crash frequency analyses have been adopted for both macro- and micro-levels. However, previous studies have explored traffic safety at either the micro- or micro-level, i.e., to the best of our knowledge no study has integrated the two levels. If traffic safety research is conducted for the same study area, macro- and micro-level crash analyses would investigate the same crashes but by different aggregation levels. Hence, we can assume that the crash counts at the two levels are correlated. Particularly, the total number of crashes in each zone (macro-level) is supposed to be the same as the total number of crashes from all road entities including segments and intersections (micro-level) located in the zone of interest. Therefore, an integrated crash frequency analysis might improve the model performance and can help in better understanding

the crash mechanisms as well. As a result, more effective and efficient countermeasures can be provided for both macro and micro levels to enhance transportation safety.

This study aims to propose an integrated model to deal with the following issues: (1) to investigate transportation safety problems at macro- and micro-levels, simultaneously; (2) to handle the potential correlation of crash counts between macro- and micro-levels based on the spatial interactions between the two different aggregation levels; (3) to consider the spatial autocorrelation of the road entities (i.e., segments and intersections) by employing a joint model structure at the micro-level.

7.2 Methodology

7.2.1 Bayesian non-integrated spatial model

(1) Bayesian non-integrated spatial model at the macro-level

Traditional Poisson and negative binomial models have been widely used in the previous macro-level traffic safety literature. Nevertheless, the models do not consider a possible spatial correlation of traffic crash counts between adjacent zones, which may yield biased modeling results (Hadayeghi et al., 2010; Quddus, 2008). By incorporating an error term for possible spatial autocorrelation, the Bayesian spatial Poisson lognormal model with Conditional Autoregressive (CAR) prior can provide more appropriate analysis results and has been widely adopted in macro-level crash analysis (Miaou et al., 2003; Quddus, 2008; Huang et al., 2010; Siddiqui et al., 2012; Lee et al., 2015; Qing et al., 2017a).

The spatial model for the macro-level can be expressed as:

$$y_i^{zone} \sim Poisson\ (\lambda_i^{zone}) \tag{7-1}$$

$$\log(\lambda_i^{zone}) = \beta^{zone} x_i^{zone} + \theta_i^{zone} + \phi_i^{zone} \tag{7-2}$$

where y_i^{zone} is the number of total crashes in zone i, λ_i^{zone} is the expected value of y_i^{zone}. $x_i^{zone}$is a set of explanatory variables while β^{zone} is the corresponding parameters. θ_i^{zone} is a random effect accounting for the unstructured over-dispersion that follows a normal distribution:

$$\theta_i^{zone} \sim N(0, \frac{1}{\tau_h}) \tag{7-3}$$

where τ_h is the precision parameter (the inverse of the variance) which follows a prior gamma (0.001, 0.001).

$\phi_i^{zone}$is a random effect term which is used to deal with the spatial autocorrelation among zones. ϕ_i^{zone}follows a normal distribution with CAR prior suggested by Besag et al. (1991):

$$\phi_i^{zone} \sim N(\frac{\sum_{i \neq j} w_{ij}^{zone} \phi_j^{zone}}{\sum_{i \neq j} w_{ij}^{zone}}, \frac{1}{\tau_c \sum_{i \neq j} w_{ij}^{zone}}) \tag{7-4}$$

in which w_{ij}^{zone} is the binary entries of proximity matrix with a value of 1 if zones i and j share border or 0 otherwise. τ_c is the precision parameter, which also follows a prior gamma (0.001,0.001).

The proportion of variability in the random effects due to spatial autocorrelation can be calculated as:

$$\alpha^{zone} = \frac{sd(\phi_i^{zone})}{sd(\theta_i^{zone}) + sd(\phi_i^{zone})} \tag{7-5}$$

where sd(·) represents the empirical marginal standard deviation function.

(2) Bayesian non-integrated spatial model at the micro-level

At the micro-level, road entities located in close proximity may also share similar factors, resulting in spatial autocorrelation of traffic crashes among road entities. Compared with solely spatial autocorrelation between segments or intersections, the spatial correlation effects between adjacent segments and intersections may be more significant if they are directly connected with each other. To this end, Zeng and Huang (2014) proposed a Bayesian spatial joint model that simultaneously analyzes the crash frequency of segments and intersections. The model introduced an indicator γ_m to distinguish whether a road entity is a segment or an intersection since the segments and intersections should have different exogenous factors affecting traffic safety. Specifically, the value of γ_m is 1 if road entity m is a segment and γ_m is 0 if the road entity is an intersection. Then, the model at micro-level is as follows:

$$y_m^{Entity} \sim Poisson\ (\lambda_m^{entity}) \tag{7-6}$$

$$\log\left(\lambda_m^{entity}\right) = \gamma_m \times \left(\beta^{seg} x_m^{seg} + log(length_m^{seg})\right) + (1-\gamma_m) \times \left(\beta^{inter} x_m^{inter}\right) + \theta_m^{entity} + \phi_m^{entity} \tag{7-7}$$

where y_m^{entity} is the number of crashes at segment or intersection m. x_m^{seg} and x_m^{inter} denote the set of explanatory variables of segments and intersections while β^{seg} and β^{inter} are the corresponding parameters. $\log(length_m^{seg})$ is logarithm of the length of road entity m if it is a segment, otherwise it is zero. Similar to the spatial model at the macro-level, θ_m^{entity} and ϕ_m^{entity} represent the two random effects which are used to account for the unstructured over-dispersion effect and spatial correlation effect, separately. The spatial random effect ϕ_m^{entity} is also assumed

to have a CAR prior. If two road entities m and n directly connect with each other the weight in the spatial proximity matrix w_{mn}^{entity} is set to be 1, otherwise, the weight is 0. This approach in the joint model can not only capture the spatial correlation of road entities of the same type but also the two different types of road entities including segments and intersections.

7.2.2 Bayesian integrated spatial model at the two levels

Figure 1 presents three GIS layers illustrating the spatial relations between crashes, road entities (micro-level), and zones (macro-level). As shown in Figure 7-1, the same crashes in the study area are aggregated at the macro- and micro-levels for the crash analyses. Hence, the crash count of a zone is supposed to be the same as the total crashes of all road entities in the same zone of interest. Let a matrix W denote the relation of spatial interaction between zones and road entities. The spatial interaction matrix w_{mi} is assigned a value of 1 if a road entity m is located in zone i or 0 otherwise. If $\hat{\imath}$ zones and $\hat{m}$ road entities included in the study, a $\hat{m} \times \hat{\imath}$ spatial dependence matrix can be generated. Then, the relation between observed crashes at the macro- and micro-levels can be expressed as follows:

$$y_i^{zone} = \sum_{m=1}^{k} y_m^{entity} w_{mi} \tag{7-8}$$

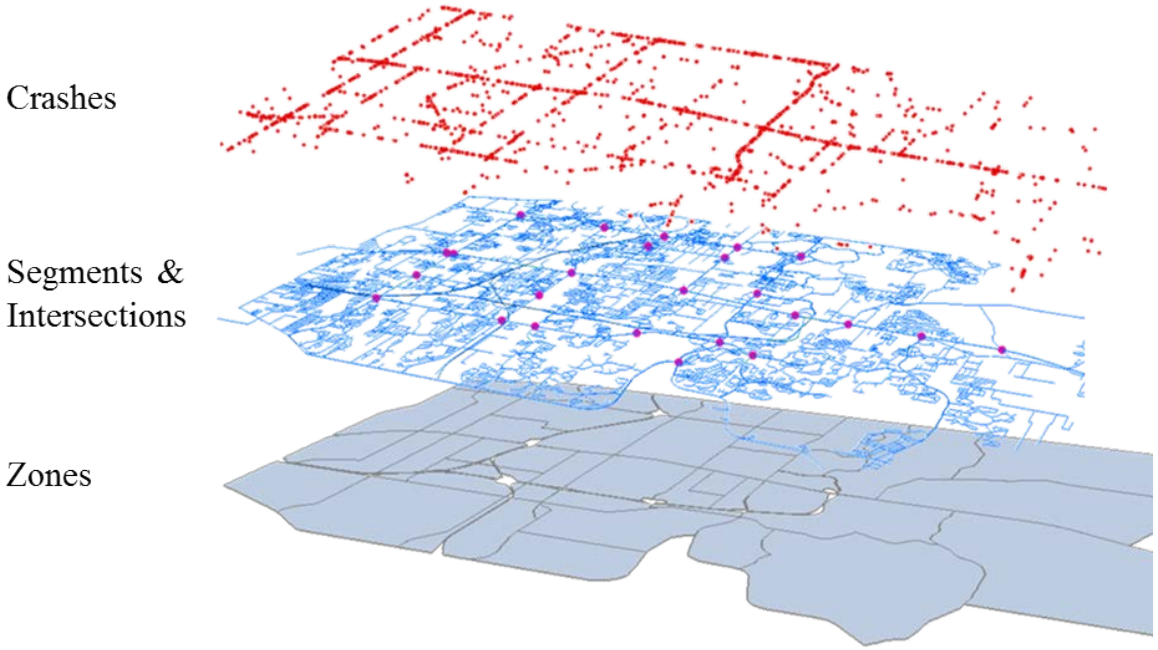


Figure 7-1 Illustration of spatial relation among crashes, road entities, and zones

Based on the equivalence relation presented in Equation (6-9), the non-integrated models for the macro- and micro-levels can be linked. However, the expected crash counts at the macro-level might not be the same as the total expected number of crashes at the micro-level since they are estimated at different levels with different explanatory variables. Therefore, an adjusted factor is introduced to relax the equivalence constraint. The link function between the macro- and micro-levels can be specified as:

$$u_i^{zone} = \sum_{m=1}^{k} \lambda_m^{entity} w_{mi} \tag{7-9}$$

$$\lambda_i^{zone} = u_i^{zone} \times ADJ_i \tag{7-10}$$

$$ADJ_i = exp(\beta'^{zone} x_i'^{zone} + \theta_i'^{zone} + \phi_i'^{zone}) \tag{7-11}$$

where u_i^{zone} is the total expected crashes (λ_m^{entity}) of all road entities in zone i and the λ_m^{entity} can be estimated based on the non-integrated spatial model at the micro-level (Equation (7)). ADJ_i is the adjustment factor of u_i^{zone} and λ_i is the expected number of crashes in zone i based on the

non-integrated spatial model at the macro-level (Equation (7-2)). The adjustment factor can represent that how many different crashes will happen in a zone given the same road network but with different socio-demographic characteristics. Hence, only macro-level socioeconomic variables are adopted for the estimation of the adjust factor ADJ_i. Also, $\theta_i'^{zone}$ and $\phi_i'^{zone}$ are two random terms to capture the unobserved and spatial autocorrelation effects at the macro-level. In the integrated approach, the expected crash counts of road entities (λ_m^{entity}) are estimated by equation (7) subjected to the relation with the crash count of zones shown in equations (7-9) and (10). Meanwhile, the expected crash frequencies of zones are the product of the total expected crash counts of all road entities and the adjustment factors (see equations (7-10) and (7-11)). Hence, based on the integrated model structure with Equations (7-1), (7-6)-(7-8), and (7-9)-(7-11), the crashes at the macro- and micro-levels can be investigated, simultaneously.

