2014

Development of Traffic Safety Zones and Integrating Macroscopic and Microscopic Safety Data Analytics for Novel Hot Zone Identification

JaeYoung Lee
University of Central Florida

DEVELOPMENT OF TRAFFIC SAFETY ZONES AND INTEGRATING MACROSCOPIC AND MICROSCOPIC SAFETY DATA ANALYTICS FOR NOVEL HOT ZONE IDENTIFICATION

by

JAEYOUNG LEE

B. Eng. Ajou University, Korea, 2007
M.S. Ajou University, Korea, 2009

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 the University of Central Florida
Orlando, Florida

Spring Term
2014

Major Professor: Mohamed Abdel-Aty

ABSTRACT

Traffic safety has been considered one of the most important issues in the transportation field. With consistent efforts of transportation engineers, Federal, State and local government officials, both fatalities and fatality rates from road traffic crashes in the United States have steadily declined from 2006 to 2011.Nevertheless, fatalities from traffic crashes slightly increased in 2012 (NHTSA, 2013). We lost 33,561 lives from road traffic crashes in the year 2012, and the road traffic crashes are still one of the leading causes of deaths, according to the Centers for Disease Control and Prevention (CDC). In recent years, efforts to incorporate traffic safety into transportation planning has been made, which is termed as transportation safety planning (TSP). The Safe, Affordable, Flexible Efficient, Transportation Equity Act – A Legacy for Users (SAFETEA-LU), which is compliant with the United States Code, compels the United States Department of Transportation to consider traffic safety in the long-term transportation planning process.

Although considerable macro-level studies have been conducted to facilitate the implementation of TSP, still there are critical limitations in macroscopic safety studies are required to be investigated and remedied. First, TAZ (Traffic Analysis Zone), which is most widely used in travel demand forecasting, has crucial shortcomings for macro-level safety modeling. Moreover, macro-level safety models have accuracy problem. The low prediction power of the model may be caused by crashes that occur near the boundaries of zones, high-level aggregation, and neglecting spatial autocorrelation.

In this dissertation, several methodologies are proposed to alleviate these limitations in the macro-level safety research. TSAZ (Traffic Safety Analysis Zone) is developed as a new zonal system for the macroscopic safety analysis and nested structured modeling method is suggested to improve the model performance. Also, a multivariate statistical modeling method for multiple crash types is proposed in this dissertation. Besides, a novel screening methodology for integrating two levels is suggested. The integrated screening method is suggested to overcome shortcomings of zonal-level screening, since the zonal-level screening cannot take specific sites with high risks into consideration. It is expected that the integrated screening approach can provide a comprehensive perspective by balancing two aspects: macroscopic and microscopic approaches.

ACKNOWLEDGMENT

The author would like to thank his advisor, Dr. Mohamed Abdel-Aty, for his invaulable guidance, advice and support and encouragement toward successful completion of his doctoral course. The author wishes to acknowldege the support of my committee members, Dr. Essam Radwan, Dr. Keechoo Choi, Dr. Boo Hyun Nam, and Dr. Pei-Fen Kuo.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

AADT	Average Annual Daily Traffic
AIC	Akaike Information Criterion
BG	Block Group
BIC	Bayesian Information Criterion
BPLSEM	Bayesian Poisson Lognormal Spatial Error Model
CAR	Crash Analysis Reporting
CAR	Conditional Autoregressive
CB	Census Block
CC	Cold (macro) and Cold (micro) zone
CDC	Centers for Disease Control and Prevention
CH	Cold (macro) and Hot (micro) zone
CN	Cold (macro) and Normal (micro) zone
CO	Cold (macro) and no data (micro) zone
CRP	Continuous Risk Profile
CT	Census Tract
CTTP	Census Transportation Planning Product
DIC	Deviance Information Criterion
DOT	Department of Transportation
DUI	Driving Under the Influence
EB	Empirical Bayes
EPDO	Equivalent Property Damage Only

FARS	Fatality Analysis Reporting System
FDOT	Florida Department of Transportation
FHWA	Federal Highway Administration
F.S.	Florida Statutes
FSB	Freeway-and-expressway State road Boundary crash
FSI	Freeway-and-expressway State road Interior crash
GIS	Geographic Information System
GWPR	Geographically Weighted Poisson Regression
HC	Hot (macro) and Cold (micro) zone
HH	Hot (macro) and Hot (micro) zone
HN	Hot (macro) and Normal (micro) zone
HO	Hot (macro) and no data (micro) zone
HSM	Highway Safety Manual
LOSS	Level of Safety Service
LR	Log-Likelihood
LRTP	Long Range Transportation Plan
MAD	Mean Absolute Deviation
MAUP	Modifiable Areal Unit Problem
MCAR	Multivariate Conditional Autoregressive
MPO	Metropolitan Planning Organization
MTR	Mass Transit Railway
MV	Multivariate model without spatial error term

MVS	Multivariate model with Spatial error term
NB	Negative Binomial
NBPLSEM	Nested Bayesian Poisson-Lognormal Spatial Error Model
NC	Normal (macro) and Cold (micro) zone
NH	Normal (macro) and Hot (micro) zone
NHTS	National Highway Travel Survey
NHTSA	National Highway Traffic Safety Administration
NN	Normal (macro) and Normal (micro) zone
NO	Normal (macro) and no data (micro) zone
NSB	Non-State Boundary crash
NSI	Non-State Interior crash
OSB	Other State Boundary crash
OSI	Other State Interior crash
PCU	Passenger Car Unit
PDO	Property Damage Only
PMAD	Percent Mean Absolute Deviation
PSI	Potential for Safety Improvements
PSL	Posted Speed Limit
RMSE	Root Mean Squared Errors
SAD	Sum of Absolute Deviation
SAFETEA-LU	Safe, Affordable, Flexible Efficient, Transportation Equity Act-A Legacy for Users
SCP	Safety-Conscious Planning

SPF	Safety Performance Function
TAD	Traffic Analysis District
TAZ	Traffic Analysis Zone
TSAZ	Traffic Safety Analysis Zone
TSP	Transportation Safety Planning
UV	Univariate model
UVS	Univariate model with Spatial error term
VIN	Vehicle Identification Number
VKT	Vehicle-Kilometers-Traveled
VMT	Vehicle-Miles-Traveled
ZCTA	ZIP Code Tabulation Area

CHAPTER 1 INTRODUCTION

1.1 Research Motivation

Traffic safety has been considered one of the most important issues in the transportation field. With consistent efforts of transportation engineers, Federal, State, and local government officials, both fatalities and fatality rates from road traffic crashes in the United States have been steadily declining over the last five years. Nevertheless, we still lost 32,310 lives of people from road traffic crashes in the year of 2011 (NHTSA, 2012), and the road traffic crashes are still one of the leading cause of deaths, according to the CDC statistics (Hoyert and Xu, 2011).

In recent years, efforts to incorporate traffic safety into transportation planning has been made, which is termed as transportation safety planning (TSP). The Safe, Affordable, Flexible Efficient, Transportation Equity Act – A Legacy for Users (SAFETEA-LU) (FHWA, 2005), which is compliant with the United States Code, compels the United States Department of Transportation to consider traffic safety in the long-term transportation planning process. Many macroscopic safety models have been developed to facilitate the implementation of TSP. Most of these models were estimated based on current zone systems such as traffic analysis zones (TAZs), census block (CB) based block groups (BGs) or census tracts (CTs). Nevertheless, no researchers have focused on developing a new zone system exclusively for the macroscopic traffic safety analysis.

Similarly, few researchers have attempted zonal-level screenings although microscopic network screening methodologies have been developed by many researchers. However, the microscopic network screening only concentrates on specific intersections, segments, or corridor. Thus, it is not adequate for the area-wide screening. Therefore, it is necessary to develop a new zonal-level screening method to provide a comprehensive perspective for policy-makers.

1.2 Problem Statement

Generally transportation planners assemble data for collection and processing by TAZs for the travel demand modeling. Although TAZs are designed for the transportation planning purpose, it has also been used in the macro-level crash modeling. Of course other geographic units such as BGs, CTs, county and state are used depending on the scope of studies. TAZs have been extensively used in traffic crash studies since they are the only transportation related areal units. Moreover, TAZs are delineated by State Department of Transportation (State DOT) and/or Metropolitan Planning Organizations (MPOs) for developing their long range transportation plans. Thus, TAZs seem to be preferred areal units in the practical point of view. Nevertheless, few studies questioned the validity of TAZs for the macroscopic safety analysis so far. Therefore, the main objective of this study is to explore possible limitations of TAZs as basic spatial units of traffic safety modeling, and developing Traffic Safety Analysis Zones (TSAZs), if TAZs are found not appropriate for the safety studies.

Furthermore, macroscopic crash models have limited accuracy in predicting traffic crashes. The limited accuracy is caused by several factors. For instance, it is believed that crashes occurring near or on the boundaries of zones are not influenced by a single zone, whereas crashes occurring completely inside a zone are affected by only one zone. However, only few researchers addressed the boundary issue in the macroscopic safety analysis. In order to solve boundary problem, two methods are suggested in this study. First, a nested structure that separates boundary and interior crashes by different roadway types was proposed. It enables estimating boundary and interior crash models individually. Second, a variable transformation was suggested to relate boundary crashes with adjacent multiple zones. It allows developing boundary crash models with zonal factors from neighboring zones. The other issue is a spatial autocorrelation problem. Spatial autocorrelation is a technical term for the fact that spatial data from near locations are more likely to be similar than data from distant location (O'Sullivan and Unwin, 2002). The existence of spatial autocorrelation in the crash data may invalidate the assumption of the random distribution (LeSage and Pace, 2004). Using Moran's *I* statistics, it was found that the spatial autocorrelation exists in the collected data set in this study. Thus, spatial autocorrelation effects were addressed in the modeling process.

Lastly, few researchers have attempted zonal-level screenings although microscopic network screening methodologies have been developed by many researchers. In this study, a novel screening methodology for integrating both macroscopic and microscopic safety data analytics was suggested. Because the zonal-level screening cannot take specific sites with high risks into

consideration, a novel approach to integrate macroscopic and microscopic screening results was proposed to overcome shortcomings of zonal-level screening.

1.3 Research Objectives

The dissertation focuses on creating the new zone system exclusively for the transportation safety planning, analyzing the traffic safety at the macroscopic level based on the new system, and developing a novel methodology for the integrated screening using both macroscopic and microscopic safety data analytics. The specific objective will be achieved by the following procedures:

1. Conducting preliminary safety studies at the macroscopic level;
2. Developing TSAZs using the regionalization technique;
3. Estimating high-accuracy SPFs based on the new zone system, and;
4. Developing an integrated screening methodology.

The first objective is analyzing traffic safety at the macroscopic level and it was achieved by following tasks:

a) Analyzing relations with residence characteristics with the number of at-fault drivers. It helps to select candidate variables for the final model (Chapter 3), and;

b) Adopting multivariate modeling. The multivariate modeling for motor vehicle, bicycle, and pedestrian crashes are estimated to improve macroscopic safety models in Chapter 4.

The second objective is the main goal of this study and it was achieved by the following tasks:

c) Analyzing effects of different zone systems. Models based on TAZs, BGs, and CTs are developed and compared in Chapter 5.

d) Exploring zonal effects from different regionalization processes. Regionalization can be conducted in keeping homogenous crash patterns in each zone. Also optimal zone scale for TSAZs was determined using Brown-Forsythe tests. The regionalization task is presented in Chapter 6, and;

e) Comparing models based on TAZs and TSAZs (Chapter 6).

The following tasks were implemented to achieve the third objective:

f) Comparing different spatial autocorrelation conceptualization methods. Four conceptualization methods are compared and evaluated in Chapter 7;

g) Applying a nested structure and variable transformation to account for boundary crashes (Chapter 8), and;

h) Developing macroscopic SPFs based on TSAZs (Chapter 9).

The final objective, which is also one of the main goals of this study, was achieved by following tasks:

i) Calculating PSIs and identifying hot zones/spots at both macroscopic and microscopic levels (Chapter 10), and;

j) Developing a novel methodology combining both macroscopic and microscopic screening results (Chapter 11).

1.4 Organization of the Dissertation

The dissertation is organized as follows: Chapter 2, following this chapter, summarizes literature review on previous macroscopic traffic safety studies. Current issues of macroscopic safety

researches and related studies and their limitations are discussed. Additionally, it will be also explained how to address limitations in these studies. Chapter 3 analyzes residence characteristics of at-fault drivers based on ZIP codes. Several socio-demographic factors from residence ZIP areas are attempted to find out relations with the number of at-fault drivers. Chapter 4 attempted to adopt multivariate modeling to improve the model predictability at the macroscopic level. Crash types by travel modes (i.e., motor vehicle, bicycle, and pedestrian crashes) are modeled simultaneously accounting for unobserved common factors among the different crash types. Chapter 5 provides the comparison of models based on different three zone systems (i.e., TAZs, BGs, and CTs). Chapter 6 suggests the development of TSAZs. TSAZs were developed using regionalization technique by aggregating current TAZs with homogenous traffic crash patterns. The Brown-Forsythe test is applied to determine the optimal zone scale. Chapter 7 compares several conceptualization methods of spatial autocorrelations. Chapter 8 provides two effective methodologies to handle boundary crashes in the modeling. First method is using the nested structure that separates boundary and interior crashes by roadway types. Second method is the variable transformation that enables to relate boundary crashes with multiple adjacent zones. Chapter 9 suggests modeling results based on the nested structure suggested in Chapter 8. Chapter 10 provides the macroscopic screening results using PSI calculated models from Chapter 9. Also, microscopic screening results for intersections and segments are briefly provided. Chapter 11 suggests a novel screening methodology combining both macroscopic and microscopic screening results. Finally, Chapter 12 summarizes the overall dissertation and proposes a set of recommendations and follow-up studies.

CHAPTER 2 LITERATURE REVIEW

2.1 Macroscopic Safety Studies

Macroscopic safety studies aim to find out statistical association of zonal characteristics with the aggregated number of crashes in zones (Levine et al., 1995). Thus, macroscopic safety researchers do not focus on individual crashes occurring at specific locations, instead they aggregate traffic crashes into spatial units. Nicholson (1985) asserted that the aggregation of traffic crashes eliminates some of fluctuations over years in traffic crashes which occur at individual locations. Moreover, the sum is more stable than those produced for small zones (so called “Law of Large Numbers”).

Nevertheless, Levine et al. (1995) claimed that two biases are produced by aggregation. First, crashes assigned to zone, rather than specific locations, producing spatial error. Second, zones assume that the risk of crashes is uniform at all locations with the zone, a situation which is frequently not correct. However, by choosing zones which are small and relatively homogenous, the advantages of grouping, namely the ability to associate characteristics of the zone with crashes outweigh the biases produced by aggregation. In general, spatial units are based on census zones such as BGs, CTs or TAZs. Counties or states can be used as basic areal units if more macroscopic level analyses are required. Sometimes, the postal code, as known as ZIP code, was used since it has own advantages. Spatial units which are widely used in the macroscopic studies and targeted crash types in each study will be briefly explained in following sub-chapters.

2.1.1 Geographical units for studies

Census-based units include CBs, BGs, and CTs. A CB is the smallest geographic unit used by United States Census Bureau for the collection and tabulation of decennial census data (U.S. Census Bureau, 1994). Detailed information is not available based on CBs due to confidentiality requirement. Moreover, CBs are very small, especially in the urban area. Averagely there are only 85 people in one CB and even one building can be divided into several blocks. Due to the lack of information and the small size, typically CBs are not used for the aggregate-level safety studies.

A BG is the next level above CBs and a BG is a combination of CBs. Each BG contains 39 CBs in average. Population in a BG ranges between 600 and 3,000 people. Several macroscopic studies were conducted based on BGs (Levine et al., 1995; Abdel-Aty et al., 2013). A CT is a combination of BGs. A CT is a statistical subdivision of a county that may include from 2,500 to 8,000 people. A CT is designed to keep homogenous socioeconomic status. A few safety studies have been done using CTs (LaScala et al., 2000; Loukaitou-Sideris et al., 2007; Wier et al. 2009; Ukkusuri et al., 2011; Abdel-Aty et al., 2013; Wang and Kockelman, 2013). Figure 2-1 compares CBs, BGs and CTs in the downtown of Orlando, Florida. As shown in the Figure 2-1, CBs have the smallest area whereas BGs and CTs have much larger size of zones.

Census Blocks (CBs)

Block Groups (BGs)

Census Tracts (CTs)

Figure 2-1: Comparison of CBs, BGs and CTs in the downtown of Orlando

Abdel-Aty et al. (2013) conducted interesting research regarding the effect from different zone systems. Authors compared crash models based on three different areal units BGs, CTs and TAZs. Authors discovered that the BG based model had the larger number of significant variables for total and severe crashes compared to models based on other geographical units. The detailed explanation of this study will be suggested in the MAUP part.

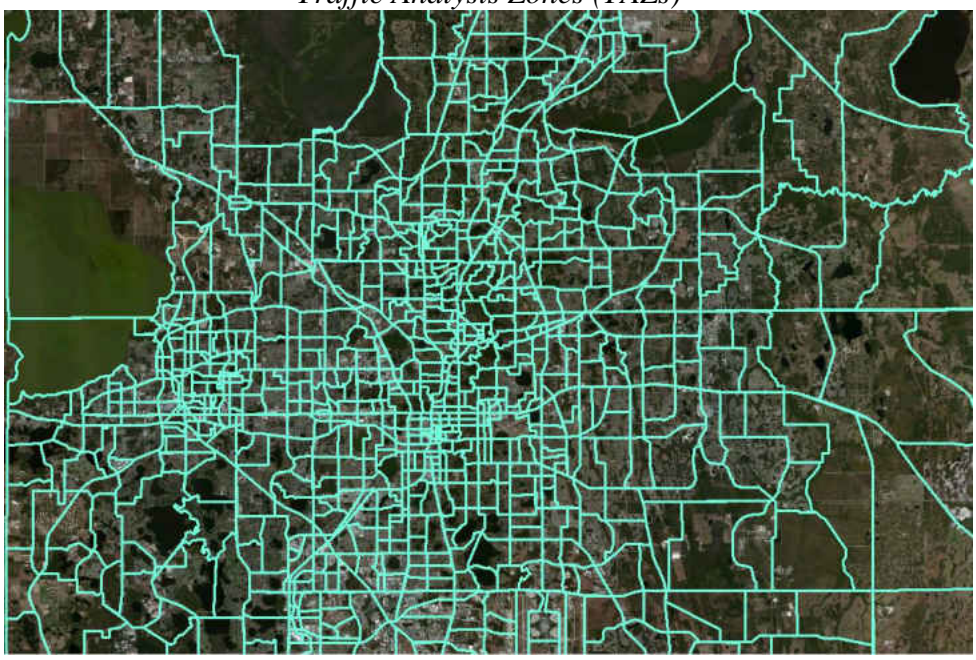
ZIP code is a system of postal codes made and used by United States Postal Service since 1963. As a matter of fact, ZIP codes are not geographical, features but a collection of mail delivery routes. U.S. Census created ZIP Code Tabulation Areas (ZCTAs), which are generalized areal representations of ZIP code service areas. ZCTA based data are also provided from U.S. Census Bureau. Many studies using ZIP code have been conducted (Blatt and Furman, 1998; Stamatiadis and Puccini, 2000; Lerner et al., 2001; Clark, 2003; Romano et al., 2006; Males, 2009; Girasek and Taylor; 2010; Lee et al., 2013; Lee et al., 2014).

ZIP codes are mostly used for the residence information, instead of the crash location. This is because the residence information is only provided as ZIP code in most cases. Many researchers only focused on road users only involved in fatal crashes since FARS (Fatality Analysis Reporting System) offers ZIP codes of drivers involved in fatal crashes. Thus, ZIP code based studies using FATS ZIP code data only focused on fatal crashes (Blatt and Furman, 1998; Stamatiadis and Puccini, 2000; Clark, 2003; Romano et al., 2006; Males, 2009; Girasek and Taylor; 2010). Nevertheless, some other researchers collected from different sources and thus

they could analyze injury crashes (Lerner et al., 2001) and total crashes (Lee et al., 2013; Lee et al., 2014).

TAZs are special purpose geographic entities delineated by state and local transportation officials for tabulating traffic-related data, especially journey-to-work and place-of-work statistics (U.S. Census Bureau, 2011). Since TAZs are the only traffic related zone system, TAZs have been most popularly used in the macroscopic safety literature (Ng et al., 2002; Hadayeghi et al., 2003, 2010a, 2010b; Guevara et al., 2004; Hadayeghi et al., 2006; Naderan and Shashi, 2010; Abdel-Aty et al., 2011; Siddiqui and Abdel-Aty, 2012; Siddiqui et al., 2012; Wang et al., 2012; Pirdavani et al., 2012; Huang et al., 2013; Pirdavani et al., 2013a, 2013b; Pulugurtha et al., 2013; Abdel-Aty et al., 2013). However, there are possible limitations of TAZs for the macroscopic safety analysis due to their zoning criteria. This issue will be addressed in the following MAUP sub-chapter. TADs are new, higher-level geographic entity for traffic analysis (U.S. Census Bureau, 2011). TADs are created by aggregating existing TAZs. TADs may cross county boundaries, but they must nest within MPOs.

Traffic Analysis Zones (TAZs)



Traffic Analysis Districts (TADs)

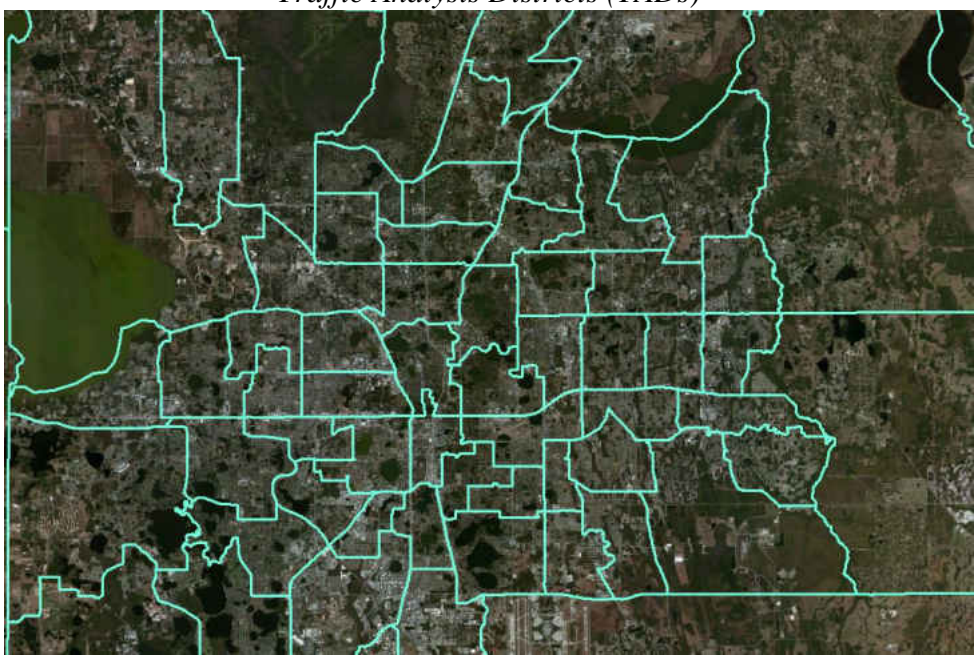


Figure 2-2: Comparison of TAZs and TADs in Orlando

For higher level of the macroscopic analysis, counties or states are also used as spatial units. Amoros et al. (2003), Noland and Oh (2004), Aguero-Valverde and Jovanis (2006) and Huang et al. (2010) aggregated data into county-levels and analyzed crashes. Noland (2003) conducted state-level study. Some researchers used unique spatial units that are not usually used in the United States or macroscopic studies. Noland and Quddus (2004) used standard statistical regions and census wards, respectively. In addition, Kim et al. (2006) employed grid-based units for their study. One of advantages using the fixed grid structure is all units were even sized different from other areal units. Lastly, MacNab (2004) used Local Health Areas to investigate injury crashes.

2.1.2 Studies on total crashes

Generally, macroscopic crash studies have analyzed crashes by severity level as total, property damage only (PDO) crashes, injury, severe, or fatal crashes. Crashes related to particular transportation mode users as bicyclists or pedestrians are also widely explored. However, Specific types of crashes such as head-on crashes or intersection crashes are not usually targeted in the macroscopic safety analysis. In this sub-chapter, contributing factors, for the specific types or severity levels of the crashes, from previous studies will be discussed. Levine et al. (1995), Ng et al. (2002), Hadayeghi et al. (2003, 2006), Noland and Quddus (2004), Kim et al. (2006), Huang et al. (2010), Naderan and Shashi (2010), Abdel-Aty et al. (2011, 2013), Pulugurtha et al. (2013), and Lee et al. (2013) focused on the total crash.

Levine et al. (1995) investigated the spatial relationship between trip generating activities and total traffic crashes using BGs of City and County Honolulu for 1990. Authors revealed that the number of total traffic crashes is also affected by the characteristics of neighborhoods and area, not only on the roadway network. Authors found out population, manufacturing employment, retail trade employment and service employment are positively associated with total crashes occurred. In contrast, financial employment and military employment are negatively related to the number of crashes. Moreover, three roadway variables as the freeway segment, major arterial length, and miles of freeway ramp/freeway access were also positively related to total crash frequencies.

Ng et al. (2002) used land-use explanatory variables for the modeling. Authors found that cinema seats, commercial floor areas, market stalls, mass transit railway catchment area, other specified uses were positively associated with total crashes and they were significant at 90% confidence level. In contrast, the container storage area and the country park were negatively related to total crashes. Hadayeghi et al. (2003) asserted that the vehicle-kilometers-traveled (VKT), the major roadway length, the total number of employment and the minor roadway length had positive relationships with the total crashes during the morning period. It was also found that the estimate of major road length is approximately 10 times larger than that of minor road length. In contrast, the volume to capacity and the posted speed limit were negatively associated with the total crash in the morning.