All the models were coded and estimated by using WinBUGS, which is a popular programming platform for Bayesian inference. The significant explanatory variables were determined based on 95% certainty of Bayesian credible intervals (BCIs). Deviance information criterion (DIC) was used to measure models' performance and determine the best set of parameters for each model. DIC is a common measurement for Bayesian model comparison and a lower DIC value is preferred. Roughly, differences of more than ten might indicate that the model with lower DIC performs better (El-Basyouny and Sayed, 2009).

7.3 Measurement of model comparison

Besides the DIC mentioned above, two additional measures were employed to compare the model performance at both the macro- and micro-levels. MAE (Mean Absolute Error) computes the mean of absolute errors with the following equation:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'| \tag{7-12}$$

where N is the number of observations, y_i and y_i' are the observed and predicted number of crashes of site *i* at the macro- and micro-levels.

Root Mean Squared Errors (RMSE) calculates the square root of the sum of the squared error divided by the number of observations as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} \tag{7-13}$$

7.4 Empirical data

Dataset were elaborately collected based on 78 TADs in Orlando, Florida to demonstrate the empirical application of the proposed model. In the same study area, totally 3,316 road entities including 2,434 segments and 882 intersections were identified for the analysis (Figure 7-2). It is noteworthy that there are more segments and intersections in the study area. Unfortunately, the traffic data were not available for all segments and intersections. Thus, only segments and

intersections with available traffic data were selected and crashes occurred on the selected road entities were aggregated at the macro- and micro- levels for the analysis. However, the proposed model can be easily extended to include all the crashes once all road entities have available explanatory data.

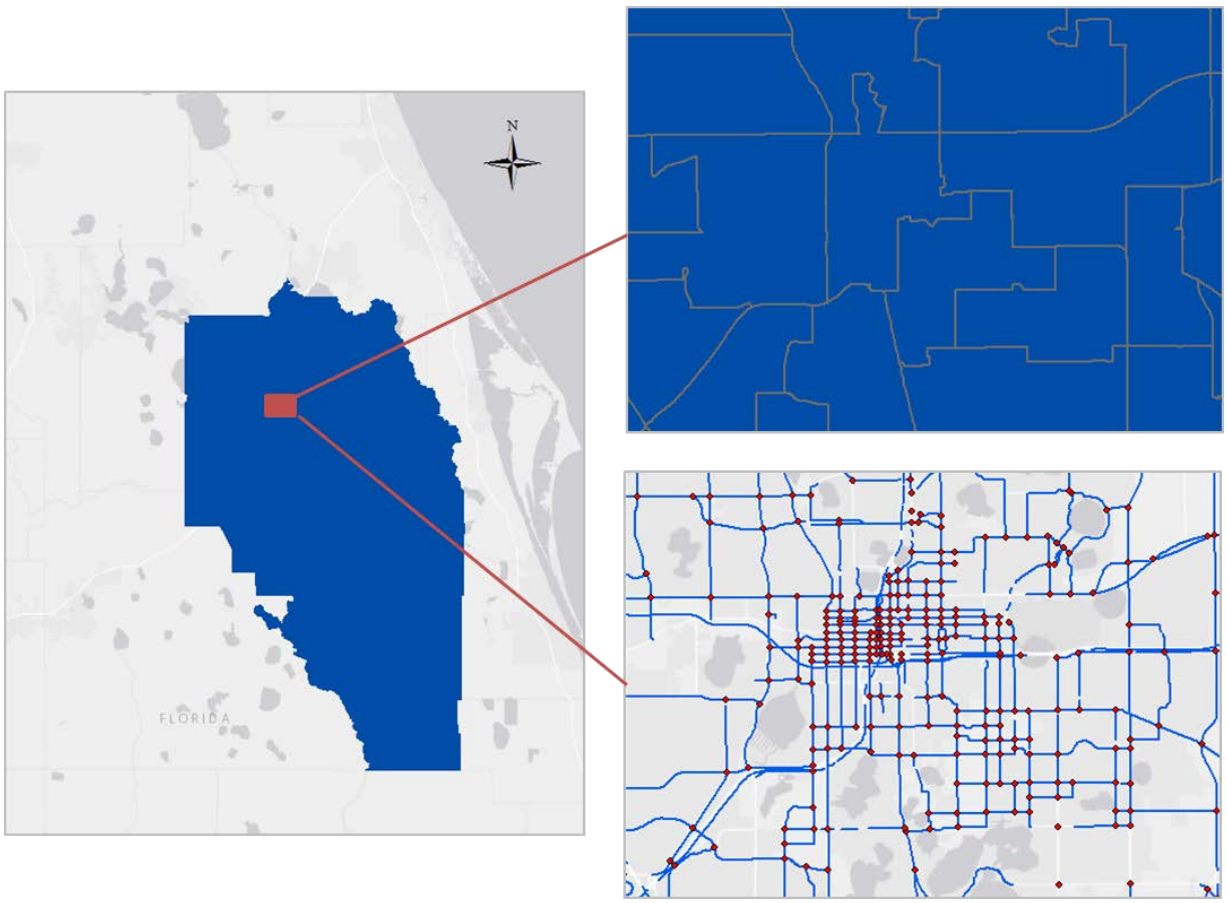


Figure 7-2 Selected TADs and road network in Orlando, Florida: overall study area (left); TADs (upper right) and road network (bottom right) in Downtown Orlando

The spatial interaction between TADs and road entities were processed by using ArcGIS 10.2 (ESRI) based on the digital maps provided by the U.S. Census Bureau (USCB) and Florida Department of Transportation (FDOT). As noted above, a lot of segments and intersections are located on the boundaries of TAZs since one of the zoning criteria of TAZs is to recognize physical boundaries such as arterial (Lee et al., 2014; Cai et al., 2017a) and the size of a TAZ is

quite small (on average 5.50 square miles in Orlando). However, the TADs were developed by combining the existing TAZs and the size of a TAD is sufficiently larger (on average 36.59 square miles). Hence, most of road entities could be located inside of TADs. If a road entity is located on the boundaries of two or more TADs, the geospatial method was applied to assign them into TADs. Specifically, each intersection was assigned into a TAD if the intersection is located within the digital boundary of the TAD. Meanwhile, each segment was allocated into a TAD if the segment is most proportionally in the corresponding TAD. Hence, the one-to-one spatial interaction between TADs (macro level) and road entities (micro level) can be obtained. A 3316×78 spatial dependence matrix can be generated corresponding to the 3316 road entities and 78 TADs. Also, the spatial autocorrelation matrix only for TADs or road entities can be obtained by applying spatial join features in ArcGIS. The descriptive statistics for the spatial relations are presented in Table 7-1. Remarkably, all TADs have adjacent TADs and each TAD has at least 5 road entities. Besides, the maximum number of neighbors among road entities is 21, which might be because some long segments connect a lot of intersections and other segments.

Table 7-1 Descriptive statistics for spatial relations

Variables	Definition	Mean	S.D.	Min.	Max.
Spatial autocorrelation between TADs					
N_TAD_NEI	Number of neighbors among TADs	5.80	1.55	2	10
Spatial autocorrelation between road entities					
N_ENTITY_NEI	Number of neighbors among road entities	3.03	2.09	0	21
Spatial dependence between TADs and road entities					
N_TAD_ENTITY	Number of road entities in each TAD	42.51	29.13	5	189

The crashes that occurred in Orlando during 2010-2012 were collected from the Florida Department of Transportation (FDOT)'s Crash Analysis Reporting System (CARS) and Signal Four Analytics (S4A) database. In the database, crashes occurring within 50 feet and 250 feet away from the intersection are defined as "crashes at intersection" and "crashes influenced by

intersection", respectively. According to this principle, a 250 feet buffer around each intersection were created and crashes in the buffers were collected and classified as intersection-related crashes while other crashes were categorized as segment-related crashes. Then, the crashes in each TAD can be obtained by summing up the crash counts of all road entities in the corresponding TAD according to the spatial interaction.

A host of explanatory variables were considered for the analysis, including traffic data, roadway, demographic, and socioeconomic factors. The traffic and road data in the road entities were first collected from FDOT and then spatially attached to the corresponding TADs in a similar way as crashes. The socio-demographic data were attained from the USCB. These census tracts-based data were aggregated to TADs since a TAD is a combination of multiple census tracts (Cai et al., 2017a). The descriptive statistics of the collected data based on TADs and road entities are summarized in Tables 7-2 and 7-3, respectively.

Table 7-2 Descriptive statistics of collected data for TADs (macro-level)

Variables	Definition	Mean	S.D.	Min.	Max.
CRASH	Three-year crash count for each TAD	257.03	213.17	18	1038
DVMT	Daily vehicle-miles traveled (in thousand)	494.53	440.19	23.30	2210.21
Segment-related variables					
ROAD_LENGTH	Total road length in each TAD (mi)	23.60	29.72	1.53	248.65
P_FREEWAY	Proportion of segment length of freeway	0.14	0.17	0	0.71
P_ARTERIAL	Proportion of segment length of arterial	0.40	0.21	0	0.74
P_COLLECTOR	Proportion of segment length of collector	0.46	0.22	0	1
P_LOCALROAD	Proportion of segment length of local road	0.01	0.03	0	0.23
P_LANE1_2	Proportion of segment length with 1 or 2 lanes	0	0.00	0	0.03
P_LANE3_4	Proportion of segment length with3 or 4 lanes	0.39	0.22	0	0.87
P_LANE5MORE	Proportion of segment length with 5 lanes or over	0.16	0.17	0	0.74
P_MEDIANROAD	Proportion of segment length having median	0.68	0.22	0.10	1
Intersection-related variables					
INTER_DENS	Number of intersections per mile (/mile)	1.70	0.57	1	4.33
P_SINGAL	Proportion of signalized intersections	0.78	0.24	0	1
P_LEG3	Proportion of intersections with 3 legs	0.32	0.17	0	0.73
P_LEG4	Proportion of intersections with 4 legs	0.67	0.18	0	1
Socio-demographic variables					
POP_DENS	Population density (in thousand)	2.38	1.49	0.02	6.56
P_AGE1524	Proportion of population aged 15-24	0.16	0.05	0.09	0.38
P_AGE65MORE	Proportion of population aged 65 or over	0.10	0.03	0.04	0.18
COMMUTERS_DENS	Commuters density (/mi^2)	1163.12	728.39	9.32	3103.77
MEDIAN_INC	Median household income (in thousand)	63.40	19.47	33.99	122.77
DIS_URBAN	Distance to the nearest urban area (mi)	1.40	1.71	1.00	14.12

Table 7-3 Descriptive statistics of collected data for road entities (micro-level)