Noland and Quddus (2004) investigated the effect of infrastructure and demographic change on traffic-related fatalities and crashes using Illinois county level data from 1987 to 1990. Authors estimated three models: 1) total crash model without time correction, 2) total crash model with year dummy variables, and 3) total crash model with year dummy variables and demographic variables. Based on the likelihood of the models, the model with year dummy and demographic variables had the best fit. They found that the infrastructure variables were not statistically significant, except for the mean number of lanes and mean outside shoulder width. Among the demographic variables, only the population was found significantly associated with increased crashes at the 90% level of significance. Authors also tried to include per capita income into the model but it was not significant. Many studies already found a correlation between crashes (or dangerous attitudes) and economic status, or income (Lerner et al., 2001; Noland, 2003; Romano et al., 2006; Males, 2009; Girasek and Taylor, 2010; Huang et al., 2010; Abdel-Aty et al., 2013; Lee et al., 2013; Lee et al., 2014) but it was shown that the income had no effect on crashes at the county-level.

Hadayeghi et al. (2006) looked into the temporal transferability of the morning period crash prediction models. Significant variables found in 1996 and 2001 crash models were found exactly same. The socioeconomic variable as total employments, and traffic related variable as volume-to-capacity, and roadway related variables including major/minor roadway length and the posted speed were significant for the total crashes. While the vehicle-miles-traveled (VMT), the major roadway length, the number of households, the speed, and the volume-to-capacity were significant for the severe crash. The results presented that both total and severe models were not

temporally transferable. Kim et al. (2006) created the 0.1 mi^2 grid structure, and estimated various regression models based on the grid. Authors identified the specific nature of the relationships between zonal characteristics and for four types of traffic crashes as total crashes, vehicle-to-vehicle, pedestrian, and bicycle related crashes.

Huang et al. (2010) investigated both total and severe crashes at the county level. Authors employed Bayesian spatial model to account for county-level variations of crash risk in Florida. Authors found out there is no difference in safety effects of risk factors in total and severe crashes. Counties with the larger traffic and population concentration, higher level of urbanization were related to the higher crash risk. Authors also revealed that the young drivers tend to involved in more crashes, whereas the elderly population decreases the crash risk. Moreover, it was discovered that counties with lower income, educational level and employment rate also contributed to the higher crash risk. Naderan and Shashi (2010) introduced the concept of crash generation model in their study. Authors utilized trip generation data in a generalized linear regression with the assumption of a negative binomial (NB) error structure. Total crash production model, which only used trip generation factors as explanatory variables, showed that work trip, college trip productions, shopping trip and non-home-based trip productions were positively associated with total crashes, whereas school trip, recreational trip and pilgrimage trip productions were negatively related. Moreover, total crash attraction model, which used trip attraction variables, revealed that educational trip, shopping trip, recreational trip and personal trip attractions had positive relationships with total crashes, while work trip attraction density and non-home-based attractions had negative associations with total crashes. Last model used

both total productions and it showed estimates of both variables had positive signs but the production factor had a slightly larger magnitude than the attraction factor. Among three models, the crash production model was a better fit than other two models.

Abdel-Aty et al. (2011) also employed trip generation factors in safety models. Authors estimated various models using 1) trip generation related factors, 2) total trip productions and attractions, 3) roadway related factors, and 4) all factors. Authors discovered that the model estimated using only total trip productions and attractions performed best. This result is inconsistent with the result of Naderan and Shashi (2010). Moreover, Abdel-Aty et al. (2011) found that the estimate of the trip attraction had an approximately twice larger magnitude compared to that of the trip production, which was quite different from the work of Naderan and Shashi (2010). Abdel-Aty et al. (2013) used Bayesian Poisson-lognormal models for total crashes by different areal units such as CTs, BGs and TAZs. Authors compared significant variables of three models based on different geographical units. It was shown that three models have several common significant variables such as the VMT, the roadway length with a speed limit of 35, 55 and 65mph, the number of intersections, the proportion of the minority population, commuters by walking, and the median household income. On the other hand, the population of the youngest age group (0 to 15 years), and roadway length with the high speed limit (65mph) were significant solely in the CT based model. Moreover, it is interesting that both workers commuting by public transportation and the density of housing units were negatively associated in the CT/BG based models; however, they were not significant in the TAZ based model.

Pulugurtha et al. (2013) used land use variables to estimate crash models based on TAZs with data in Charlotte, North Carolina. Authors found the mixed use development area, the urban residential area, the business area, the multi-family residential area, the office district area, the institutional area, the roadway area, the planned unit development area and, research district area increase the total number of crashes, whereas the single-family residential area decrease the total crashes. Lee et al. (2013) did not analyze zones of crash locations, but authors explored residence zones of at-fault drivers using NB models. At-fault drivers are defined as drivers who were responsible for the traffic crash. The result displayed that drivers living in zones with more Hispanic people, more households in the urban area had higher probability to cause traffic crashes, regardless of severity level or types. While drivers living in zones with higher median age, greater median household income, and more workers in the primary industry field are less likely to cause any kind of traffic crashes.

2.1.3 Studies on PDO and/or injury crashes

A few studies looked into PDO crashes separately, at the macroscopic level (Guevara, 2004; Nederan and Shashi, 2010; Pulugurhta et al., 2013). Guevara et al. (2004) estimated a NB model for PDO crashes. Authors found that the population density, the number of employees, the intersection density, the percentage of other principal arterial, minor arterial and urban collectors were significant for PDO crashes, and all of these variables had positive signs of estimates.

In the study of Naderan and Shashi (2010), PDO crash production model showed that work trip, college trip productions, shopping trip and non-home-based trip productions were positively

associated with PDO crashes, whereas school trip, recreational trip and pilgrimage trip productions were negatively related. In addition, PDO crash attraction model presented that shopping trip attractions, educational trip attraction density, recreational trip attraction density, and personal trip attraction density were positively related to PDO crashes, while working attraction trip density and non-home-based attraction trip density were negatively associated with PDO crashes. Furthermore, both total productions and attractions were significant in the PDO crash generation model, and they had positive signs. Based on goodness-of-measure, it was found that the PDO crash production model had the better fit compared to other two models. Pulugurtha et al. (2013) asserted that the mixed use development area, the urban residential area, the business area, the multi-family residential area, the office district area, the roadway area, the innovative area (non-traditional and new type of land use), and the planned unit development area were significantly positively associated with PDO crashes, while the single-family residential area was negatively related to PDO crashes.

Some injury crash studies at the macro-level have been conducted by Lerner et al. (2001), Noland (2003), Guevara et al. (2004), Noland and Oh (2004), MacNab (2004), Aguero-Valverde and Jovanis (2006), Quddus (2008), Nederan and Shashi (2010), and Pirdavani et al. (2013a); Pulugurtha (2013).

Lerner et al. (2001) conducted a retrospective chart review from patients of a trauma center for injuries from traffic crashes. Age, gender, race and ZIP code were used to identify significant factors of seatbelt use. ZIP code was a proxy for socioeconomic status by using census data. At

last, a logistic model revealed that younger people, male, African American, people with lower income and passengers are less likely to use seatbelts. Noland (2003) investigated the effect of changes in infrastructure on injury crashes. Total lane miles, average interstate lanes, average arterial lanes, the percentage of arterial lane miles, the percentage collector lane miles were positively contributed to injury crashes, whereas the percentage of interstate lane miles, the percentage of arterial with 9ft or less lane widths, the percentage of 10ft lane widths, the percentage of collectors with 9ft or less lane widths, and the percentage of collectors with 11ft lane widths were negatively affected to injury crashes. Moreover, the author concluded that the infrastructure improvements reduced injury crashes. Guevara et al. (2004) estimated the injury crash model and it was found that significant variables in the injury crash model were same in the PDO crash model. Thus, the population density, the numbers of employees, the intersection density, the percentage of other principal arterial, minor arterial and urban collectors were positively associated with the injury crash model.

MacNab (2004) used Local Health Area to explore relationships between injury rates and regional characteristics. The result indicated a large regional variation in the crash injury in males aged 0-24, and the socioeconomic influence on crash injury was apparent in young males more common in regions with deprived socio-economic status. Moreover, high adult male crime rates were significantly related with high injury rates of males aged from 1 to 14 years old. Aguero-Valverde and Jovanis (2006) focused on injury crashes in Pennsylvania. Authors showed that counties with lower daily VMT, higher roadway mileage, higher mileage roadway density,

higher percentage of federal aid roads, higher percentage of age groups from 0 to 14, 15 to 24 and 65 or more, and higher precipitations increased risk of injury crashes.

Quddus (2008) estimated spatial error models both for total slight injury crashes and motorized slight injury crashes using London crash data. It was shown that the traffic flow (PCU km/h), the motorway length, the minor road length, the number of employees, and the number of households without vehicles increased the total slight injury crashes, whereas the population aged 60 or over decrease the total slight injury crashes. Meanwhile, that the A-road length, and the minor road length was positively related to motorized slight injury crashes, whereas the population aged 60 or over was negatively associated with the motorized slight injury crashes. Pirdavani et al. (2013) investigated the safety effects of a fuel cost increase. Authors estimated injury crash models for both the null and the fuel-cost increase scenario. The result showed that a 20% increase of in fuel price decrease the annual VKT by 5.02 billion (11.57%), and thus the decrease of traffic volume reduced the total number of injury crashes by 2.83%.

Pulugurtha et al. (2013) found several significant land use factors for the injury crashes. Mixed use development, the urban residential, the business, the multi-family residential, the office district, the institutional, the roadway area and research district area had positive signs in the model whereas the industrial, and the single-family residential area had negative signs. Generally severe crashes are defined as crashes with incapacitating injuries or fatalities. Several researchers investigated the severe crashes and their contributing factors (Hadayeghi et al., 2003,

2006, 2010b; Quddus, 2008; Naderan and Shashi, 2010; Huang et al., 2010; Abdel-Aty et al., 2011; Abdel-Aty et al., 2013).

Hadayeghi et al. (2003) estimated macro-level severe crash prediction models in Toronto, Canada. Authors discovered that the VKT, the major road length, the number of household were positively related to the severe crash in the morning. In contrast, the volume-to-capacity and the posted speed limit had negative associations with the severe crash for the morning period. Hadayeghi et al. (2006) investigated the severe crashes during the morning period. Traffic related variables such as the VKT, the major road length, and the socioeconomic variable as number of employments was found positively associated with the severe crash. Other traffic related variables including volume-to-capacity and speed had the negative signs of estimates. Quddus (2008) found that the traffic related factors such as the traffic flow (PCU km/h) and average speed, roadway factors including the motorway length, A-road length and the minor road length, and socio-demographic factors as the resident population aged 60 or older, the number of employees and the number of households without vehicles were significant for the severe crash. Among these factors, only the average speed and the resident population aged 60 or over had negative signs, whereas all other variables had positive signs.

In the study of Naderan and Shashi (2010), the severe crashes were explained best by the crash generation model, which only used the total productions and attractions. The meaning of severe crashes in this study was different from other researches. They included fatal and injury crashes, thus minor injury crashes were also included in severe crashes in this study. Authors asserted that

the severe crash models performed better because data on injury or fatal crashes are more accurately collected and maintained by the police. Authors also found that the number of significant variable was also decreased compared to total and PDO crash generation models. Huang et al. (2010) found no significant difference in safety effects of risk factors on total and severe crashes at the county-level but increasing truck volume had tendencies to cause more severe crashes.

In the study of Abdel-Aty et al. (2011), authors showed that the severe crash model using only trip related variables were found best, among other models. The result revealed that home-based work attractions and heavy truck productions were negatively associated with the severe crashes. However, the home-based shop attractions, light truck productions and external-internal attractions had positive relationships with the severe crashes. Abdel-Aty et al. (2013) estimated the severe models based on CTs, BGs and TAZs, separately. The results found that the VMT, the roadway length with 35, 55 and 65mph, the number of intersections and median household income, were found also common significant variables in all severe crash models, regardless of geographical units. It is interesting that the variable, workers with shorter commute time (5-9 min), was also significant and negatively associated with severe crashes in all severe crash models. While the workers longer commute time (30 min or over) variable was positively associated with severe crashes in both the BG and TAZ based models. This result presented that the workers with the longer commute time were more vulnerable to severe crashes compared to workers with shorter commute time. It was also found that the population from 16 to 64 has a positive relationship with severe crashes in BG and TAZ based models. Furthermore, the number

of homeworkers and housing density were negatively related with severe crashes in BGs and TAZ based models. To sum up, many variables were commonly significant in BG/TAZ based models, but CT based model tended to have fewer and different significant variables compared to BG/TAZ models. Lee et al. (2013) explored the residential characteristics of at-fault drivers who caused severe crashes. It was revealed that zones with lower median age, smaller number of owner occupied households, more workers commuting by passenger cars or public transportation, and lower income had higher probability to have more at-fault drivers for severe crashes.

2.1.4 Studies on fatal crashes

Many studies have concentrated on fatal crashes, since they cause not only negative personal impacts (such as the pain, suffering, and economic hardship of victims and their friends and relatives), but also negative society impacts (including the public distress and the political, social, and economic problems), very seriously (Keeney, 1980).

Fatal crash studies at the macroscopic level usually focused on zonal characteristics where fatal crashes occurred (Blatt and Furman, 1998; Ng et al., 2002; Noland, 2003, Guevara et al., 2004; Noland and Oh, 2004; Aguero-Valverde and Jovanis, 2006; Quddus, 2008), while some researcher analyzed the residence of drivers (Blatt and Furman, 1998; Stamatiadis and Puccini, 2000; Clark, 2003; Romano et al., 2006; Males, 2009; Girasek and Taylor, 2010). Moreover, Blatt and Furman (1998) examined both the residence of drivers and crash locations. They explored the residence types of drivers involved in fatal crashes using ZIP codes of drivers from FARS based on county-level aggregation. Authors concluded that not only the majority of fatal

crashes occurred in rural area but also rural residents are more likely to be involved in fatal crashes.

Furthermore, Stamatiadis and Puccini (2000) concentrated on the Southeast United States which has higher fatality rates compared to other regions using ZIP codes and corresponding census data from FARS and the U.S. Census Bureau, respectively. Authors showed that higher percentage of the population below poverty levels, rural area and lower educated people affected the fatal crash rates in the Southeast. These socioeconomic factors were found significant for single vehicle fatal crash rates; however, they were not significant for multi vehicle fatal crash rates. Ng et al. (2002) revealed that the building materials storage floor area, the cinema seats, the commercial floor area, and market stalls had positive relationships between fatal crashes, whereas the container storage container storage area, hotel rooms, the country park, and the residential floor area were negatively associated with fatal crashes.

In addition, Moreover, Clark (2003) found out the population density of drivers' residence (using ZIP code), populations at crash location, age, seat belt use, vehicle speed and rural locations significantly affect the mortality after crashes. Noland (2003) revealed that total lane miles, average collector lanes, the percentage of arterial lane miles, the percentage of collector lane miles, and the percentage of collectors with 12ft or greater land widths had positive and significant effects on fatal crashes. On the other hand, the percentage of arterials with 10ft lane widths, the percentage of collectors with 9ft or less lane widths and collectors with 11ft lane widths had negative and significant effects on fatal crashes. The author also figured out that the

infrastructure improvement contributed to the reduction of fatal crashes. Furthermore, the author also found that demographic changes in age cohorts, increased seat-belt use, reduced alcohol consumption and increases in medical technology have contributed to overall reduction in fatal crashes, ceteris paribus.

Besides, Guevara et al. (2004) figured out that the population density, the percentage of minor population and intersection density were found significant in the fatal crash model. The population density was positively, and both the percentage of minor population and the intersection density were negatively related to the fatal crash. In the research of Noland and Oh (2004), only the mean lane width and the population were found significant. These two variables were significantly associated to fatalities in the model with year dummy variables and demographic variables. Aguero-Valverde and Jovanis (2006) conducted spatial analysis of fatal crashes in Pennsylvania. Authors found that the daily VMT and the total precipitation were negatively associated with the fatal crash, whereas the infrastructure mileage, the percentage of daily VMT on federal aid roads, the percentage of persons under poverty and the percentage of persons aged 0 to 14 were positively associated with the fatal crash their full Bayes model.

Romano et al. (2006) investigated the effect of race/ethnicity, language skills, income levels and education levels on alcohol-related fatal crashes. They collected fatal crash data including drivers' ZIP code and socioeconomic data from FARS and the U.S. Census Bureau, respectively. The authors confirmed that people with lower income and less education are more vulnerable to alcohol-related fatal crashes. Quddus (2008) revealed that the traffic flow (PCU km/h), the

motorway length, and the number of households without vehicle were significant and they are positively related to the fatal crash.

Males (2009) focused on the relationship between poverty and young drivers' fatal crashes. The author revealed that driver age itself is not a significant predictor of fatal crash risk once other factors associated with high poverty condition such as more occupants per vehicle; smaller vehicle size, older vehicle, lower state per-capita income and so forth were controlled. These factors were significantly associated with each other and with higher crash involvement among drivers from other age groups as well. Girasek and Taylor (2010) looked into the relationship between socioeconomic status based on ZIP code and vehicle characteristics such as crash test rating, electronic stability control, side impact air bags, vehicle age and weight. Specific vehicle data were collected from the Insurance Institute for Highway Safety using vehicle identification numbers (VINs). Authors revealed that lower income groups experience more risk since it is more likely that their vehicles are not safe enough.

2.1.5 Studies on bicycle and pedestrian crashes

There have been many efforts to analyzed bicycle and/or pedestrian crashes so far. Noland and Quddus (2004), Siddiqui et al. (2012) and Lee et al. (2013) analyzed the bicycle related crashes, and Abdel-Aty et al. (2011) estimated models for the combination of bicycle crashes and pedestrian crashes. Noland and Quddus (2004) found out several significant medical, roadway, vehicle, socio-demographic factors for bicycle related crashes by the injury severity level. Authors revealed that shorter length of inpatient stay in the hospital, larger national health

service staffs per population, higher percentage of motorway, trunk road density, older vehicle, higher percentage of households without cars, lower income, larger per capita expenditure on alcohol, larger population, lower percentage of population aged 45-64, and higher percentage of population aged 65 or over increased severe bicycle crashes. Authors also found that shorter length of inpatient stay in the hospital, more persons waiting for hospital treatment, and smaller trunk road density increased the number of cyclists with minor injuries.

Siddiqui et al. (2012) used the Bayesian Poisson-lognormal model accounting for spatial correlation for the bicycle crashes. Authors found out shorter roadway with 15 mph speed limits, longer roadway with 35 speed limits, more intersections, larger dwelling units, higher population density, urbanized area, the percentage of households with 0 or 1 vehicle, and total employment increased the bicycle crashes. Lee et al. (2013) investigated the residence of bicyclists who involved in the traffic crash. Authors estimated a NB model to analyzed contributing factors for the residence of bicyclists involved in the crash. It was revealed that the median age, the average travel time to work, the household income, and workers in primary industry fields were negatively associated with the residence of crash involved bicyclists. Meanwhile, Hispanic people, workers commuting by the bicycle, households in the urban area, and older buildings were positively associated with the residence of bicyclists who involved in the crash. Abdel-Aty et al. (2011) estimated four NB models for the sum of bicycle crashes and pedestrian crashes using different explanatory sets. Authors found that the pedestrian/bicycle crash model was best fit using only roadway factors. Authors discovered that the roadway length with 35 mph was

negatively associated with the pedestrian/bicycle crashes, but the number of intersections increased the number of pedestrian/bicycle crashes.

Pedestrian crashes have been considered serious issue, particularly in the urban area. Many researchers have conducted the pedestrian crashes at the macro-level, so far (LaScala et al., 2000; Ng et al., 2002; Noland and Quddus, 2004; Loukaitou-sideris et al., 2007; Wier et al., 2009; Cotrill and Thakuriah, 2010; Ukkusuri et al., 2011; Siddiqui et al., 2012; Siddiqui and Abdel-Aty, 2012; Wang and Kockelman, 2013; Abdel-Aty et al., 2013; Lee et al., 2013). LaScala et al., (2000) examined pedestrian injury rates across 149 CTs in the city of San Francisco. Authors found out the pedestrian injury rates were associated with traffic flow, population density, age composition of the local population, unemployment, gender and education. Ng et al. (2002) revealed that the number of cinema seats, commercial area, flatted factory area, market stall, and MTR catchment area were positively affected to the pedestrian crashes. Meanwhile, the greenbelt area, specialized factory area, and school places had negative relationships with pedestrian crashes in Hong Kong.

Noland and Quddus (2004) developed two pedestrian crash models for severe crashes and minor injury crashes. Authors figured out that the percentage of other road, the income, and the percentage of aged 45-64 decreased the severe pedestrian crashes, whereas the total population was negatively related with the severe pedestrian crashes. In regards to the minor injury crash of pedestrians, more persons waiting for hospital treatment, higher percentage of trunk road, higher income, and the percentage of population aged 45-64 had positive associations with minor injury

pedestrian crashes. On the other hand, the percentage of motorway, and the trunk road density were negatively associated with pedestrian related minor injury crashes. Loukaitou-sideris et al. (2007) explored the pedestrian collision based on CTs in the city of Los Angeles. Authors found out that pedestrian collisions are more likely to occur in neighborhoods with high population and employment density, high traffic volumes, and a large concentration of commercial/retail and multifamily residential land uses. Moreover, zones with high concentration of Latino population had a higher chance to have more pedestrian crashes per capita. Wier et al. (2009) investigated pedestrian crashes using 176 CTs of San Francisco. Authors showed that the traffic volume, arterials without transit, the proportion of land area zoned for neighborhood commercial and residential neighborhood commercial uses, employee and resident populations, and the proportion of people living in poverty, were found significant and positively affected to bicycle crashes. In contrast, land areas, and the proportion of population aged 65 or over had negative signs in the bicycle crash model.

Furthermore, Cotrill and Thakuriah (2010) analyzed the pedestrian crashes in deprived areas with many low-income and minority populations. Authors corrected the underreporting problem using a Poisson model, and found out the exposure including the suitability of the area for walking and transit accessibility), crime rates, transit availability, income, and presence of children were found significant for the pedestrian crashes. Ukkusuri et al. (2011) used CTs of New York City and discovered several socioeconomic and environmental factors for the frequency of pedestrian crashes using the NB model with random parameter. Siddiqui et al. (2012) found that the roadway length with 35 mph, the intersection, dwelling units, the

population density, the percentage of households with 0 or 1 vehicle, long term parking cost, and total employments had positive relationship with the number of pedestrian crashes, whereas the income reduced pedestrian crashes, from their Bayesian Poisson-lognormal model with a spatial error component. Siddiqui and Abdel-Aty (2012) estimated pedestrian crash models for interior crashes and boundary crashes, separately. Authors pointed out that the models could capture several unique explanatory variables explicitly related with interior and boundary crashes. For instance, total roadway length with 35 mph speed limit and long term parking cost were not significant in the interior pedestrian crash model but they were significant in the boundary. It was also found that hotel units were positively associated with interior crashes whereas it had a negative sign in the boundary crash model.

In recent, Wang and Kockelman (2013) studied the relationship between pedestrian crash frequency and land use, network and demographic attributes at the CT level. Authors revealed that the higher shares of residences near transit stops are associated with pedestrian crash risks. In addition, the provision of sidewalk is associated with lower pedestrian crash rates. Abdel-Aty et al. (2013) compared the pedestrian crash models based on different spatial units as CTs, BGs and TAZs. It was found that VMT and the number of intersections, the number of workers commuting by public transportation, the workers commuting by walking and the proportion of minority population were significant all models Moreover, roadways with relatively lower speed limits were positively associated with the pedestrian crashes in BGs/TAZ based models. Furthermore, the roadway with high speed limit (65mph) was significant and negatively associated with pedestrian crashes solely in the CT based model. Furthermore, the population of

the children aged from 0 to 15 was negatively related with pedestrian crashes in the BG/TAZ based models whereas the density of children (K to 12th grade) was positively associated with pedestrian crashes only in the TAZ based model. Variables related to workers by commute time were not significant in the pedestrian models except for the variable of workers with commute time 15 to 19 minutes, significant only in the TAZ based model. Furthermore, the number of home workers had a negative relationship with pedestrian crashes for the BG/TAZ based models. Lee et al. (2013) investigated the residential characteristics of pedestrians who involved in the traffic crash using the NB model. Authors discovered that people lived in the ZIP code area with lower median age, larger number of Hispanic people, more workers commuting by the public transportation, shorter travel time to work, lower income, older buildings, smaller number of workers in the primary industry field were more likely to be involved in pedestrian crashes.

As shown above, many studies have contributed to analyzing traffic crashes at the macro-level. Nevertheless, there are still several important issues in the macroscopic safety field. Following sub-chapter will discuss about the current issues of macroscopic safety modeling.

2.2 Current Issues of Macroscopic Safety Modeling

There have been several key issues in macroscopic safety studies. First issue is the spatial dependence of traffic crashes. Most statistical models assume that the values of observations in each sample are independent or randomly distributed. However, a positive spatial autocorrelation may violate this assumption, if the samples were collected from nearby areas (Lai et al., 2008). Many researchers found spatial autocorrelations in the traffic crash data. Second issue is

regarding boundary problem. Since TAZs are often delineated by arterial roads, most crashes occur on zone 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 (Siddiqui and Abdel-Aty, 2012). Last issue of the macroscopic safety research is the modifiable areal unit problem (MAUP), which is presented when artificial boundaries imposed on continuous geographical surfaces and the aggregation of geographic data cause the variation in statistical results (Openshaw, 1984). The detailed review of each issue will be conducted in the following sub-chapters.

2.2.1 Spatial autocorrelation problem

Spatial autocorrelation is a technical term for the fact that spatial data from near locations are more likely to be similar than data from distant location (O'Sullivan and Unwin, 2002). The existence of spatial autocorrelation in the crash data may invalidate the assumption of the random distribution (LeSage and Pace, 2004). Thus, it is required to test for the presence of spatial autocorrelations in the data set before the model estimation. If the spatial autocorrelation is detected, the crash model should account for spatial effect.