Variables	Definition	Mean	S.D.	Min.	Max.
Segment variables					
CRASH	Three-year crash count for each segment	6.20	12.59	0	132
LENGTH	Segment length (mile)	0.75	1.35	0.10	30.91
AADT	Average annual daily traffic (in thousand)	20.19	25.51	0.20	195.77
FREEWAY	Freeway indicator: 1 if freeway, 0 otherwise	0.11	0.31	0	1
ARTERIAL	Arterial indicator: 1 if arterial, 0 otherwise	0.39	0.49	0	1
COLLECTOR	Collector indicator: 1 if collector, 0 otherwise	0.49	0.50	0	1
LOCALROAD	Local road indicator: 1 if local road, 0 otherwise	0.01	0.11	0	1
MEDIAN	Median barrier indicator: 1 if present, 0 otherwise	0.63	0.48	0	1
LANE1_2	1 or 2 lanes indicator: 1 if yes, 0 otherwise	0.56	0.50	0	1
LANE3_4	3 or 4 lanes indicator: 1 if yes, 0 otherwise	0.30	0.46	0	1
LANE5MORE	5 or more lanes indicator: 1 if yes, 0 otherwise	0.15	0.36	0	1
URBAN	Urban indicator: 1 if in urban area; 0 otherwise	0.93	0.26	0	1
Intersection variables					
CRASH	Three-year crash count for each intersection	16.86	20.34	0	135
MAJ_AADT	AADT on major approach (in thousand)	23.72	15.76	0.60	81.50
MIN_AADT	AADT on minor approach (in thousand)	8.22	7.64	0.20	52.50
TRAFFIC_SIGNAL	Traffic signal indicator: 1 if present, 0 otherwise	0.76	0.43	0	1
LEG3	3-Leg intersection indicator: 1 if yes, 0 otherwise	0.31	0.46	0	1
LEG4	4-Leg intersection indicator: 1 if yes, 0 otherwise	0.69	0.46	0	1
URBAN	Urban indicator: 1 if in urban area; 0 otherwise	0.99	0.10	0	1

7.5 Model Estimation

7.5.1 Model Comparison

As discussed above, three models were estimated in this study, i.e., (1) a non-integrated model for the macro-level, (2) a non-integrated model for the micro-level, and (3) an integrated model for both levels. Prior to discussing the model results, we present the performance results of the estimated models in Table 4. The table presents the DIC, MAE, and RMSE for the two levels based on the results of non-integrated and integrated models. Several observations can be made according to the results presented in Table 7-4. At the macro-level, the integrated model can provide significantly smaller values of the three measures compared with the non-integrated

model. Specifically, the DIC difference for macro-level is 44.99, which indicates significant difference between the two models (El-Basyouny and Sayed, 2009). Likewise, the prediction accuracy of crash frequency for macro-level in the integrated model is improved by 27.99% and 18.57% respectively based on the MAE and RMSE. On the other hand, the integrated model can provide significantly smaller DIC for the micro-level compared with the non-integrated model as well. Besides, the goodness-of-fit for the micro-level is improved by 21.16% and 23.33% according to the values of MAE and RMSE, respectively. Hence, in terms of the comparison results, we can generally conclude that the proposed integrated model is preferable for crash frequency analysis at both macro- and micro-levels with better overall statistical fit.

Table 7-4 Comparison results of model performance

Measure	Non-Integrated Model		Integrated Model		Difference between Models	
	Macro-level	Micro-level	Macro-level	Micro-level	Macro-level	Micro-level
DIC	798.83	17524.30	753.84	17506.60	44.99	17.70
MAE	161.41	10.16	116.23	8.01	45.18	2.15
RMSE	242.28	24.43	197.30	18.73	44.98	5.70

The model comparison results discussed above indicate that the proposed integrated model can improve the crash frequency prediction and analysis at the macro- and micro-levels. The findings are somewhat not surprising. At the macro-level, a possible explanation may be the less aggregated traffic and road variables from the micro-level were adopted for the zonal crashes estimation and the explanatory factors associated with the crash risk from the micro-level may be more direct and specific to crash circumstances (Huang et al., 2017). In comparison, the non-integrated model for the macro-level crash frequency analysis adopts a list of aggregated traffic and roadway variables from the micro-level together with socio-demographic variables based on

the macro-level. Hence, the non-integrated model for macro-level cannot consider the heterogeneity of different road entities since the potential variation is neutralized by the aggregation of data. At the micro-level, a possible reason is that the integrated model analyzes the crash frequency with the prior information from the macro-level, which indicates the total crash counts of TADs where the road entities are located. Meanwhile, the macro-level socio-demographic variables can affect the parameter estimation of micro-level variables through the adjusted factors which links crash frequencies of the two levels. In conclusion, the macro- and micro-level crash frequency models indeed support each other and the integrated model can consequently improve model performance for crash prediction and analyses at the two levels.

7.5.2 Model Results

The results of three models (i.e., two non-integrated models, one integrated model) for crashes at both macro- and micro-levels are displayed in Tables 7-5, 7-6, and 7-7. The results for two non-integrated models only present the variables with significant effects on crash frequency at either macro- level or micro-level. On the other hand, the integrated model results consist of two components: (1) significant variables affecting the crash counts at the macro- and micro-levels; and (2) other socio-demographic variables at the macro-level adjusting the relation of the expected crash counts between the two levels. All micro-level significant variables in the integrated model can also be found significant in the micro-level non-integrated model. Meanwhile, the same significant socio-demographic variables can be obtained from the integrated model and the non-integrated model for the macro-level. All the significant variables are found to have consistent signs of parameter estimates in the integrated and non-integrated models. While the results summarized in the three tables, the discussions about the parameter

estimates at the two levels focuses on the integrated model which has better fit and more significant variables.

As shown in Table 7-6, totally 8 micro-level variables are statistically significant for crash frequency with 95% BCIs: 5 segment-related variables (i.e., AADT (average annual daily traffic), functional class is arterial, number of lanes is 1 or 2, presence of median barrier) and 4 intersection-related variables (i.e., AADT on major approach, AADT on minor approach, presence of traffic signal, number of legs is 3). The AADTs of segments and intersections are used as exposure variables of the crash frequency and expected to have positive effects on crashes. Compared with other road types, arterials have partially limited accesses with comparatively higher traffic volumes. Given the same road length, the arterial is supposed to have more traffic interactions and conflicts. Unsurprisingly, a road segment will have fewer crashes if it only has one or two lanes. The presence of median barriers will increase crash counts on the road segments, which is consistent with the previous studies (Anastasopoulos et al., 2012). As for the intersections, a variable related to the intersection control type and a variable about number of legs are found significant. Intersections with signalized controls are more likely to have more crashes. The signal control is usually installed at intersections with higher traffic volumes where more traffic interactions occur (Wang et al., 2016). Also, the existence of dilemma zones can lead to more crashes at the signalized intersections (Wu et al., 2015). More crashes are prone to happen at intersections with more intersecting legs (Wang and Huang, 2016). Hence, the 3-leg intersection indicator is negatively associated with the crash frequency.

As for the macro-level socio-demographic variables, the proportion of population aged 15-24 is positive while the median household income and distance to the nearest urban area are negatively

associated with crash counts for the macro-level crash counts. The finding about the young drivers is consistent with the well-known fact that young drivers prone to be involved in crashes due to the lack of driving experience (Huang et al., 2010).TADs having higher median household income would experience less traffic crashes since drivers and passengers with higher income are more likely to use seatbelts (Lerner et al., 2001) and their vehicles tend to be safer (Girasek and Taylor, 2010). As the distance of the TAD centroid from the nearest urban region increases, total traffic crash risk is reduced - a sign of low traffic exposure in the suburban regions.

The two random terms due to the spatial autocorrelation and unobserved heterogeneity are significant for crash frequency of both macro- and micro-levels. The proportions of variability due to the spatial autocorrelation at the macro- and micro-levels are 0.65 and 0.6, respectively, indicating the importance to consider the spatial effects in crash frequency analysis. Compared with the non-integrated model, the standard deviations of the spatial autocorrelation and unobserved heterogeneity for the crash frequency at the macro- and micro-levels are much smaller in the integrated model, which indicates that considering the spatial interaction between the two levels can reduce the effects of random terms.

Table 7-5 Non-Integrated model result at macro level

Variable	Definition	Mean	S.D.	BCI	
				2.50%	97.50%
Intercept		-3.33	0.09	-3.47	-3.14
DVMT	Daily vehicle-miles traveled	0.91	0.01	0.90	0.92
Segment-related variables					
P_ARTERIAL	Proportion of segment length of arterial	0.66	0.11	0.44	0.85
Intersection-related variables					
INTER_DENS	Number of intersections per mile	0.58	0.11	0.35	0.78
P_SINGAL	Proportion of signalized intersections	0.40	0.13	0.21	0.67
Socio-demographic variables					
P_AGE1524	Proportion of population aged 15-24	2.70	0.30`	2.06	3.27
MEDIAN_INC	Median household income	-0.29	0.01	-0.31	-0.28
DIS_URBAN	Distance to the nearest urban area	-0.21	0.06	-0.33	-0.10
Random effects					
$sd[\theta^{zone}]$	Standard deviation of θ^{zone}	0.11	0.05	0.03	0.20
$sd[\phi^{zone}]$	Standard deviation of ϕ^{zone}	0.36	0.02	0.30	0.40
α_{zone}	Proportion of variability due to spatial correlation	0.78	0.09	0.62	0.93

Table 7-6 Non-Integrated model result at micro level

Variable	Definition	Mean	S.D.	BCI	
				2.50%	97.50%
Segment					
Intercept		-3.34	0.09	-3.47	-3.21
AADT	Average annual daily traffic	0.55	0.01	0.54	0.56
ARTERIAL	Arterial indicator: 1 if arterial, 0 otherwise	0.27	0.03	0.22	0.34
LANG1_2	1 or 2 lanes indicator: 1 if yes, 0 otherwise	-0.41	0.03	-0.47	-0.34
MEDIAN	Median barrier indicator: 1 if present, 0 otherwise	0.11	0.03	0.05	0.16
Intersection					
Intercept		-8.18	0.08	-8.35	-7.99
MAJ_AADT	AADT on major approach	0.75	0.01	0.74	0.76
MIN_AADT	AADT on minor approach	0.29	0.01	0.27	0.31
TRAFFIC_SIGNAL	Traffic signal indicator: 1 if present, 0 otherwise	0.45	0.04	0.38	0.53
LEG3	3-Leg intersection indicator: 1 if yes, 0 otherwise	-0.51	0.04	-0.59	-0.42
Random effects					
$sd[\theta^{entity}]$	Standard deviation of ϕ^{entity}	2.73	0.17	2.40	3.07
$sd[\phi^{entity}]$	Standard deviation of θ^{entity}	3.90	0.41	3.22	4.83
α_{entity}	Proportion of variability due to spatial correlation	0.79	0.02	0.75	0.83