Many researchers showed that spatial autocorrelations are found in the traffic crash data. Levine et al. (1995) found that the introduction of the spatial lag effect improves the predictability of the total crash model. LaScala et al. (2000) discovered that there is a significant spatial relationship between pedestrian injury crashes and specific environmental and demographic characteristics of San Francisco. Hadayeghi et al. (2010a) estimated a series of spatial safety planning models

based on TAZs, and authors found that the spatial covariates from Full-Bayesian Semi-parametric Additive model showed that traffic crash frequencies aggregated by TAZs are spatially correlated. Hadayeghi (2010b) also developed the planning level transportation safety tools using Geographically Weighted Poisson Regression (GWPR). The authors concluded that the GWPR performs better than the conventional Generalized Linear Models in general. Nevertheless, Aguero-Valverde and Jovanis (2006) claimed that there is no spatial correlation in fatal crashes in the counties of Pennsylvania. In addition, Quddus (2008) asserted that the spatial autocorrelation decreases at higher levels of aggregation such as regions, counties and states whereas the spatial correlation increases at lower levels of spatial aggregations such as wards, enumeration districts postcode sectors, and super output areas. The author found that the different ward-level factors affect traffic crashes in a different way using London crash data and claimed that the Bayesian hierarchical models are more suitable for the area-wide traffic crash modeling. Huang et al. (2010) showed that the variation accounted for by spatial clustering are essential for crash risk models. The authors discovered that spatial autocorrelations were significant in traffic crashes across adjacent counties in Florida. Siddiqui and Abdel-Aty (2012) also asserted that the hierarchical Bayesian model accounting for spatial autocorrelation performs better in the pedestrian crash analysis.

2.2.2 Boundary crash problem

In the spatial analysis, boundary problems originate from the ignorance of interdependences that occurs from outside the boundary of zones (Fotheringham and Rogerson, 1993). In the macroscopic analysis, the boundary problem is much more crucial and unique. Since TAZs are

often delineated by arterial roads, most crashes occur on zone boundaries. Figure 2-3 displays an example of boundary crashes. The yellow lines in the figure are major arterials and red points show the location of crash occurred. As seen in the figure, the majority of crashes occur on or near the boundary of TAZs.

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 (Siddiqui and Abdel-Aty, 2012). Authors separated predictor sets for boundary and interior pedestrian crashes and estimated the hierarchical Bayesian Poisson-lognormal model. They found the separate considerations for interior and boundary crashes had a better model performance than the traditionally aggregated pedestrian crash model. Although the boundary crash issue is very crucial, not many studies have been done in regards to this issue. This is certainly an active area of further studies.

Figure 2-3: Example of boundary crashes in the downtown Orlando

2.2.3 Modifiable areal unit problem

MAUP is presented when artificial boundaries imposed on continuous geographical surfaces and the aggregation of geographic data cause the variation in statistical results (Openshaw, 1984). Assuming that areal units in a particular study were specified differently, it is possible that very different patterns and relationships are shown up (O'Sullivan and Unwin, 2002).

MAUP was first investigated by Gehlke and Biehl (1934). Authors found that the correlation coefficient increases as the unit area enlarges. According to Openshaw (1984), MAUP is composed of two effects: scale effects and zoning effect. Scale effects result from the different level of spatial aggregation. For example, traffic crash patterns are differently described in lower aggregation spatial units such as TAZs and higher aggregation units such as counties or states. Meanwhile zoning effects are from the different zoning configurations at a same level of the spatial aggregation. In the traffic safety field, Thomas (1996) explored how the length of segment affects the distribution of crash frequencies at the network level. The author found that the crash frequencies follow Poisson distribution for very small segments about 100 meters, they follow an intermediate empirical distribution for medium sized segments (300-2000 meters), and they are almost normally distributed for large segment (more than 200 meters). Thus, it was shown that the generalization made at one level of the spatial aggregation does not necessarily hold at another level (Thomas, 1996). This phenomenon occurs due to the scale effect of MAUP causing inconsistent statistical results based on different spatial units. Several transportation planning studies have addressed MAUP (Ding, 1998; Chang et al., 2002; Zhang and Kukadia, 2005; Viegs et al., 2009). Few studies have been conducted regarding the MAUP on

macroscopic traffic crash modeling to date. In recent, Abdel-Aty et al. (2013) compared three models based on different areal units such as CTs, BGs and TAZs. Although authors began to explore the MAUP in macroscopic traffic safety modeling, only three geographical units were used for the comparison. Moreover, since the size of two geographical units used in the study is quite comparable, it is thought that more extensive investigation of MAUP with more levels of the spatial aggregation in the macroscopic safety field is required. Abdel-Aty et al. (2013) compared models based on three different zone systems including CTs, BGs and TAZ. Authors found that the significance of explanatory variables was quite different among three models. Nevertheless, since TAZs are delineated for long term transportation plans, and two other zone systems were designed for the census, they have different zoning configurations. Moreover, their zone scales are different each other. Therefore, it was difficult to examine zonal and/or scale effects from MAUP because none of these effects were controlled.

2.2.4 Under-reporting problem

One of the minor issues, which is not only for the macroscopic safety studies but for all traffic safety fields, is the under-reporting crash problem. Hauer and Hakkert (1989) asserted that not all crashes are reportable, and not all reportable crashes were reported. Unreported data tend to produce biased estimations for crash models. Particularly, crashes without high severity, such as PDO crashes are more likely to be unreported (Ye and Lord, 2011). Some other researcher also pointed out the underreporting problem (Elvik and Mysen, 1999; Naderan and Shashi, 2010; Cotrill and Thakuriah, 2010).

The underreporting problem has been a crucial issue in the State of Florida. Two forms of crash report are used in Florida, short form and long form crash reports. A long form is used when the following criteria are met:

- Death or personal injury,
- Leaving the scene involving damage to attended vehicles or property (F.S. 316.061(1)), and
- Driving while under the influence (F.S. 316.193).

Whereas a short form is used to report other types of PDO traffic crashes. Since only long form crashes have been coded and archived in FDOT crash analysis reporting (CAR) database so far, previous researchers only could get access to long form crashes for the crash analysis of Florida. Thus, safety analysts had many missing crash data in Florida, especially for PDO crashes. Fortunately, MPO started to code short crashes recently, thus short form crash data for the three counties of Central Florida were also obtainable. Therefore, more complete data can be used in this study. As shown in Figure 2-4, crash data without short form reports (long form only data) had 52.3% of injury crashes, which is even larger than PDO crashes (43.3%). In contrast, in complete data set, only 25.5% were injury crashes whereas 72.4% were PDO crashes, which is obviously reasonable. Using data with many missing PDO crashes may result in biased model estimation. Thus, complete data including short from and long form data were used for both macroscopic and microscopic analyses in this study.

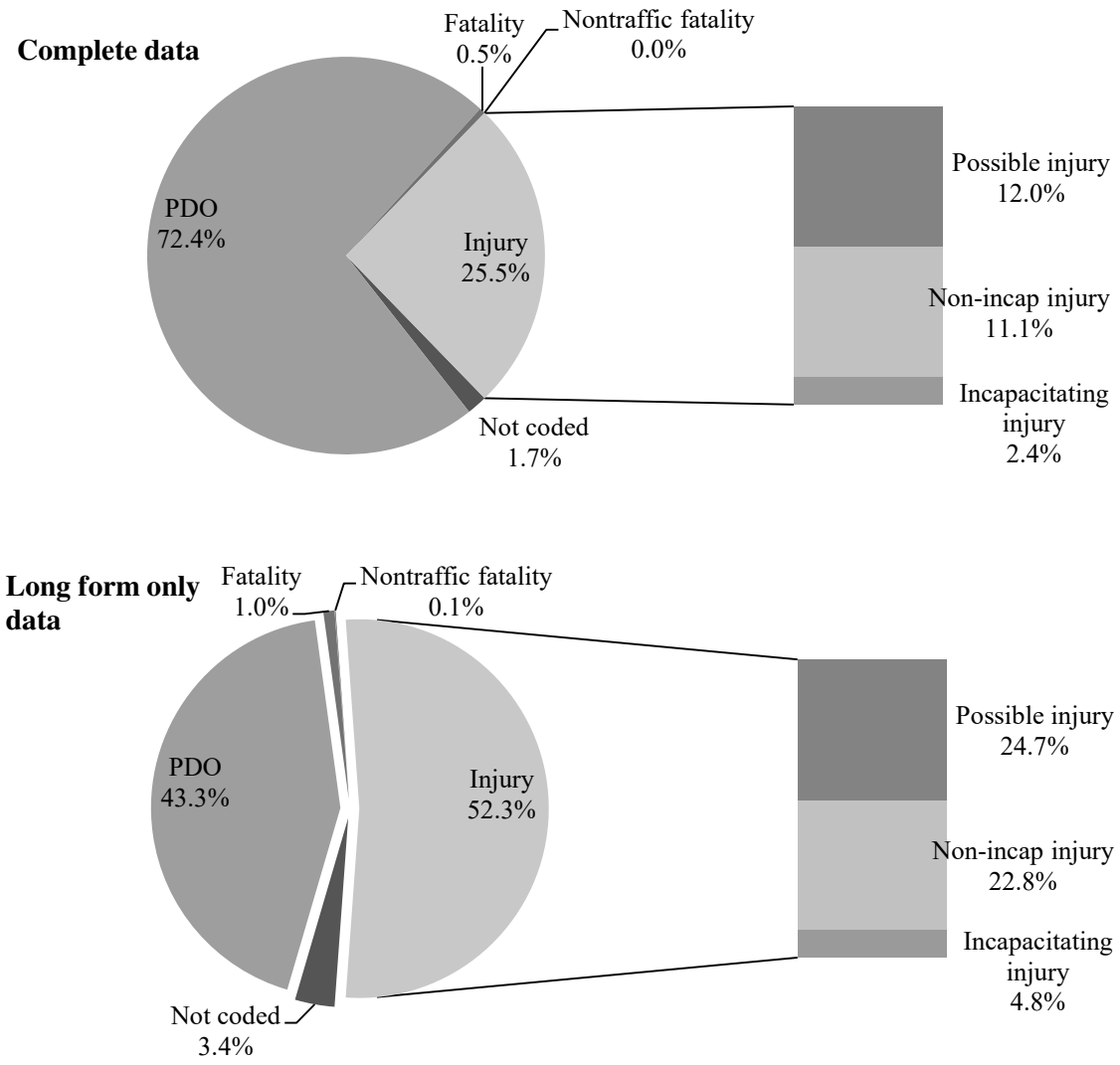


Figure 2-4: Comparison of complete crash data and long form crash data

2.3 Regionalization

As described previously, many researchers used TAZs as basic spatial units for their macroscopic safety studies (Ng et al., 2002; Hadayeghi et al., 2003, 2010a, 2010b; Guevara et al., 2004; Hadayeghi et al., 2006; Naderan and Shashi, 2010; Abdel-Aty et al., 2011; Siddiqui and Abdel-Aty, 2012; Siddiqui et al., 2012; Wang et al., 2012; Pirdavani et al., 2012; Huang et al.,

2013; Pirdavani et al., 2013a, 2013b; Pulugurtha et al., 2013; Abdel-Aty et al., 2013). TAZs may be reasonable for the traffic safety research purpose because they are only transportation based zone system.

However, it is required to investigate whether TAZs are appropriate spatial units for the macroscopic traffic safety modeling. General zoning criteria for TAZs are as follow (Baass, 1981).

1) Homogeneous socioeconomic characteristics for each zone's population.
2) Minimizing the number of intra-zonal trips.
3) Recognizing physical, political, and historical boundaries.
4) Generating only connected zones and avoiding zones that are completely contained within another zone.
5) Devising a zonal system in which the number of households, population, area, or trips generated and attracted are nearly equal in each zone.
6) Basing zonal boundaries on census zones.

Criteria 1), 4), 5) and 6) are also sensible for the macroscopic modeling. Nevertheless, possible limitation of TAZs for the crash analysis arise from criteria 2) and 3). Basically, TAZs were designed to find out origin-destination pairs of trips generated from each zone. Thus, transportation planners need to minimize trips which start and end in the same zone. It is thought that minimizing intra-zonal trips end up with the small size of TAZs. On the other hand, traffic safety analysts need to analyze traffic crashes that occurred inside the zone, so they are able to

relate zonal characteristics with traffic crash patterns of the zone. Therefore, it is possible that TAZs are too small to analyze traffic crashes at the macroscopic level. Moreover, the small size of zones makes many zones with zero crash frequency, especially for rarely occurring crashes such as severe, fatal or pedestrian crashes. Criterion 3) indicates that TAZs are usually divided based on physical boundaries, mostly arterial roadways. Considering that many of crashes occur on arterial roads, between zones, inaccurate results will be made from relating traffic crashes on the boundary of the zone to only the characteristics of that zone (Siddiqui and Abdel-Aty, 2012). A simple way to overcome these two issues from using TAZs for the safety analysis is to aggregate TAZs into sufficiently large and homogenous traffic crash patterns. This process is called regionalization (Guo and Wang, 2011).

Recently, Huang et al. (2013) conducted extensive research of the regionalization. Authors suggested the regionalization scheme for TSP in the paper. Authors recommended the partitioning scheme of 350 or 400 zones to keep the balance between model fitting and variable significance. However, it would be more beneficial if zonal effects are also explored, and if more various types of explanatory variables are used in the modeling.

2.4 Statistical Methodologies

A wide array of statistical techniques for the safety analysis has been developed. Lord and Mannering (2010) summarized the statistical methods for crash count data. In this dissertation, statistical models widely used for crash count models, including Poisson, negative binomial (NB), Poisson-lognormal, random-effects, and multivariate models will be briefly addressed.

Since crash frequency data are non-negative integers, the application of ordinary least squares regression is obviously not suitable. For crash count analyses, Poisson regression models have been used for several decades (Jovanis and Chang, 1986; Joshua and Garber, 1990; Jones et al., 1991; Miaou and Lum, 1993; Miaou, 1994).

The equation of Poisson model is as follows:

$$\mathrm{P}(y_i) = {exp(-\lambda_i)\lambda_i^{y_i}}\Big/{y_i!} \quad (1)$$

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

$$\lambda_i = exp(\beta X_i) \quad (2)$$

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

The Poisson model is the most basic and also easy to estimate, but it cannot handle over-dispersion and under-dispersion. Over-dispersion is one of characteristics of crash count data, which is that the variance exceeds the mean of the crash frequencies. Over-dispersion can violate the assumption of the equal mean-variance. Under-dispersion is that the mean of the crash counts is greater than the variance. Incorrect parameter is estimated in the existence of under-dispersed

data. Moreover, the Poisson model is largely affected by low sample-mean and small sample size bias.

The NB, or Poisson-gamma model, is an extension of the Poisson model to overcome the over-dispersion problem. NB model assumes that the Poisson parameter follows a gamma probability distribution. The NB model is derived by manipulating the relationship between the mean and the variance. The equation of NB model is as follow:

$$\lambda_i = exp(\beta X_i + \varepsilon_i) \quad (3)$$

where, i is each observation, $exp(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α. The addition of variance term α allows the variance to differ from the mean as:

$$\mathrm{VAR}[y_i] = E[y_i][1 + \alpha \mathrm{E}[y_i] = E[y_i] + \alpha \mathrm{E}[y_i]^2 \quad (4)$$

Abdel-Aty and Radwan (2000) also pointed out that both Poisson and NB regression models had been widely used since they can well represent crash count data. Additionally, the negative binomial model performs better than Poisson regression model, especially when over-dispersion existed.

Although NB model still cannot handle under-dispersion problem and it is also affected by low sample-mean and sample size bias, The NB model has been most frequently used model in crash count model (Maycock and Hall, 1984; Hauer et al,. 1988; Miaou, 1994; Persaud, 1994; Kumala, 1995; Shankar et al., 1995; Poch and Mannering, 1996; Milton and Mannering, 1998; Karlaftis and Tarko, 1998; Persaud and Nguyen, 1998; Abdel-Aty and Radwan, 2000; Carson and

Mannering, 2001; Miaou and Lord, 2003; Amoros et al., 2003; Guervara et al., 2004; Hirst et al., 2004; Abbas, 2004; Lord et al., 2005; Wang and Abdel-Aty, 2006; El-Basyouny and Sayed, 2006; Lord, 2006; Kim and Washington, 2006; Lord and Bonneson, 2007; Lord et al., 2010; Malyshkina and Mannering, 2010; Daniels et al., 2010; Cafiso et al., 2010; Naderan and Shashi, 2010; Abdel-Aty et al., 2011; Ukkusuri et al., 2011; and Lee et al., 2013).

In recent, several traffic crash studies have been conducted using Poisson-lognormal models (Miaou et al., 2003; Lord and Miranda-Moreno, 2008; Aguero-Valverde and Jovanis, 2008; Abdel-Aty et al., 2013). The Poisson-lognormal model was suggested as an alternative to the NB model for crash count data. The Poisson-lognormal model is similar to the NB model, but the $exp(\varepsilon_i)$ is a lognormal rather than gamma-distributed. Admittedly, the Poisson-lognormal can provide more flexibility compared to the NB model, it has two disadvantages: 1) model estimation is more complicated, and 2) still negatively affected by small sample sizes and low sample-mean values (Miaou et al, 2003).

The correlation among observations could arise from spatial and/or temporal considerations. Random-effects model and fixed-effects models can be considered to account for such correlation. Random effects model is considered where the common unobserved effects are assumed to be distributed over the spatial/temporal units according to some distribution and shared unobserved effects are assumed to be uncorrelated with explanatory variables. Fixed-effects models are considered where common unobserved effects are accounted for by indicator variables and shared unobserved effects are assumed to be correlated with explanatory variables.

Random effects models modify the Poisson parameter as $\lambda_{ij} = exp(\beta X_{ij})exp(\eta_j)$, where λ_{ij} is the expected number of crashes for roadway entity i belonging to group j, and $\eta_j$is a random effect for observation group j. The most common model is derived by assuming $\eta_j$is randomly distributed across group such that $exp(\eta_j)$is gamma-distributed with mean one and variance α. The Poisson model limits the mean and variance to equal ($\mathrm{E}[y_{ij}] = VAR[y_{ij}]$), but the Poisson variance to mean ratio is $1+{\lambda_{ij}}/{1/\alpha}$, with random effects.

Random effects model was firstly explored by Hausman et al. (1984) for count data, random-effects for have been used by many researchers in the traffic crash studies (Johansson, 1996; Shankar et al., 1998; Miaou et al., 2003; Yu et al., 2013).

Bivariate/multivariate models become necessary when modeling specific types of crash counts. Modeling the counts of specific types of crashes cannot be done with independent count models since the counts of specific crash types are not independent. For instance, the number of fatal crashes cannot increase or decrease without affecting the counts of PDO or injury crashes. This problem can be solved using bivariate/multivariate models since they obviously consider the correlation among the severity levels for each roadway entity (Miaou and Song, 2005; Bijleveld, 2005; Song et al., 2006; Lord and Mannering, 2010).

Multivariate models have been employed such as multivariate Poisson model (Ma and Kockelman, 2006), the multivariate NB model (Winkelman, 2003), and the multivariate Poisson-lognormal model (Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009; Park et al., 2010).

Recently, application of Bayesian approach became popular for the traffic crash studies. Bayesian methods provide a comprehensive and robust approach to model estimation. Moreover, Bayesian models do not depend on the assumption of asymptotic normality underlying classical estimation methods as maximum likelihood. Traditional estimation methods such as the least square estimation are designed to find single estimate point. However, Bayesian estimation focuses on the entire density of parameters. For instance, in classical statistics the prediction of out-of-sample data often involves calculating moments or probabilities from the assumed likelihood for *y*, $p(y|\theta_m)$, which is evaluated at the selected point estimate θ_m (Congdon, 2001). However, the information about θ is contained in the posterior density $p(\theta|y)$ in the Bayesian method; thus prediction is estimated based on averaging $p(y|\theta)$ over this posterior density (Congdon, 2001).

Congdon (2003) argued that, among the benefits of the Bayesian approach are a more natural interpretation of parameter intervals, often termed Bayesian credible or confidence intervals, and the freedom of obtaining true parameter density. On the contrary, maximum likelihood estimates rely on normality approximations based on large sample asymptotic. New estimation methods

assist in the application of Bayesian random effects models due to pooling strength across sets of related units (Congdon, 2003).

Several researchers (Miaou et al., 2003; Aguero-Valvarde and Jovanis, 2006; Quddus, 2008; Abdel-Aty, 2013) applied the Bayesian method for the macroscopic safety modeling.

2.5 Network Screening Techniques

Network screening is a process for reviewing a transportation network to identify and rank sites with respect to safety risk, then rank from most likely to least likely to realize a reduction in crash frequency with the implementation of a countermeasure.

There is a growing body of literature that has focused on the development of traffic safety network screening. The majority of these studies are specifically on the micro level, which deals with the safety screening of road segments or intersections, or other types of spots. Several methods have been developed in the last a few decades. The principal network screening methods include:

- Table C method
- Level of Safety Service (LOSS) method
- Empirical Bayes (EB) methods
- Continuous Risk Profile (CRP) for highway segments
- Screening based on high proportions
- Detection of safety deterioration over time

Table C method (Ragland et al., 2007) identifies sites that have experienced considerably greater number of crashes per unit of ADT than the average. For highway segments, the roadway segments are screened by sliding window of size 0.2 miles and in increments of 0.02 miles. On the other hand, for intersections, the influence area is 250 feet from the intersection and all crashes within the influence area are considered as intersection crashes. The criteria for a hotspot are: 1) the observed crash frequency is more than the average for the rate group with 99.5% confidence level in either the 3, 6, or 12 months period, and 2) four or more crashes in the given time period.

Level of Service of Safety was proposed by Kononov et al. (2003). LOSS method is similar to the Table C method in that the observed crash frequency is compared to an expected crash frequency and the level of deviation is measured. The Table C method considers whether the deviation is large enough for the statistical significance that more crashes occurred than would be expected for the average site. While in the LOSS method, the deviation from the expected for an average site is shown by creating 4 categories of service levels. The expected LOSS for similar sites is determined by safety performance functions (SPFs) using traffic volume, number of lanes, lane width and so forth. Through using SPFs, the LOSS method would be superior compared to Table C method by removing the use of constant crash rates.

The EB methods have started with its application in traffic safety by Abbess et al. (1981). The EB methods are defined as a suit of screening methods that are based on the EB method of estimating the long-term expected crash frequency for a location. These methods were included

as a preferred method in the Highway Safety Manual. The EB estimate of expected crash frequency for a location is a weighted combination of the prediction from an SPF and the observed crash frequency for the location. The weights are calculated based on the EB that makes use of the over-dispersion parameter that is an outcome of the SPF development using a negative binomial model.

Continuous Risk Profile (CRP) method was introduced by Chung et al. (2007). CRP deals only with observed crashes. The key concept of the CRP is that a continuous profile plot of risk along a roadway can be helpful to identify sites of high risk.

Screening based on high proportions was suggested by Heydecker et al. (1991). This method identifies and ranks the sites that have a proportion of a specific crash type relative to the total crashes that is higher than some average or threshold proportion value for similar road types.
Hauer (1996) developed a methodology that detects safety deterioration over time. In this method, two tests are conducted: The first one is to detect potential gradually increasing trend in mean crash frequency. The second is to detect a potential sudden increase in mean crash rate and can be ranked if necessary.

Lastly, Highway Safety Manual (HSM) (2010) summarized many of the previously developed methods and presented clear processes. There are five steps in micro-level network screening. First of all, establishing focus identifies the purpose or the intended outcome of the network screening analysis. In the second step, identifying network and establishing reference

populations specify the types of sites of facilities being screened (e.g. segments, intersections, etc.) and identifying grouping of similar sites or facilities. In the next step, fourteen performance measures are provided to measure the expected crash frequency or some other equivalent values at a site. Followed this step, a few methods are provided for screening. There are three principle screening methods: they are ranking method, sliding window method and peak searching method. The final step in the network screening process is to conduct the screening analysis and evaluate results.

Although several network screening methods have been developed and used, no screening methods are developed for the macroscopic level so far. One of objective of this study is to suggest zone-level screening methods with borrowing ideas from the network screening techniques.

2.6 Summary and Conclusion

Considerable researches have been conducted to analyze traffic crashes at the macroscopic level. Various spatial units are used for the macroscopic modeling, such as BGs, CTs, or TAZs. In recent TADs were delineated for more macroscopic transportation plans. The possibility of TADs for the geographic unit in the safety modeling needs to be explored in the follow-up study.

Many studies have analyzed total crashes, crashes by severity levels (PDO, injury, severe, fatal, etc.), or crashes by transportation mode (bicycle, pedestrian, etc.). There are several important issues in the macroscopic safety study. They are; 1) spatial autocorrelation, 2) boundary crashes,

3) MAUP. First, spatial autocorrelation should be test in the data set before further analysis. However, it is assumed that the effect of spatial autocorrelation is reduced as the zone scale increases (Huang et al., 2013). Second, boundary problem will remarkably decrease after the regionalization. Lastly, two effects of MAUP, scale and zonal effects will be thoroughly explored, and thus TSAZs, optimized for the macroscopic safety analysis can be built.

Several network screening methods have been used. However, zone screening methods are not developed for the macroscopic level. Zone-level screening is required to be developed with ideas from existing network screening methods.

CHAPTER 3 ANALYSIS OF RESIDENCE CHARACTERISTICS OF DRIVERS INVOLVED IN CRASHES

3.1 Introduction

As discussed previously, it is important to explore zonal-level crash patterns and their contributing factors on safety at the macroscopic level. Although many recent studies have concentrated on zones where the crash occurred, there have been few studies that focused on residence characteristics associated with the origin of the drivers causing traffic crashes, so called at-fault drivers. Intuitively, it is reasonable to assume that the number of at-fault drivers is related to socio-demographic features of at-fault drivers' residence area. Although, most of macro-level traffic safety studies focused only on crash locations, maybe it is not sensible to relate crash locations aggregated by zones with zonal characteristics, if an at-fault driver came from the zone that is far from the crash site.

One of the reasons that make it hard to focus on the residence characteristics is the difficulty of the drivers' data collection. Usually, the residence information of at-fault drivers is difficult to obtain. Fatality Analysis Reporting Systems (FARS) offers postal codes (ZIP codes) of at-fault drivers who caused fatal crashes (Stamatiadis and Puccini; 2000; Clark, 2003; Romano et al., 2006; Males, 2009), however, FARS does not provide driver information causing less severe crashes (i.e., property only damage, minor injury crash, etc.). Fortunately, Florida Department of Transportation (FDOT) provided ZIP code information of at-fault drivers of the State of Florida for the research. Therefore, it was possible to analyze the significant socio-demographic characteristics associated with the origin of the drivers causing traffic crashes using complete

data. In this chapter, overall 509,882 crashes were used for the analysis, which is considered a large enough sample size for the modeling. Moreover, various zonal-level demographic, socio-economic and commute characteristics were obtained from the U.S. Census Bureau and they were used as candidate factors. Therefore, the objective of this chapter is to investigate the relationship between at-fault drivers for all types of crashes and residence characteristics of the at-fault drivers, using a large sample with various socio-demographic variables.