Table 7-7 Integrated model result at the two levels

Variable	Definition	Mean	S.D.	BCI	
				2.50%	97.50%
Segment-related variables					
Intercept		-2.87	0.05	-2.95	-2.80
AADT	Average annual daily traffic	0.48	0.01	0.47	0.49
ARTERIAL	Arterial indicator: 1 if arterial, 0 otherwise	0.31	0.03	0.27	0.38
LANG1_2	1 or 2 lanes indicator: 1 if yes, 0 otherwise	-0.43	0.03	-0.48	-0.36
MEDIAN	Median barrier indicator: 1 if present, 0 otherwise	0.19	0.04	0.12	0.24
Intersection-related variables					
Intercept		-7.96	0.06	-8.06	-7.87
MAJ_AADT	AADT on major approach	0.74	0.01	0.72	0.76
MIN_AADT	AADT on minor approach	0.29	0.01	0.27	0.30
TRAFFIC_SIGNAL	Traffic signal indicator: 1 if present, 0 otherwise	0.45	0.06	0.35	0.57
LEG3	3-Leg intersection indicator: 1 if yes, 0 otherwise	-0.54	0.04	-0.62	-0.46
Socio-demographic variables for adjusted factor					
Intercept		3.62	0.07	3.49	3.75
P_AGE1524	Proportion of population aged 15-24	0.92	0.32	0.32	1.41
MEDIAN_INC	Median household income	-0.34	0.01	-0.35	-0.33
DIS_URBAN	Distance to the nearest urban area	-0.11	0.02	-0.16	-0.06
Random effects					
$sd[\theta^{entity}]$	Standard deviation of ϕ^{entity}	0.60	0.02	0.56	0.64
$sd[\phi^{entity}]$	Standard deviation of θ^{entity}	0.92	0.04	0.87	1.01
$sd[\theta^{zone}]$	Standard deviation of θ^{zone}	0.07	0.02	0.03	0.12
$sd[\phi^{zone}]$	Standard deviation of ϕ^{zone}	0.10	0.03	0.04	0.14
α_{Entity}	Proportion of variability due to spatial correlation at micro level	0.61	0.01	0.58	0.63
α_{zone}	Proportion of variability due to spatial correlation at macro level	0.57	0.15	0.25	0.81

7.6 Integrated Hotspots Identification Analysis

One possible application of the proposed integrated model is to identify crash hotspot, which is a top priority for safety treatment. The crash hotspot should not be simply the one with the highest crash frequency; instead, it should be the one that experiences more crashes than similar sites as a result of site-specific deficiency (Xie et al., 2017). A potential for safety improvement (PSI) was adopted in this study to identify hotspots, which is defined as the expected crash frequency

at the sites of interest minus the expected crashes in the similar sites (Aguero-Valverde and Jovanis, 2010). The spots with higher PSI are expected to have more reduced crashes after the implementation of the treatments. Based on the integrated spatial model, the PSIs for the two levels can be calculated as:

$$EXP_m^{entity} = \gamma_m \times \left(\beta^{seg} x_m^{seg} + log(length_m^{seg})\right) + (1-\gamma_m) \times \left(\beta^{inter} x_m^{inter}\right) \quad (7\text{-}14)$$

$$PSI_m^{entity} = \lambda_m^{entity} - EXP_m^{entity} \quad (7\text{-}15)$$

$$EXP_i^{zone} = \sum_{m=1}^{k} EXP_m^{entity} w_{mi} * \exp(\beta''^{zone} x_i'^{zone}) \quad (7\text{-}16)$$

$$PSI_i^{zone} = \lambda_i^{zone} - EXP_i^{zone} \quad (7\text{-}17)$$

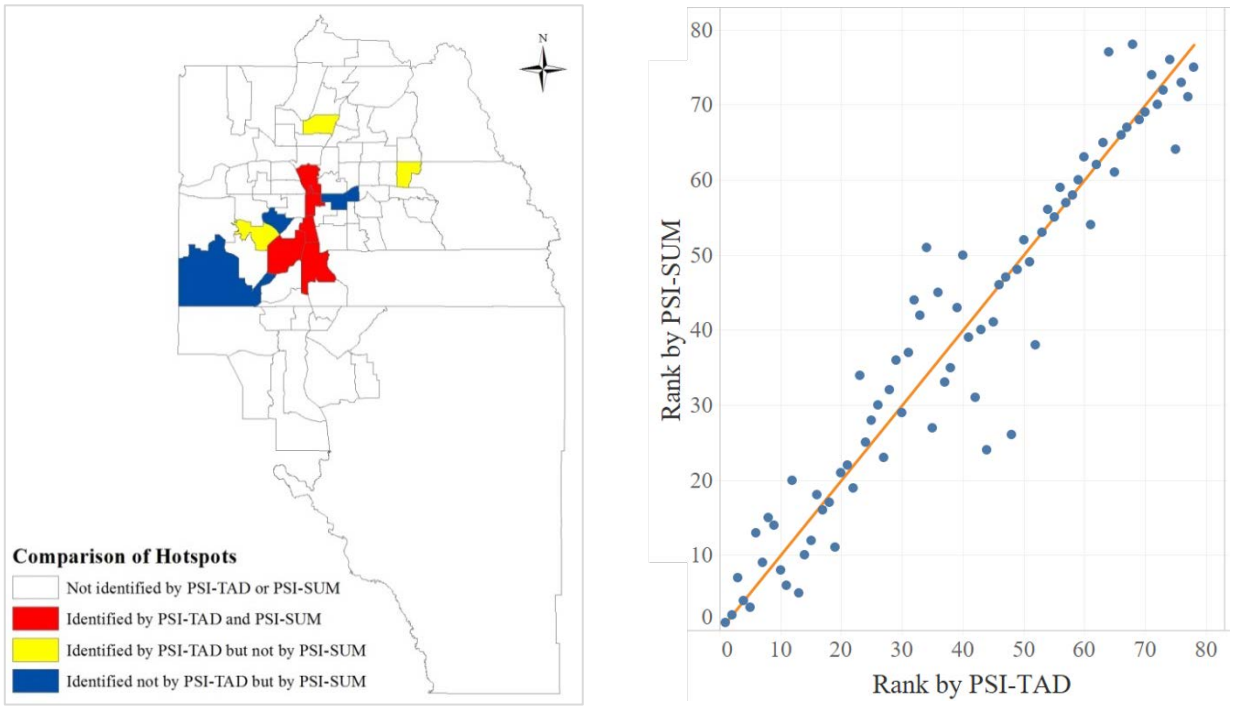
where EXP_m^{entity} and EXP_i^{zone} are the expected number of crashes at micro and macro levels while λ_m^{entity}, λ_i^{zone} are the predicted number of crashes at the two levels. PSI_m^{entity} and PSI_i^{zone} are the micro and macro PSIs. The coefficients and random terms in the equations can be obtained by Bayesian inference in the estimated model. The spots with positive PSIs could be considered as hazardous and should have the potential to be improved. However, given time and budget constraints, it is more efficient to identify hotspots which have the priority to implement treatments. In our study, all sites at the macro- and micro-levels are classified into three categories based on the calculated PSIs: hot (H), warm (W), and cold (C) sites. Hot sites are defined as those with top 10% PSIs, warm sites refer to be sites with positive PSIs but not the top 10%, and the remaining sites are cold sites. It should be noted that 10% was commonly used as the threshold to identify hotspots (Cheng and Washington, 2008; Cai et al., 2017b), and it can be increased or decreased depending on researchers' needs.

The macro- and micro-level PSIs should recognize transportation safety problems with different aspects. In favor of providing an equivalent comparison of PSIs at the macro- and micro-levels,

the PSIs at the micro-level are aggregated into the macro-level. Figure 7-3(a) shows the difference between the hot TADs identified by PSIs based on the macro-level (PSI-TAD) and sum of PSIs based on the micro-level (PSI-SUM). In summary, 5 (6.41%) TADs were identified as hotspots by both the PSI-TAD and PSI-SUM, 3 (3.85%) TADs were identified by PSI-TAD only, and 3 (3.85%) TADs were identified by PSI-SUM only. As indicated in Figure 3(a), spatial clustering of high-risk TADs can be observed. Most of the identified hot TADs are located in the downtown Orlando area, especially hot TADs identified by both PSI-TAD and PSI-SUM. Figure 3(b) illustrates the difference between the ranks by PSI-TAD and PSI-SUM. The X- and Y- axis show the rank in descending order of the PSI-TAD and PSI-SUM. The red line is the 45-degree reference line and the points on the red line represent that same ranking results can be obtained based on PSI-TAD and PSI-SUM. As shown in Figure 7-3(b), most of points are plotted around the reference line indicating that similar ranking results are obtained based on the PSIs at the two levels. However, some TADs have clearly different ranking results based on PSI-TAD and PSI-SUM, revealing that the hotspots identification based on single level may result in largely ignoring certain spots with excess crash frequency studies (Abdel-Aty et al., 2016; Huang et al., 2016). Hence, it is necessary to develop an integrated approach to identify hotspots to overcome the shortcomings of individual identification analysis.

At the macro-level, an integrated classification is suggested based on TADs to support policy making and long-term transportation planning. Given that three categories are adopted for the classification at the two levels, there are nine candidate combination classifications: HH, HW, HC, WH, WW, WC, CH, CW, and CC. The former letter represents the safety at the macro-level while the latter letter denotes the combined crash risk based on the micro-level. For example, the ‘HH’ refers the TADs with serious safety problem at both macro- and micro-levels. Table 7-7

summarizes the number of TADs by the integrated category and only 6 classifications can be obtained for the 78 TADs. There are 5 (6.41%) TADs are classified as ‘HH’ which are the top priority for safety treatments since they have highest safety risks at the two levels. The integrated classification result is illustrated in Figure 5. Since the number of ‘HW’ and ‘WH’ TADs are small, they are merged together for the purpose of brevity. Hence, five categories are presented, i.e., ‘HH’, ‘HW/WH’, ‘WW’, ‘WC’, and ‘CC’. As demonstrated in Figure 7-4, spatial clustering of high-risk zones can be observed. Special attention should be paid in Downtown Orlando since most of zones with high crash risk are located in this area. The zones with moderate crash risk cluster in the north corner of the study area while the safe zones are rather spatially isolated.



(a) **(b)**

Figure 7-3 Comparisons of hot TADs identified by PSI at macro and micro levels

Beside integrated classification at the macro-level, an integrated classification analysis is also conducted at the micro-level to help provide appropriate engineering treatments to reduce crashes in specific road entities. Similar to the macro-level integration approach, all sites (segments and intersections) are classified into nine categories including two scale groups (micro and macro) and three risk levels (hot, warm, and cold). Hence, for example, the 'HH' indicates that a road entity has safety problem and it is located in a TAD with serious safety issues. For such road entity, both appropriate engineering treatments and enforcement strategies should be implemented. As summarized in Table 7-8, most road entities with high risk are in the dangerous area. Moreover, Figure 7-5 presents which road entities should be targeted in downtown Orlando since the area has most zones of interest.

Table 7-8 TADs and road entities by integrated category7

Sites	Category	HH	HW	WH	HC	CH	WW	WC	CW	CC
TAD	Counts	5	3	3	0	0	49	1	0	17
	Percentage	6.41%	3.85%	3.85%	0.00%	0.00%	62.82%	1.28%	0.00%	21.79%
Intersection	Counts	23	74	142	1	26	356	84	114	62
	Percentage	2.61%	8.39%	16.10%	0.11%	2.95%	40.36%	9.52%	12.93%	7.03%
Segments	Counts	83	146	295	5	34	913	249	397	312
	Percentage	3.41%	6.00%	12.12%	0.21%	1.40%	37.51%	10.23%	16.31%	12.82%

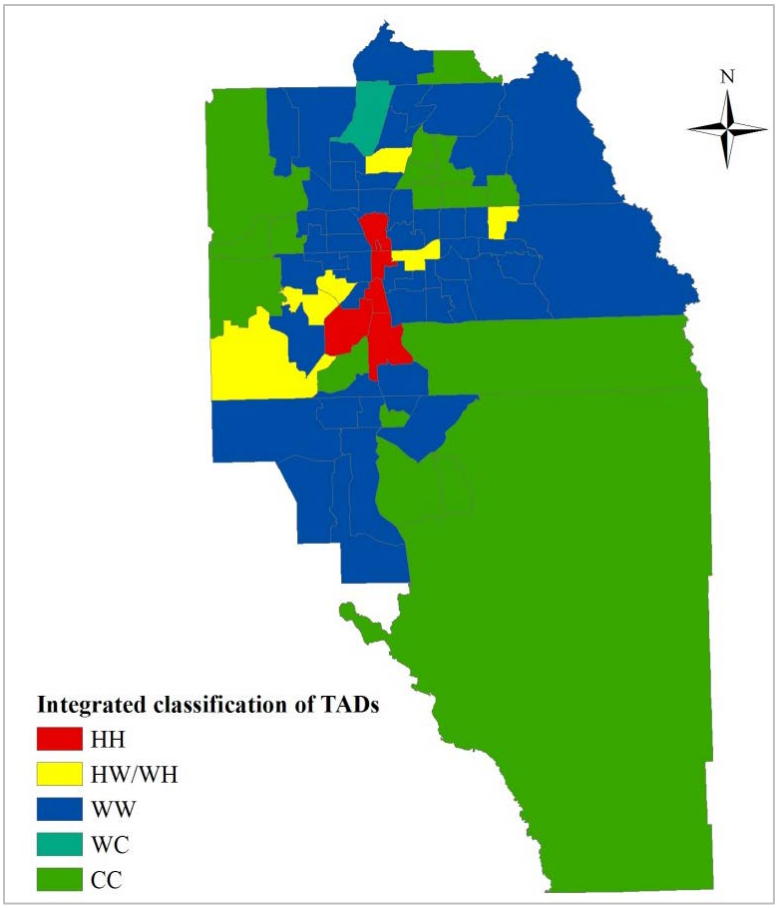


Figure 7-4 Spatial distribution of hot TADs based on the integrated classification

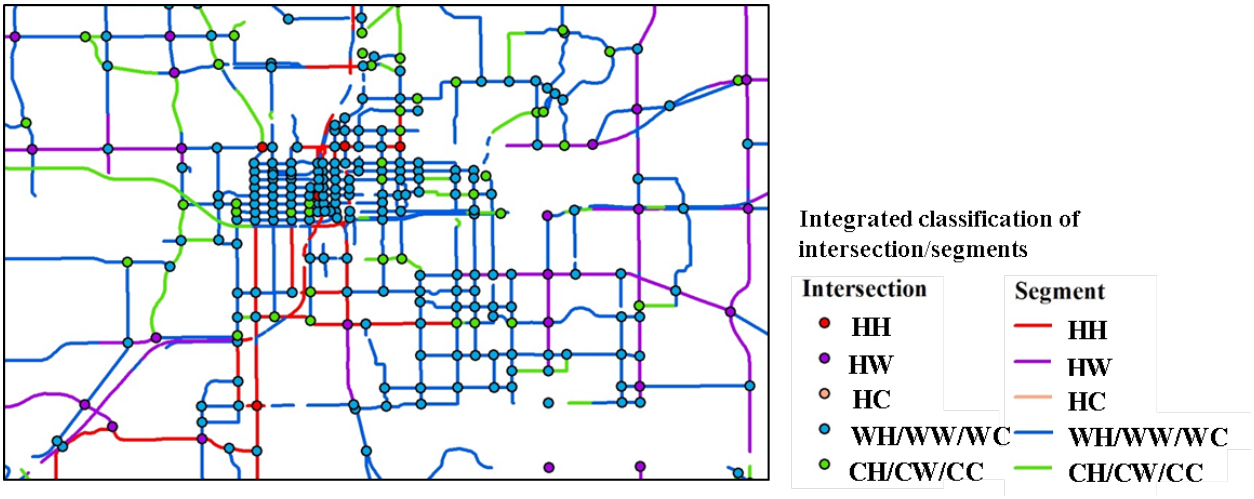


Figure 7-5 Spatial distribution of road entities based on the integrated classification in Downtown Orlando

7.7 Summary and Conclusion

The crash frequency modeling analysis plays an essential role in transportation safety as it can estimate the effects of macro- and micro-level factors on safety and identify hotspots, which have safety issues. This study formulated and estimated a Bayesian integrated spatial model to analyze crash frequency at the macro- and micro-levels, simultaneously. Based on the spatial interaction between zones and road entities, the expected crash counts at the macro- and micro-levels were linked by an adjustment factor. The adjustment factor was estimate by using a set of macro-level socio-demographic variables, which indicates how many more crashes occur at the macro-level given the same road network but with the different socio-demographic characteristics. Besides the spatial interaction, the spatial autocorrelations at zones and road entities were considered in the model. Especially, the spatial autocorrelation at micro-level was considered for different types of road entities (i.e., segments and intersections) with a joint structure. Two independent non-integrated models were also estimated for comparison. The crashes that occurred on both segments and intersections in Orlando, Florida during 2010-2012 were selected for the empirical analysis. Then, the selected crashes were aggregated at both macro- and micro-levels and a comprehensive set of exogenous variables from the two levels were selected for the model estimation.

The results of the integrated model clearly highlighted the existence of spatial interaction between the macro- and micro-level crash counts and confirmed the benefit of integrating modeling analysis of crash counts for the two levels. The comparison results indicated that the integrated model significantly outperformed non-integrated model at the macro-level while the integrated model provided a slightly better model performance for micro-level crash frequency

analysis. The integrated model provided a combination of significant variables from both micro- and macro-levels including segment-based variables (e.g., AADT, arterial indicator, 1 or 2 lanes indicator), intersection-based variables (e.g., AADT on major and minor approaches, traffic signal control indicator), and TAD-based socioeconomic variables (e.g., proportion of population aged 15-24, median household income). The identification of significant macro-level variables can help undertake planning process to enhance transportation safety while we can suggest engineering solution to reduce traffic crashes based on micro-level contributing factors. Therefore, the proposed model can be employed as a useful tool that links the transportation safety planning and traffic engineering countermeasures.

This study further contributed to the literature by proposing a novel integrated method to identify hotspots of crashes at both macro- and micro-levels. The PSI was adopted as a measure to identify the hotspots for the two levels. The macro-level hotspot identification can detect zones with area-wide planning-level safety problems while the micro-level approach is capable of identifying specific road entities with high risks. Since the sole hotspot identification may ignore certain spots with excess crash frequency, an integrated hotspot identification approach was suggested. Both TADs and road entities were classified into nine categories with the consideration of two levels (macro- and micro-levels) and three crash risk levels (hot, warm, and cold). With the integrated hotspot identification approach, better classification results can be obtained for both TADs and road entities with a comprehensive transportation planning and traffic engineering perspectives.

CHAPTER 8: CONCLUSIONS

8.1 Summary

This dissertation mainly focused on the crash frequency analysis at both the macroscopic and microscopic levels. The main objectives of this study are to 1) suggest statistical methodologies to improve macroscopic traffic safety analysis, 2) determine the optimal zonal system for macro-level crash analysis, 3) investigate macro-level effects on the crashes at segments and intersections, and 4) develop an integrated model to simultaneously analyze macroscopic and microscopic crashes.

The study in Chapter 3 contributes to safety literature by conducting a macro-level analysis for pedestrian and bicycle crashes at the traffic analysis zone (TAZ) level. The study considers both single-state (negative binomial (NB)) and dual-state count models (zero-inflated negative binomial (ZINB) and hurdle negative binomial (HNB)) for analysis. In addition, the research proposes the consideration of spatial spillover effects of exogenous variables from neighboring TAZs. The model development exercise involved estimating 6 model structures each for pedestrians and bicyclists. These include NB models with and without spatial effects, ZINB models with and without spatial effects and HNB models with and without spatial effects. The model comparison exercise for pedestrians and bicyclists highlighted that models with spatial spillover effects consistently outperformed the models that did not consider the spatial effects. Across the three models with spatial spillover effects, the ZINB model offered the best fit for pedestrian and bicyclists. The model results clearly highlighted the importance of several variables including traffic (such as VMT and heavy vehicle mileage), roadway (such as

signalized intersection density, length of sidewalks and bike lanes, etc.) and socio-demographic characteristics (such as population density, commuters by public transportation, walking and cycling) of the targeted and neighboring TAZs.

In Chapter 4, a new method for the comparison between different zonal systems for macro-level crash analysis was suggested by adopting grid structures of different scales. The Poisson lognormal (PLN) models without and Poisson lognormal conditional autoregressive model (PLN-CAR) with consideration of spatial correlation for total, severe, and non-motorized mode crashes were developed based on census tracts (CTs), traffic analysis zones (TAZs), and a newly developed traffic-related zone system - traffic analysis districts (TADs). Based on the estimated models, predicted crash counts for the three zonal systems were computed. Considering the average area of each geographic unit, ten sizes of grid structures with dimensions ranging from 1 mile to 100 square miles were created for the comparison of estimated models. The observed crash counts for each grid were directly obtained with GIS while the different predicted crash counts were transformed into the grids that each geographic unit intersects with. The weighted mean absolute error (MAE) and root mean square error (RMSE) were calculated for the observed and different transformed crash counts of different grid structures. By comparing the MAE and RMSE values, the best zonal system as well as model for macroscopic crash modeling can be identified with the same sample size. The comparison results indicated that the models based on TADs offered the best fit for all crash types. Based on the modeling results and the motivation for developing the different zonal systems, it is recommended TADs for transportation safety planning. Also, the comparison results highlighted that models with the consideration of spatial effects consistently performed better than the models that did not consider the spatial effects. The modeling results based on different zonal systems had different significant variables, which

demonstrated the zonal variation. Besides, the results clearly highlighted the importance of several explanatory variables such as traffic (i.e., VMT and heavy vehicle mileage), roadway (e.g., proportion of local roads in length, signalized intersection density, and length of sidewalks, etc.) and socio-demographic characteristics (e.g., population density, commuters by public transportation, walking as well as cycling, median household income, etc.).

Chapter 5 conducted a further study about pedestrian and bicycle crashes based on traffic analysis districts (TADs), which are suggested as the optimal geographic units for crash analysis in Chapter 4. This paper formulated and estimated models based on count and proportion models to investigate the effects of exogenous factors on pedestrian and bicycle crashes at the Traffic Analysis District (TAD) level in Florida. In order to identify potentially different impacts of exogenous variables on vehicle drivers and non-motorists, a joint model combining the negative binomial (NB) model and the logit model was suggested. More specifically, the NB model part is for the total crash counts to explore the effects on vehicle drivers while the logit model part is for the proportion of non-motorist crashes to investigate the influences on non-motorists. The model was estimated employing a comprehensive set of exogenous variables: traffic measures, roadway information, socio-demographic characteristics, and commuting variables. Also, a traditional NB model was developed and compared with the joint model. The results of the joint model obviously highlighted the existence of different impact of exogenous factors on drivers and non-motorists for pedestrian and bicyclist crashes. The model comparison indicates that the proposed joint model can provide better performance over the NB model. In addition, more significant variables such as signalized intersection density and proportion of population age 65 or over could be observed in the proposed model. Moreover, the result of the joint modeling emphasized that the importance of several other variables including traffic (e.g., VMT, proportion of heavy

vehicle mileage, etc.), roadway (e.g., length of local road, length of sidewalk, etc.), socio-demographic characteristics (e.g., population density, median household income, etc.), and commuting variables (e.g., commuters by public transportation and those by bicycle). To provide a clear quantitative comparison of the variables' impact, elasticity effects for the NB and joint models are computed. The results revealed that the same significant variables in the two models would have the same signs of elasticity effects on the non-motorist crashes. Also, the elasticity effect calculation allows us to determine the factors that substantially increase crash risk for crashes involving pedestrians and bicyclists.