Many macroscopic safety studies have been conducted using a wide array of spatial units. Levine et al. (1995) and Abdel-Aty et al. (2013) used a census block group as a basic spatial unit in their studies. Also a census tract which is larger than a block group is also used in the macro-level studies (LaScala et al., 2001; Abdel-Aty et al., 2013). Political boundaries were used in several studies, as counties (Aguero-Valverde and Jovanis, 2006) and states (Noland, 2003). Furthermore, a traffic analysis zone, which is only related to transportation/traffic, has been widely used in macro-level safety studies recently (Siddiqui, 2009; Washington et al., 2010; Naderan and Shahi, 2010; Abdel-Aty et al., 2011; Abdel-Aty et al., 2013).

These prior studies focused on crash locations aggregated by specific geographic units. On the other hand, some researcher focused on the residence of drivers involved in crashes (Blatt and Furman, 1998; Lerner et al., 2001; Clark, 2003; Romano et al., 2006; Males, 2009; Stamatiadis and Puccini, 2000; Girasek and Taylor, 2010), instead of the crash location. Most of these studies used ZIP codes as geographical units for the analysis because the residence information is typically provided as a form of ZIP code. For example, FARS offers ZIP codes of drivers

involved in crashes. Blat and Furman (1998) examined the residence types of drivers involved in fatal crashes using residence ZIP codes of drivers obtained from FARS, based on county-level aggregation. The authors concluded that not only the majority of fatal crashes occurred in rural area but also rural residents are more likely to be involved in fatal crashes. Lener et al. (2001) conducted a retrospective chart review from patients of a trauma center for injuries from traffic crashes. Age, gender, race and ZIP code were used to identify significant factors of seatbelt use. ZIP code was a proxy for socioeconomic status by using census data. A logistic model revealed that younger population, male, African American, people with lower income and passengers are less likely to use seatbelts.

Moreover, Clark (2003) found that the population density of drivers' residence (using ZIP code), populations at crash location, age, seat belt use, vehicle speed and rural locations significantly affect the mortality after crashes. Romano et al. (2006) investigated the effect of race/ethnicity, language skills, income levels and education levels on alcohol-related fatal crashes. They collected fatal crash data including drivers' ZIP code and socioeconomic data from FARS and the U.S. Census Bureau, respectively. The authors pointed out that people with lower income and less education are more likely to cause alcohol-related fatal crashes. Males (2009) focused on the relationship between poverty and young drivers' fatal crashes. The author revealed that driver age itself is not a significant predictor of fatal crash risk once other factors associated with high poverty condition such as more occupants per vehicle; smaller vehicle size, older vehicle, lower state per-capita income and so forth were controlled for. These factors were significantly

associated with each other and with higher crash involvement among drivers from other age groups as well.

Furthermore, Stamatiadis and Puccini (2000) concentrated on the Southeast United States which has higher fatality rates compared to other regions using ZIP codes and corresponding census data from FARS and the U.S. Census Bureau, respectively. The authors showed that higher percentage of population below poverty levels, rural area and lower educated people affected the fatal crash rates in the Southeast. These socioeconomic factors were found significant for single vehicle fatal crash rates; however, they were not significant for multi-vehicle fatal crash rates. Girasek and Taylor (2010) looked into the relationship between socioeconomic status based on ZIP code and vehicle characteristics such as crash test rating, electronic stability control, side impact air bags, vehicle age and weight. Specific vehicle data were collected from the Insurance Institute for Highway Safety using vehicle identification numbers (VINs). Authors revealed that lower income groups experience more risk since it is more likely that their vehicles are not safe enough.

In addition, a Bayesian Poisson-lognormal model was adopted in this study. Bayesian estimation concentrates on estimating the entire density of a parameter whereas classical estimation methods target finding a single fixed estimate. In recent traffic safety research studies, Bayesian models have been popular since they have advantages over traditional likelihood based inference methods (Park and Lord, 2007; Aguero-Valverde and Jovanis, 2008, 2009; Ma et al., 2008; El-Basyouny and Sayed, 2009, 2010; Abdel-Aty et al., 2013).

To sum up, there have been only a few studies that have focused on the relationship between the number of at-fault drivers and residence zonal characteristics, although many macro-level studies have been conducted. While few studies have focused on residence characteristics, however, these studies only investigated fatal crashes (Stamatiadis and Puccini, 2000; Clark, 2003; Romano et al., 2006), young driver related fatal crashes (Males, 2009), safety equipment usage (Girasek and Taylor, 2010). Thus, it is important to examine at-fault drivers in all types of crashes with a large sample size and accounting for various new explanatory variables.

3.2 Data Preparation

In order to examine the residence characteristics of drivers involved in traffic crashes, two types of data are required. First, we need the aggregated number of at-fault drivers based on a specific spatial unit. At-fault drivers are defined as the drivers who caused traffic crashes. After police officers investigate traffic crashes, they issue citations to drivers who are responsible for traffic crashes. All at-fault drivers, regardless of crash types, are recorded in the crash report with their personal information. However, detailed address is not coded due to privacy concerns. Only ZIP code information of road users involved in traffic crashes have been coded and archived by FDOT. Thus, solely ZIP codes were possible to be used as a spatial unit for the analysis. Second, we need the corresponding demographic and socioeconomic data based on the same spatial unit. These data were obtained from the U.S. Census Bureau, which provides various socio-demographic data as well as commute pattern data.

The data preparation process is presented in Figure 3-1. According to the crash report data from FDOT 518,008 drivers received citations in traffic crashes during 3 years. After excluding observations with missing ZIP codes or from out-of-state ZIP codes, 509,882 observations for driver residence were retained and they were used in the analysis. As potential independent variables for the models, overall 16 candidate independent variables were prepared. They include demographic data such as proportions of age groups, socioeconomic data such as educational attainment and median family income, and commute patterns. Table 3-1 summarizes the variable description of the data. In modeling log transformation of total population was used since they have very large values.

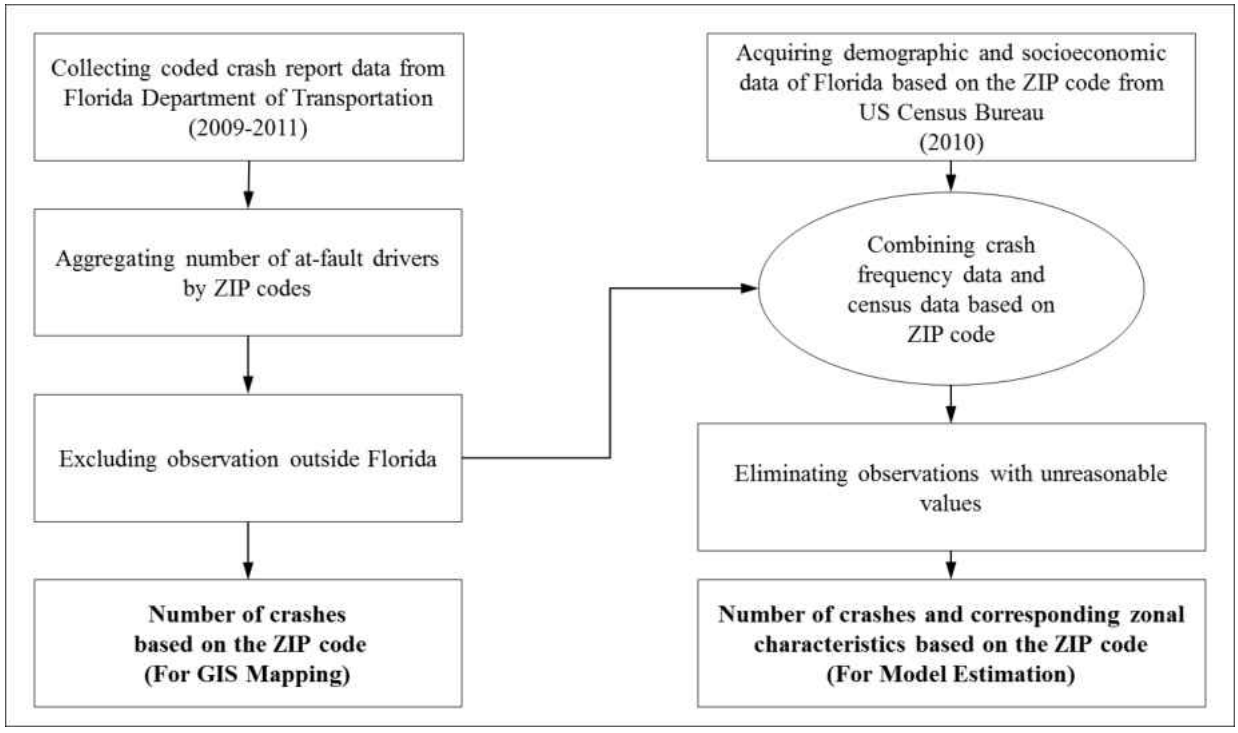


Figure 3-1: Data preparation process

Demographic variables prepared are ‘Proportion of young people’ between 20 and 24 years old, and ‘Proportion of elderly people’ aged 65 years old or more. Also, commute pattern variables such as ‘Proportion of commuters using public transportation’ and ‘Proportion of commuters using non-motorized modes’, which means proportion of people commuting by bicycle or walking. Moreover, ‘Proportion of people working at home’ was also considered because they do not need to commute though they have jobs; so their travel pattern should be different from other people who need to commute. Commute time variables were also included as possible contributing factors. In order to consider, commute times both ‘Proportion of commuters whose commute time less than 15 min’ and ‘Proportion of commuters whose commute time more than 45 min’ were prepared.

Additionally, occupation related variables were prepared. The primary sector is the retrieval and production of raw materials (i.e., agriculture, forestry, fishing, etc.). The second sector is the transformation of raw (or intermediate) materials into goods (i.e., manufacturing, construction, etc.) Lastly, the tertiary sector involves supply of services to consumers and businesses (i.e., transportation, wholesale trade, finance, public administration, etc.) (Kenessey, 1987). For ‘Whether median year of structure built before 1984’, one is assigned if the median year of structure built is before 1984, whereas the value is zero if it is 1984 or after. 1984 is the median year of structure built in the whole Florida.

Before the model estimation, multi-collinearity between variables was investigated using Pearson correlation coefficients. It was found that there are several correlations among variables. It was also revealed that ‘Proportion of people whose educational attainment lower than high

school' is somewhat correlated with 'Proportion of workers in the primary sector' (r=0.532), 'Proportion of workers in the tertiary sector' (r=-0.520) and 'Median family income' (r=-0.517). Regarding the occupation sectors, 'Proportion of workers in the tertiary sector' is somewhat related with 'Proportion of workers in primary sector' (r=-0.553). Any variables that have potential correlations between other variables were not included simultaneously in the same model, and careful consideration of the correlated variables separately was attempted before reaching the final model.

Figure 3-2 displays ZIP code areas with top 15% of at-fault drivers per population. Even though the number of at-fault drivers is normalized by population, hot zones for at-fault drivers are located in the urban area. However, these hot zones are not only in large metropolitan areas such as Miami, Tampa/St. Petersburg and Jacksonville but also in mid-sized cities such as Pensacola, Palm Beach and Gainesville.