In Chapter 6, crash frequency analysis was conducted at the micro-level for both segments and intersections. A Bayesian hierarchical model was proposed to investigate the potential macro-level effects on crashes at the micro-level. Macro-level factors including both macro-level explanatory variables such as socio-economic characteristics and the total crashes aggregated at macro-level were employed for the micro-level crash analysis. Besides, a joint modeling structure was introduced for the potentially spatial autocorrelation between segments and intersections. The results clearly suggested that considering macro-level effects can improve the model performance for micro-level crash analysis. The proposed model considering both macro-level explanatory and total crashes of zones could further enhance the model performance. A set of variables from both macro- and micro-levels were found significant for crashes at segments and intersections including segment-based variables (e.g., AADT, arterial indicator, 1 or 2 lanes indicator), intersection-based variables (e.g., AADT on major and minor approaches, traffic signal control indicator), and macro-level variables (e.g., proportion of segment length of arterial, intersection density, proportion of population aged 15-24, median household income).

In Chapter 7, an integrated study was conducted at both the macro- and micro-levels. This study formulated and estimated a Bayesian integrated spatial model to analyze crash frequency at the macro- and micro-levels, simultaneously. Based on the spatial interaction between zones and road facilities, the expected crash counts at the macro- and micro-levels were linked by an adjustment factor. The adjustment factor was estimate by using a set of macro-level socio-demographic variables, which indicates how many more crashes occur at the macro-level given the same road network but with the different socio-demographic characteristics. The results of the integrated model clearly highlighted the existence of spatial interaction between the macro- and micro-level crash counts and confirmed the benefit of integrating modeling analysis of crash counts for the two levels. The comparison results indicated that the integrated model significantly outperformed non-integrated model for crash frequency analysis at both the macro- and micro-level. Subsequently, a novel integrated method to identify hotspots of crashes at the two levels. Both TADs and road facilities were classified into nine categories with the consideration of two levels (macro- and micro-levels) and three crash risk levels (hot, warm, and cold). With the integrated hotspot identification approach, better classification results can be obtained for both TADs and road facilities with a comprehensive transportation planning and traffic engineering perspectives.

It would be useful to note that the method to integrate the macro-level effect in micro-level crash analysis proposed in Chapter 6 could be also regarded as an integrated modeling analysis at the two levels. From the model performance in Chapter 6 and 7, it is indicated that the method suggested in Chapter 6 could provide better analysis result for micro-level crash analysis, which is expected since more macro-level factors will be used for micro-level crash analysis.

8.2 Implications

The findings from Chapter 3 suggest that the dual-state models are appropriate to analyze macro-level crashes with excess zeros. Although several researchers questioned the basic dual-state assumption for crash occurrence and have conducted analysis at the micro-level, which indicated that the development of models with dual-state process is not consistent with crash data at the micro-level. However, based on the results in Chapter 3, dual-state models should be applicable for macro-level crashes if excess zeros exist. With the appropriate model adopted, the importance of several variables for pedestrian and bicycle crashes were revealed including traffic (such as VMT and heavy vehicle mileage), roadway (such as signalized intersection density, length of sidewalks and bike lanes, etc.) and socio-demographic characteristics (such as population density, commuters by public transportation, walking and cycling) of the targeted and neighboring TAZs. Besides, this study suggested consideration of exogenous variables from neighboring zones for accounting for spatial autocorrelation. This approach, referred to as spatial spillover model, is easy to implement and allows practitioners to understand and quantify the influence of neighboring units on crash frequency.

Chapter 4 has important implications for both researchers and practitioners. First, a novel method was suggested to compare different zonal system for macro-level crash frequency analysis. One of difficulties is to compare models based on different geographic units of which number of zones is not the same. This study proposes an innovative method for the comparison between different zonal systems by adopting a grid based framework. The number of grids remains the same for all models based on different zonal systems thereby providing a common comparison platform. Second, this study recommended traffic analysis districts (TADs), which are newly

developed traffic-related geographic units by aggregating existing traffic analysis zones, for researchers and practitioners to analyze crashes.

Chapter 5 also carries two important implications for traffic safety researchers and practitioners: First, this study contributed to the study on pedestrian and bicycle safety by suggesting a joint model to explore exogenous factors effecting pedestrian and bicycle crashes at the macroscopic level. The proposed joint model could analyze pedestrian and bicycle crashes with a new perspective. Specifically, the results of the proposed joint model can identify potentially different impacts of exogenous variables on vehicle drivers and non-motorists. It is supposed that more efficient countermeasures can be suggested to enhance pedestrian and bicycle safety since more significant variables can be detected with more detailed information. Second, the joint screening results could reveal hot zones for non-motorists into three types: hot zones with more dangerous driving environment only, hot zones with more hazardous walking and cycling conditions only, and hot zones with both. Hence, the joint screening method could help decision makers, transportation officials, and community planners more proactively improve pedestrian and bicyclist safety.

Chapter 6 conducted crash analysis at the micro-level, and suggested that considering macro-level data for micro-level crash analysis could improve modeling performance and reduce the variance of random effects. Besides, more accurate models can be developed at the micro-level if both macro-level explanatory variables and total crashes aggregated based on zones are employed. Finally, although many studies considered spatial effects at the micro-level, few studies have considered the potentially spatial autocorrelation between segments and their connected intersections. The result in this chapter clearly suggested that spatial correlations exist

among segments and intersection, suggesting the employment of the joint modeling structure to analyze traffic safety for various types of road facilities.

Chapter 7 provides many essential implications for traffic safety researchers. An innovative integrated model was suggested, which firstly linked the macroscopic and microscopic crash analysis. It was indicated that the crash analysis at the two levels can support each other. In other words, better analysis results by the integrated approach for both macro- and micro-levels. Besides, the integrated model revealed a combination of significant variables from both micro- and macro-levels including segment-based variables (e.g., AADT, arterial indicator, 1 or 2 lanes indicator), intersection-based variables (e.g., AADT on major and minor approaches, traffic signal control indicator), and TAD-based socioeconomic variables (e.g., proportion of population aged 15-24, median household income). The identification of significant macro-level variables can help undertake planning process to enhance transportation safety while we can suggest engineering solution to reduce traffic crashes based on micro-level contributing factors. Therefore, the proposed model can be employed as a useful tool that links the transportation safety planning and traffic engineering countermeasures. In addition, the results at the micro-level further suggested, as highlighted in Chapter 6, that segments and intersections are spatially correlated. Finally, the integrated screening approach can provide a comprehensive perspective by balancing macroscopic and microscopic screening results. With the integrated screening approach, better classification results can be obtained for both macroscopic and microscopic levels with a comprehensive transportation planning and traffic engineering perspectives.

REFERENCE

Abdel-Aty, M., Ekram, A.-A., Huang, H., Choi, K., 2011a. A study on crashes related to visibility obstruction due to fog and smoke. Accident Analysis & Prevention 43 (5), 1730-1737.

Abdel-Aty, M., Lee, J., Siddiqui, C., Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. Transportation Research Part A: Policy and Practice 49, 62-75.

Abdel-Aty, M., Siddiqui, C., Huang, H., 2011. Integrating trip and roadway characteristics in 1 managing safety at traffic analysis zones. Transportation Research Record: Journal of the Transportation Research Board 2213, 20-28.

Abdel-Aty, M.A., Lee, J., Eluru, N., Cai, Q., Amili, S.A., Alarifi, S., 2016. Enhancing and generalizing the two-level screening approach incorporating the highway safety manual (HSM) methods, Phase 2.

Abdel-Aty, M., Radwan, A., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention 32 (5), 633-642.

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

Agbelie, B.R., Roshandeh, A.M., 2015. Impacts of signal-related characteristics on crash frequency at urban signalized intersections. Journal of Transportation Safety & Security 7 (3), 199-207.

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

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

Aguero-Valverde, J., Jovanis, P.P., 2007. Identifying road segments with high risk of weather-related crashes using full Bayesian hierarchical models. Presented at the 86th Annual Meeting of the Transportation Research Board, Washington DC..

Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. Accident Analysis & Prevention 43(4), 1581-1589.

Alaluusua, S., Calderara, P., Gerthoux, P.M., Lukinmaa, P.-L., Kovero, O., Needham, L., Patterson Jr, D.G., Tuomisto, J., Mocarelli, P., 2004. Developmental dental aberrations after the dioxin accident in Seveso. Environmental health perspectives, 1313-1318.

Amoh-Gyimah, R., Saberi, M., Sarvi, M., 2017. The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. Analytic methods in accident research 13, 28-51.

American Association of State Highway and Transportation Officials (AASHTO), 2010. Highway Safety Manual, AASHTO, Washington, D.C.

Amoros, E., Martin, J.L., Laumon, B., 2003. Comparison of road crashes incidence and severity between some French counties. Accident Analysis & Prevention 35 (4), 537-547.

Anastasopoulos, P., Mannering, F., Shankar, V., Haddock, J., 2012. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 45 (1), 628-633.

Anselin, L., 2013. Spatial econometrics: Methods and models Springer Science & Business Media.

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.

Bhat, C.R., Born, K., Sidharthan, R., Bhat, P.C., 2014a. A count data model with endogenous covariates: Formulation and application to roadway crash frequency at intersections. Analytic Methods in Accident Research 1, 53-71.

Bhat, C.R., Paleti, R., Singh, P., 2014b. A spatial multivariate count model for firm location decisions. Journal of Regional Science 54(3), 462-502.

Brijs, T., Offermans, C., Hermans, E., Stiers, T., 2006. Impact of weather conditions on road safety investigated on hourly basis. In Proceedings of the Transportation Research Board 85th Annual Meeting.

Bullough, J.D., Donnell, E.T., Rea, M.S., 2013. To illuminate or not to illuminate: Roadway lighting as it affects traffic safety at intersections. Accident Analysis & Prevention 53, 65-77.

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.

Cai, Q., Wang, Z., Zheng, L., Wu, B., Wang, Y., 2014. Shock wave approach for estimating queue length at signalized intersections by fusing data from point and mobile

sensors. Transportation Research Record: Journal of the Transportation Research Board 2422, 79-87

Cai, Q., Lee, J., Eluru, N., Abdel-Aty, M., 2016. Macro-level pedestrian and bicycle crash analysis: incorporating spatial spillover effects in dual state count models. Accident Analysis & Prevention 93, 14-22.

Cai, Q., Abdel-Aty, M., Lee, J., Eluru, N., 2017a. Comparative analysis of zonal systems for macro-level crash modeling. Journal of Safety Research 61, 157-166.

Cai, Q., Abdel-Aty, M., Lee, J., 2017b. Bayesian joint approach of frequency and proportion modeling at macro level: Case study of crashes involving pedestrians or bicyclists. Compendium of papers CD-ROM, Transportation Research Board 94th Annual Meeting, Washington, D.C.

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accident Analysis & Prevention 39(4), 657-670.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on ice-accident frequencies and severities. Accident Analysis & Prevention 33 (1), 99-109.

Castro, M., Paleti, R., Bhat, C.R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. Transportation research part B: methodological 46(1), 253-272.

Census Bureau, U., 1992. Geographic areas reference manual. Washington, DC.

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.

Chin, H.C., Quddus, M.A., 2003. Modeling count data with excess zeroes an empirical application to traffic accidents. Sociological methods & research 32 (1), 90-116.

Chiou, Y.C., Fu, C., Chih-Wei, H., 2014. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. Analytic methods in accident research 2, 1-11.

Cho, G., Rodríguez, D.A., Khattak, A.J., 2009. The role of the built environment in explaining relationships between perceived and actual pedestrian and bicyclist safety. Accident Analysis & Prevention 41 (4), 692-702.

Dong, C., Clarke, D.B., Richards, S.H., Huang, B., 2014a. Differences in passenger car and large truck involved crash frequencies at urban signalized intersections: An exploratory analysis. Accident Analysis & Prevention 62, 87-94.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014b. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accident Analysis & Prevention 70, 320-9.

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

Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. Accident Analysis & Prevention 82, 192-198.

Dong, N., Huang, H., Lee, J., Gao, M., Abdel-Aty, M., 2016. Macroscopic hotspots identification: a Bayesian spatio-temporal interaction approach. Accident Analysis & Prevention 92, 256-264.

Dubin, R.A., 1988. Estimation of regression coefficients in the presence of spatially autocorrelated error terms. Review of Economics and Statistics 70, 466-474.

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

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis & Prevention 41(5), 1118-1123.

FHWA, Census Transportation Planning Products (CTPP), 2011. 2010 Census Traffic Analysis Zone Program MAF/TIGER Partnership Software Participant Guidelines.

FHWA, Census Transportation Planning Products (CTPP), 2014. Federal Highway Administration Website. http://www.fhwa.dot.gov/planning/census_issues/ctpp/status report/sr1214/index.cfm. Accessed 21.06, 2016.

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. Accident Analysis & Prevention 35(6), 991-1004.

FMCSA, U.S., 2012. Moving ahead for progress in the 21st century act ("MAP-21"). US Department of Transportation, Federal Motor Carrier Safety Administration.

Garnowski, M., Manner, H., 2011. On factors related to car accidents on German Autobahn connectors. Accident Analysis & Prevention 43(5), 1864-1871.

Gelman, A., Hill, J., 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Geyer, J., Raford, N., Ragland, D., Pham, T., 2006. The continuing debate about safety in numbers—data from Oakland, CA. Compendium of papers CD-ROM, Transportation Research Board 93rd Annual Meeting, Washington, D.C.

Girasek, D.C., Taylor, B., 2010. An exploratory study of the relationship between socioeconomic status and motor vehicle safety features. Traffic Injury Prevention 11, 151-155.

Graham, A.L., Hayward, A.D., Watt, K.A., Pilkington, J.G., Pemberton, J.M., Nussey, D.H., 2010. Fitness correlates of heritable variation in antibody responsiveness in a wild mammal. Science 330 (6004), 662-665.

Graham, D.J., Glaister, S., 2003. Spatial variation in road pedestrian casualties: The role of urban scale, density and land-use mix. Urban Studies 40 (8), 1591-1607.

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

Ha, H.-H., Thill, J.-C., 2011. Analysis of traffic hazard intensity: A spatial epidemiology case study of urban pedestrians. Computers, Environment and Urban Systems 35 (3), 230-240.

Hadayeghi, A., Shalaby, A., Persaud, B., 2003. Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record: Journal of the Transportation Research Board 1840, 87-95.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010. Development of planning level transportation safety tools using geographically weighted Poisson regression. Accident Analysis & Prevention 42 (2), 676-688.

Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. Journal of safety research 41 (4), 347-357.

Haleem, K., Gan, A., 2011. Identifying traditional and nontraditional predictors of crash injury severity on major urban roadways. Traffic injury prevention 12 (3), 223-234.

Haque, M.M., Chin, H.C., Huang, H., 2010. Applying Bayesian hierarchical models to examine motorcycle crashes at signalized intersections. Accident Analysis & Prevention 42 (1), 203-212.

Hauer, E., Council, F., Mohammedshah, Y., 2004. Safety models for urban four-lane undivided road segments. Transportation Research Record: Journal of the Transportation Research Board 1897, 96-105.

Hosseinpour, M., Prasetijo, J., Yahaya, A.S., Ghadiri, S.M.R., 2013. A comparative study of count models: Application to pedestrian-vehicle crashes along Malaysia federal roads. Traffic injury prevention 14 (6), 630-638.

Hosseinpour, M., Yahaya, A.S., Sadullah, A.F., 2014. Exploring the effects of roadway characteristics on the frequency and severity of head-on crashes: Case studies from Malaysian federal roads. Accident Analysis & Prevention 62, 209-222.

Houston, R. W., 1998. The transportation Equity Act for the 21st century. Institute of Transportation Engineers. ITE Journal 68(7), 45.

Hu, S.-R., Li, C.-S., Lee, C.-K., 2011. Assessing casualty risk of railroad-grade crossing crashes using zero-inflated Poisson models. Journal of Transportation Engineering 137 (8), 527-536.

Huang, H., Abdel-Aty, M., Darwiche, A., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transportation Research Record: Journal of the Transportation Research Board 2148, 27-37.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis & Prevention 42(6), 1556-1565.

Huang, H., Chin, H.C., 2010. Modeling road traffic crashes with zero-inflation and site-specific random effects. Statistical Methods & Applications 19 (3), 445-462.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography 54, 248-256.

Huang, H., Zhou, H., Wang, J., Chang, F. and Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. Analytic methods in accident research 14, 10-21.

Jang, H., Lee, S., Kim, S.W., 2010. Bayesian analysis for zero-inflated regression models with the power prior: Applications to road safety countermeasures. Accident Analysis & Prevention 42 (2), 540-7.

Jones, A.P., Jørgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. Accident Analysis & Prevention 35(1), 59-69.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transportation planning and Technology 15 (1), 41-58.

Jovanis, P.P., Chang, H.-L., 1986. Modeling the relationship of accidents to miles traveled. Transportation Research Record 1068, 42-51.

Karlaftis, M.G., Tarko, A.P., 1998. Heterogeneity considerations in accident modeling. Accident Analysis & Prevention 30 (4), 425-433.

Kim, D.-G., Washington, S., 2006. The significance of endogeneity problems in crash models: An examination of left-turn lanes in intersection crash models. Accident Analysis & Prevention 38 (6), 1094-1100.

Kim, D.-G., Washington, S., Oh, J., 2006a. Modeling crash types: New insights into the effects of covariates on crashes at rural intersections. Journal of Transportation Engineering 132 (4), 282-292.

Kim, K., Brunner, I., Yamashita, E., 2006b. Influence of land use, population, employment, and economic activity on accidents. Transportation Research Record: Journal of the Transportation Research Board 1953, 56-64.

Kim, K., Yamashita, E., 2002. Motor vehicle crashes and land use: Empirical analysis from Hawaii. Transportation Research Record: Journal of the Transportation Research Board 1784, 73-79.

Kim, D.G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models. Accident Analysis & Prevention 39 (1), 125-134.

Kononov, J., Bailey, B., Allery, B., 2008. Relationships between safety and both congestion and number of lanes on urban freeways. Transportation Research Record: Journal of the Transportation Research Board 2083, 26-39.

Kumala, R., 1995. Safety at rural three and four-arm junctions: Development of accident prediction models. Espoo: Technical Research Centre of Finland, VTT Publications.

Kumara, S.S., Chin, H.C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevention 4 (1), 53-57.

Kweon, Y.-J., 2011. Development of crash prediction models with individual vehicular data. Transportation Research Part C: Emerging Technologies 19 (6), 1353-1363.

Ladron De Guevara, F., Washington, S., Oh, J., 2004. Forecasting crashes at the planning level: Simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board (1897), 191-199.

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

Lascala, E.A., Gerber, D., Gruenewald, P.J., 2000. Demographic and environmental correlates of pedestrian injury collisions: A spatial analysis. Accident Analysis & Prevention 32 (5), 651-658.

Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accident Analysis & Prevention 37(4), 775-786.

Lee, J., Abdel-Aty, M., Choi, K., Huang, H., 2015. Multi-level hot zone identification for pedestrian safety. Accident Analysis & Prevention 76, 64-73.

Lee, J., Abdel-Aty, M., Choi, K., Siddiqui, C., 2013. Analysis of residence characteristics of drivers, pedestrians, and bicyclists involved in traffic crashes. In: Proceedings of the Transportation Research Board 92nd Annual Meeting.

Lee, J., Abdel-Aty, M., Jiang, X., 2014. Development of zone system for macro-level traffic safety analysis. Journal of Transport Geography 38, 13-21.

Lee, J., Abdel-Aty, M., Jiang, X., 2015b. Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accident Analysis & Prevention 78, 146-154.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis. Accident Analysis & Prevention 34 (2), 149-161.

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M.A., Cai, Q., 2016. A macroscopic analysis of crash proportion by mode: A fractional split multinomial logit modeling approach. In Proceedings of the Transportation Research Board 95th Annual Meeting, Washington D.C.

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Accident Analysis & Prevention 102, 213-226.

Lerner, E.B., Jehle, D.V., Bilittier, A.J., Moscati, R.M., Connery, C.M., Stiller,G., 2001. The influence of demographic factors on seatbelt use by adults injured in motor vehicle crashes. Accident Analysis & Prevention 33, 659-662.

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

Li, W., Carriquiry, A., Pawlovich, M., Welch, T., 2008. The choice of statistical models in road safety countermeasure effectiveness studies in Iowa. Accident Analysis & Prevention 40 (4), 1531-1542.

Li, Z., Wang, W., Liu, P., Bigham, J.M., Ragland, D.R., 2013. Using geographically weighted Poisson regression for county-level crash modeling in California. Safety science 58, 89-97.

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

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective. Safety Science 46 (5), 751-770.

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

Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accident Analysis & Prevention 39 (1), 53-57.

Loukaitou-Sideris, A., Liggett, R., Sung, H.-G., 2007. Death on the crosswalk a study of pedestrian-automobile collisions in Los Angeles. Journal of Planning Education and Research 26 (3), 338-351.

Ma, J., Kockelman, K., 2006. Bayesian multivariate Poisson regression for models of injury count, by severity. Transportation Research Record: Journal of the Transportation Research Board 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 & Prevention 40 (3), 964-975.

Macnab, Y.C., 2004. Bayesian spatial and ecological models for small-area accident and injury analysis. Accident Analysis & Prevention 36 (6), 1019-1028.

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.

Malyshkina, N.V., Mannering, F.L., 2009. Markov switching multinomial logit model: an application to accident-injury severities. Accident Analysis & Prevention 41(4), 829-838.

Maycock, G., Hall, R., 1984. Accidents at 4-arm roundabouts. TRRL Laboratory Report 1120, Transportation and Road Research Laboratory, Crow Thorne, UK.

Mayora, J.M.P., Rubio, R.L., 2003. Relevant variables for crash rate prediction in Spains two lane rural roads. In Proceedings of the Transportation Research Board 82nd Annual Meeting, Washington D.C.

Meyer, M.D., Miller, E.J., 1984. Urban transportation planning: A decision-oriented approach.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis & Prevention 26 (4), 471-482.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transportation Research Record: Journal of the Transportation Research Board (1840), 31-40.

Miaou, S.-P., Hu, P.S., Wright, T., Rathi, A.K., Davis, S.C., 1992. Relationship between truck accidents and highway geometric design: a Poisson regression approach. Transportation Research Record 1376.

Miaou, S.-P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis & Prevention 25 (6), 689-709.

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, 33-58.

Mitra, S., Chin, H.C., Quddus, M.A., 2002. Study of intersection accidents by maneuver type. Transportation Research Record: Journal of the Transportation Research Board 1784 (1), 43-50.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accident Analysis & Prevention 49, 439-448.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of econometrics 33 (3), 341-365.

Narayanamoorthy, S., Paleti, R., Bhat, C.R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transportation Research Part B: Methodological 55, 245-264.

National Highway Traffic Safety Administration, 2016. Traffic safety facts 2015. Department of Transportation.

Noland, R.B., 2003. Traffic fatalities and injuries: The effect of changes in infrastructure and other trends. Accident Analysis & Prevention 35 (4), 599-611.

Noland, R.B., Oh, L., 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: A case study of Illinois county-level data. Accident Analysis & Prevention 36 (4), 525-532.

Noland, R.B., Quddus, M.A., 2004. A spatially disaggregate analysis of road casualties in England. Accident Analysis & Prevention 36 (6), 973-984.

Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis & Prevention 41(4), 683-691.

Park, E.S., Park, J., Lomax, T.J., 2010. A fully Bayesian multivariate approach to before–after safety evaluation. Accident Analysis & Prevention 42(4), 1118-1127.

Park, J., Abdel-Aty, M., Wang, J.H., Lee, C., 2015. Assessment of safety effects for widening urban roadways in developing crash modification functions using nonlinearizing link functions. Accident Analysis & Prevention 79, 80-87.

Park, J., Abdel-Aty, M., Lee, J., Lee, C., 2015. Developing crash modification functions to assess safety effects of adding bike lanes for urban arterials with different roadway and socio-economic characteristics. Accident Analysis & Prevention 74, 179-191.

Peng, Y., Lord, D., 2011. Application of latent class growth model to longitudinal analysis of traffic crashes. Transportation Research Record: Journal of the Transportation Research Board 2236, 102-109.

Permpoonwiwat, C.K., Kotrajaras, P., 2012. Pooled time-series analysis on traffic fatalities in Thailand. World Review of Business Research 2 (6), 170-182.

Persaud, B.N., 1994. Accident prediction models for rural roads. Canadian Journal of Civil Engineering 21 (4), 547-554.

Pirdavani, A., Brijs, T., Bellemans, T., Wets, G., 2013. Spatial analysis of fatal and injury crashes in Flanders, Belgium: application of geographically weighted regression technique. Proceedings of the Transportation Research Board 92th Annual Meeting, Washington D.C.

Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequencies. Journal of Transportation Engineering 122 (2), 105-113.

Pulugurtha, S.S., Duddu, V.R., Kotagiri, Y., 2013. Traffic analysis zone level crash estimation models based on land use characteristics. Accident Analysis & Prevention 50, 678-687.

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

Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis & Prevention 36 (2), 183-191.

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

Rose, C.E., Martin, S.W., Wannemuehler, K.A., Plikaytis, B.D., 2006. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. Journal of Biopharmaceutical Statistics 16(4), 463-481.

Schneider Iv, W., Savolainen, P., Moore, D., 2010. Effects of horizontal curvature on single-vehicle motorcycle crashes along rural two-lane highways. Transportation Research Record: Journal of the Transportation Research Board 2194, 91-98.

Schneider, R., Diogenes, M., Arnold, L., Attaset, V., Griswold, J., Ragland, D., 2010. Association between roadway intersection characteristics and pedestrian crash risk in alameda county, California. Transportation Research Record: Journal of the Transportation Research Board 2198, 41-51.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accident Analysis & Prevention 29 (6), 829-837.

Shankar, V., Albin, R., Milton, J., Mannering, F., 1998. Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. Transportation Research Record: Journal of the Transportation Research Board 1635, 44-48.

Shankar, V.N., Chayanan, S., Sittikariya, S., Shyu, M.-B., Juvva, N.K., Milton, J.C., 2004. Marginal impacts of design, traffic, weather, and related interactions on roadside crashes. Transportation Research Record: Journal of the Transportation Research Board 1897, 156-163.

Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological), 683-690.

Siddiqui, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis & Prevention 45, 382-391.

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.

Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J., 2012. The role of built environment on pedestrian crash frequency. Safety Science 50 (4), 1141-1151.

US Congress., 2012. Moving ahead for progress in the 21st century act. Washington, DC.

U.S. Department of Transportation., 2015. The fixing America's surface transportation act. https://www.transportation.gov/fastact (accessed 04.24.2017)

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

Venkataraman, N., Ulfarsson, G., Shankar, V., Oh, J. and Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: Exploratory insights from random parameter negative binomial approach. Transportation research record: Journal of the Transportation Research Board 2236, 41-48.

Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accident Analysis & Prevention 59, 309-318.

Venkataraman, N., Shankar, V., Ulfarsson, G., Deptuch, D., 2014. Modeling the effects of interchange configuration on heterogeneous influences of interstate geometrics on crash frequencies. Analytic Methods in Accident Research 2, 12-20.

Vogt, A., 1999. Crash models for rural intersections: Four-lane by two-lane stop-controlled and two-lane by two-lane signalized. Publication FHWA-RD-99-128. FHWA, U.S. DOT.

Wang, J., Huang, H., 2016. Road network safety evaluation using Bayesian hierarchical joint model. Accident Analysis & Prevention 90, 152-158.

Wang, J., Huang, H., Zeng, Q., 2017. The effect of zonal factors in estimating crash risks by transportation modes: Motor vehicle, bicycle and pedestrian. Accident Analysis & Prevention 98, 223-231.

Wang, J.H., Abdel-Aty, M.A., Park, J., Lee, C., Kuo, P.F., 2015. Estimating safety performance trends over time for treatments at intersections in Florida. Accident Analysis & Prevention 80, 37-47.

Wang, L., Abdel-Aty, M., Lee, J., 2017. Safety analytics for integrating crash frequency and real-time risk modeling for expressways. Accident Analysis & Prevention 104, 58-64.

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

Wang, X., Abdel-Aty, M., 2008. Modeling left-turn crash occurrence at signalized intersections by conflicting patterns. Accident Analysis & Prevention 40 (1), 76-88.

Wang, X., Abdel-Aty, M., Brady, P., 2006. Crash estimation at signalized intersections: Significant factors and temporal effect. Transportation Research Record: Journal of the Transportation Research Board 1953, 10-20.

Wang, X., Fan, T., Chen, M., Deng, B., Wu, B., Tremont, P., 2015. Safety modeling of urban arterials in Shanghai, China. Accident Analysis & Prevention 83, 57-66.

Wang, X., Yuan, J., Scheltz, G., Meng, W., 2016. Investigating safety impacts of roadway network features of suburban arterials in Shanghai, China. Presented at the 95th Annual Meeting of the Transportation Research Board, Washington DC.

Wang, Z., Cai, Q., Wu, B., Zheng, L., Wang, Y., 2016. Shockwave-based queue estimation approach for undersaturated and oversaturated signalized intersections using multi-source detection data. Journal of Intelligent Transportation Systems,1-12.

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

Washington, S., Van Schalkwyk, I., Mitra, S., Mayer, M., Dumbaugh, E., Zoll, M., 2006. NCHRP Report 546: Incorporating Safety into Long-Range Transportation Planning. Transportation Research Board, Washington, DC.

Wedagama, D.P., Bird, R.N., Metcalfe, A.V., 2006. The influence of urban land-use on non-motorised transport casualties. Accident Analysis & Prevention 38 (6), 1049-1057.

Wier, M., Weintraub, J., Humphreys, E.H., Seto, E., Bhatia, R., 2009. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. Accident Analysis & Prevention 41 (1), 137-145.

Wu, Y., Ding, Y., Abdel-Aty, M., Jia, B., 2015. A comparative analysis of different dilemma zone countermeasures on signalized intersections' safety based on cellular automaton model. Presented at the 96th Annual Meeting of the Transportation Research Board, Washington DC.

Wu, Y., Abdel-Aty, M., Lee, J., 2017. Crash risk analysis during fog conditions using real-time traffic data. Accident Analysis & Prevention (in press).

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 & Prevention 54, 90-98.

Xie, Y., Zhao, K., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accident Analysis & Prevention 47, 36-44.

Xie, K., Wang, X., Huang, H., Chen, X., 2013. Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accident Analysis & Prevention 50, .25-33.

Xie, K., Ozbay, K., Kurkcu, A., Yang, H., 2017. Analysis of traffic crashes involving pedestrians using big data: Investigation of contributing factors and identification of hotspots. Risk Analysis.

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. Accident Analysis & Prevention 75, 16-25.

Xu, P., Huang, H., Dong, N., Abdel-Aty, M., 2014. Sensitivity analysis in the context of regional safety modeling: Identifying and assessing the modifiable areal unit problem. Accident Analysis & Prevention 70, 110-120.

Xu, P., Huang, H., Dong, N., Wong, S.C., 2017. Revisiting crash spatial heterogeneity: A Bayesian spatially varying coefficients approach. Accident Analysis & Prevention 98, 330-337.

Yasmin, S., Eluru, N., Bhat, C.R., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Analytic Methods in Accident Research 1, 23-38.

Yasmin, S. and Eluru, N., 2016. Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. Accident Analysis & Prevention 95,157-171.

Yasmin, S., Eluru, N., Lee, J. and Abdel-Aty, M., 2016. Ordered fractional split approach for aggregate injury severity modeling. Transportation Research Record: Journal of the Transportation Research Board 2583, 19-126.

Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Science 47 (3), 443-452.

Yu, R., Abdel-Aty, M., 2013. Multi-level Bayesian analyses for single-and multi-vehicle freeway crashes. Accident Analysis & Prevention 58, 97-105.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accident Analysis & Prevention 50, 371-376.

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

Zhang, C., Ivan, J., 2005. Effects of geometric characteristics on head-on crash incidence on two-lane roads in Connecticut. Transportation Research Record: Journal of the Transportation Research Board 1908, 159-164.

Zhang, Y., Bigham, J., Li, Z., Ragland, D., Chen, X., 2012, January. Associations between road network connectivity and pedestrian-bicyclist accidents. In 91st Annual Meeting of the Transportation Research Board, Washington DC, Washington, DC.

Zhang, Y., Bigham, J., Ragland, D., Chen, X., 2015. Investigating the associations between road network structure and non-motorist accidents. Journal of transport geography 42,34-47.

Zou, Y., Zhang, Y., Lord, D., 2013. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. Accident Analysis & Prevention 50, 1042-1051.

Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Analytic methods in accident research 1, 39-52.