Table 3-1: Descriptive statistics of variables

Variables (N=983)	**Mean**	**Stdev**	**Median**	**Min**	**Max**
Number of at-fault drivers by ZIP codes	518.7	547.9	356	1	4549
Population	19126.4	14561.9	16875	0	72248
Proportion of young people (20-24 years old)	0.063	0.049	0.055	0.000	0.643
Proportion of elderly people (65 years old or more)	0.189	0.116	0.161	0.000	1.000
Proportion of commuters using public transportation	0.017	0.046	0.006	0.000	1.000
Proportion of commuters using non-motorized modes	0.026	0.052	0.015	0.000	1.000
Proportion of people working at home	0.054	0.070	0.042	0.000	1.000
Proportion of commuters whose commute time less than 15 min	0.263	0.143	0.244	0.000	1.000
Proportion of commuters whose commute time more than 45 min	0.157	0.108	0.135	0.000	1.000
Proportion of workers in the primary sector	0.115	0.084	0.100	0.000	1.000
Proportion of workers in the secondary sector	0.093	0.059	0.087	0.000	1.000
Proportion of workers in the tertiary sector	0.780	0.138	0.800	0.000	1.000
Proportion of households without available vehicle	0.027	0.034	0.018	0.000	0.462
Proportion of people whose educational attainment lower than high school	0.152	0.107	0.133	0.000	1.000
Proportion of unemployed people	0.102	0.053	0.099	0.000	0.545
Median family income (in 1,000 of US Dollars)	50.023	18.796	46.770	9.979	250.00
Whether median year of structure built is before 1984	0.507	0.500	1.000	0.000	1.000

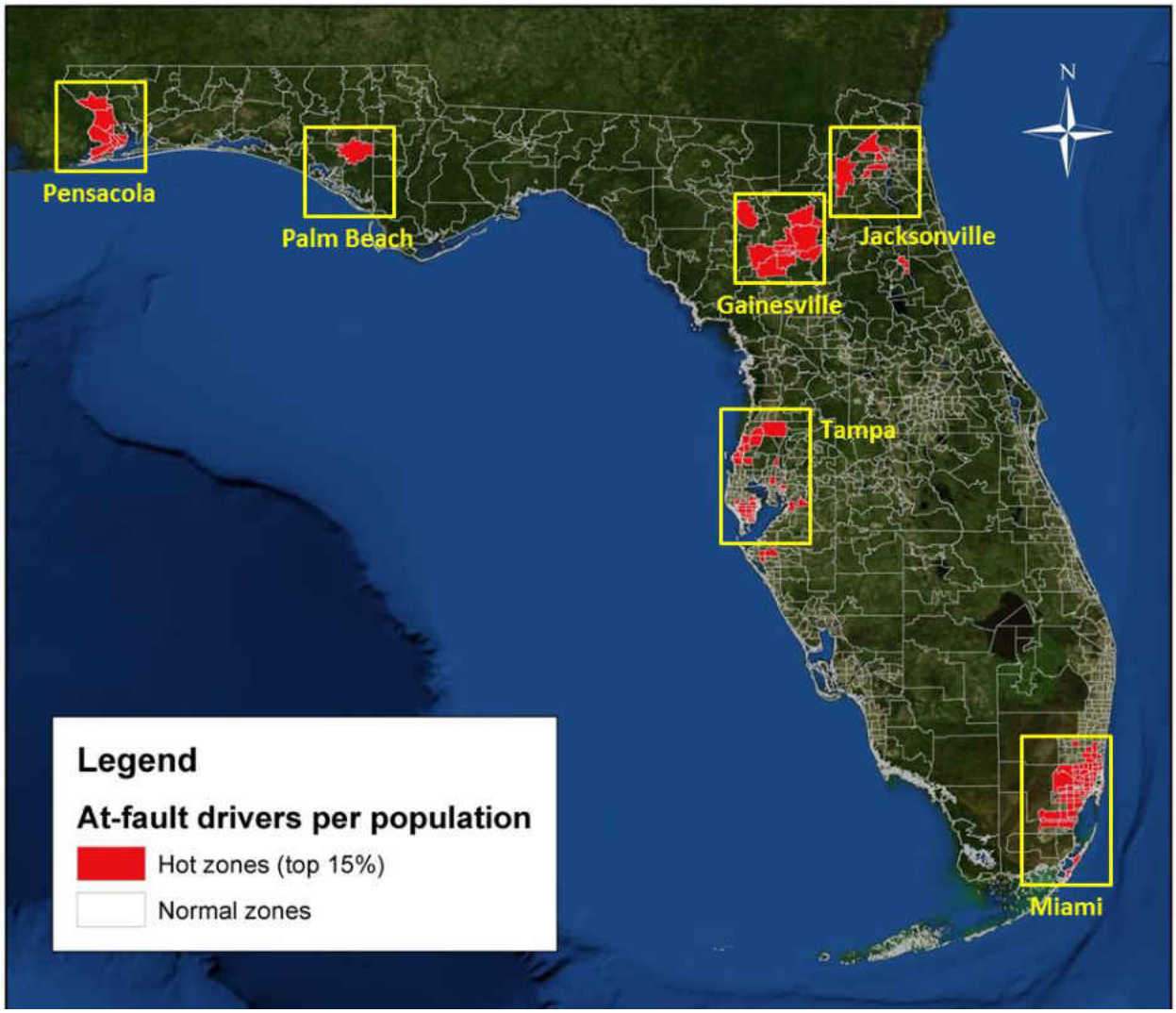


Figure 3-2: Hot zones with top 15% at-fault driver per population by ZIP codes

3.3 Methodology

Since crash count data are non-negative integers, it is not appropriate to apply ordinary regression models, which are used for continuous response variables. For crash count analyses, Poisson regression models have been used over the last several decades. The Poisson model is the most basic and also easy to estimate, however, it cannot handle over-dispersion. Over-dispersion is one of the characteristics of crash frequency data, which is that the variance exceeds

the mean of the crash frequency. Over-dispersion can violate the assumption of the equal mean and variance. Poisson-lognormal is an extension of the Poisson model to overcome the over-dispersion problem. Poisson-lognormal assumes that the Poisson parameter follows a lognormal distribution. Poisson-lognormal can provide more flexibility compared to a Poisson-gamma model which is most widely used in traffic safety studies (Lord and Mannering, 2010). Poisson-lognormal models have been widely applied in recent traffic safety modeling (Park and Lord, 2007; Aguero-Valverde and Jovanis, 2008, 2009; Ma et al., 2008; El-Basyouny and Sayed, 2009, 2010; Abdel-Aty et al., 2013).

In this chapter, Poisson-lognormal models were fitted in a Bayesian framework for the number of at-fault drivers. In recent years, an application of the Bayesian approach became popular in traffic safety research. Bayesian methods provide a comprehensive and robust approach to model estimation. Moreover, Bayesian models do not depend on the assumption of asymptotic normality underlying classical estimation methods as maximum likelihood. Classical estimation methods such as the least square estimation are designed to find single estimate point. In contrast, Bayesian estimation focuses on the entire density of parameters. For instance, the prediction of out-out-sample data in traditional statistics often involves calculating moments or probabilities from the assumed likelihood for y, $p(y|\theta m)$, which is evaluated at the selected point estimate θm. However, the information about θ is contained in the posterior density $p(\theta|y)$ in the Bayesian method; thus prediction is estimated based on averaging $p(y|\theta)$ over this posterior density (Congdon, 2001). Congdon (2003) argued that, among the benefits of the Bayesian approach are a more natural interpretation of parameter intervals, often termed Bayesian Credible Intervals

(BCI), and the freedom of obtaining true parameter density. On the contrary, maximum likelihood estimates rely on normality approximations based on large sample asymptotic. Novel estimation methods assist in the application of Bayesian random effects models due to pooling strength across sets of related units. Thus, Bayesian models are thought to have several advantages compared to the classical likelihood based inference methods.

A Poisson-lognormal model is specified as follows:

$$y[i] \sim Poisson(\mu[i]) \quad (5)$$

$$log(\mu[i]) = \beta_0 + \beta X_i + \theta[i] \quad (6)$$

$$\theta[i] \sim Normal\ (0, \tau_\theta) \quad (7)$$

where,

$\beta_0 =$ intercept term,

βs are the coefficient estimates of model covariates X_i

$\theta[i] =$ error component of the model, and

$\tau_\theta =$ precision parameter which is inverse of the variance; τ_θ follows a prior gamma (0.5, 0.005)

This variance ($1/\tau_\theta$) provides the amount of variation not explained by the Poisson assumption (Lawson et al., 2003).

The models should control for exposure by including 'Log of population' variable in each model. Without controlling for exposure, it is not reasonable to interpret other variables since it is

expected that zones with more population should have more at-fault drivers. Thus, 'Log of total population' should be included as a surrogate of the exposure measure for all candidate models. The model was run considering a non-informative Normal (0, 1000000) prior for β s, including β_0. The significance of the model parameters were judged based on BCI. BCI infers on the true parameters value. For instance, a 95% BCI will contain the true parameter value with approximately 95% certainty. Deviance Information Criterion (DIC) was used to decide the best variable set. We determined the final model which has the smallest DIC among candidate models. DIC was calculated using the following equation (Spiegelhalter et al., 2002).

$$DIC = 2 \times \overline{D} - \widehat{D} \quad (8)$$

where,

$\overline{D} =$ posterior mean of deviance

$\widehat{D} = -2 \times \log(p(y|\bar{\theta}))$, and

$\bar{\theta} =$ posterior mean of θ, respectively.

3.4 Modeling Results

Table 3-2 presents three candidate models for at-fault drivers for traffic crashes. Independent variables only with 90% BCI were included in the candidate models. All models have non-significant intercepts. It implies that the number of at-fault driver can be well explained by variables in the models.

Initially, all candidate variables except for variables with multi-collinearity were included in the preliminary model, and non-significant variables without 90% certainty are all excluded in

Model 1. Model 1 has 7 variables including ‘Proportion of elderly people’, ‘Proportion of people working at home’, ‘Proportion of workers people working at home’, ‘Proportion of commuters whose commute time is less than 15 minutes’, ‘Proportion of workers in the tertiary sector’, ‘Proportion of households without available vehicle’ and ‘Median year of structure built’.

In Model 2, commute mode related factors such as ‘Proportion of commuters using public transportation’ and ‘Proportion of commuters using non-motorized modes were attempted. However, ‘Proportion of commuters using public transportation’ was not significant, while ‘Proportion of commuters using non-motorized modes’ was statistically significant with 90% certainty. Furthermore, ‘Proportion of households without available vehicle’ became non-significant after including ‘Proportion of commuters using non-motorized modes’. It seems there is an interaction between ‘Proportion of commuters using non-motorized modes’ and ‘Proportion of households without available vehicle’, although the Pearson correlation coefficient indicates there is no high correlation (r=0.189). Thus, both ‘Proportion of commuters using public transportation’ and ‘Proportion of households without available vehicle’ were not included. Finally, Model 2 has 7 variables and they are ‘Log of population’, ‘Proportion of elderly people’, ‘Proportion of commuters using non-motorized modes’, ‘Proportion of people working at home’, ‘Proportion of commuters whose commute time less than 15 min’, ‘Proportion of workers in the tertiary sector’ and ‘Whether median year of structure built before 1984’.

Model 3 attempted to include ‘Median family income’. However, ‘Median family income’ became non-significant. It was shown that the income factor has a very slight correlation with

‘Whether median year of structure built before 1984’ (r=0.289). Also, ‘Proportion of workers in the tertiary sector’ was slightly correlated with the income variable (r=0.366). Although Pearson correlation coefficients indicate there is no high correlation, there could be interactions among these variables. After excluding both ‘Whether median year of structure built before 1984’ and ‘Proportion of workers in the tertiary sector’, the income variable became significant with 90% certainty. Eventually, Model 3 has 6 significant variables including ‘Log of population’, ‘Proportion of elderly people’, ‘Proportion of commuters using non-motorized modes’, ‘Proportion of people working at home’, ‘Proportion of commuters whose commute time less than 15 min’, ‘Median family income’.

Even though all three candidate models have significant variables that can explain at-fault driver counts, Model 3 has been selected as the final model since it has the smallest DIC. According to Spiegelhalter et al. (2003), differences of more than 10 might rule out the model with higher DIC. Differences between 5 and 10 are considered substantial. Since the differences of DIC of Model 1 and Model 2 between Model 3 are 7.44 and 5.24, respectively, it is concluded that Model 3 outperforms other two models and thus it was chosen as the best model. Lastly, Table 3-3 presents the Pearson correlation coefficients between variables used in the final model, Model 3. It was shown that no obvious correlations were observed between the independent variables included in the final model.

Table 3-2: Result of at-fault driver model estimation

Variable	Model 1				Model 2				Model 3			
	Mean	Stdev	BCI		Mean	Stdev	BCI		Mean	Stdev	BCI	
			5%	95%			5%	95%			5%	95%
Intercept	-1.362[#]	1.119	-2.036	1.245	-1.371[#]	1.190	-2.043	1.423	-1.045[#]	1.365	-1.817	2.688
Log of population	0.832[*]	0.083	0.671	0.881	0.842[*]	0.089	0.706	0.898	0.789[*]	0.105	0.514	0.851
Proportion of elderly people (65 years old or more)	-0.986[*]	0.278	-1.505	-0.642	-0.928[*]	0.304	-1.400	-0.581	-1.077[*]	0.401	-2.027	-0.667
Proportion of commuters using non-motorized modes					0.852[*]	0.515	0.060	1.735	1.158[*]	0.468	0.398	1.946
Proportion of people working at home	-0.610	0.388	-1.224	-0.079	-0.673[*]	0.405	-1.334	-0.140	-0.915[*]	0.326	-1.457	-0.382
Proportion of commuters whose commute time is less than 15 min	-1.187[*]	0.318	-1.739	-0.875	-1.309[*]	0.386	-2.070	-0.941	-1.277[*]	0.510	-2.727	-0.829
Proportion of workers in the tertiary sector	-0.555[*]	0.370	-1.598	-0.122	-0.646[*]	0.370	-1.602	-0.273				
Proportion of households without available vehicle	1.723[*]	0.664	0.712	2.708								
Median family income (in 1,000s of US Dollars)									-0.003	0.004	-0.013	-0.00004
Whether median year of structure built before 1984	0.235[*]	0.057	0.117	0.310	0.223[*]	0.066	0.102	0.300				
Standard deviation of θ_i	0.564[*]	0.060	0.527	0.658	0.565[*]	0.067	0.527	0.668	0.595[*]	0.077	0.550	0.779
DIC	9288.48				9286.34				9281.04			

[#] not significant at 10%, [*] significant at 5%, and all other variables are significant at 10%

Table 3-3: Pearson correlation coefficients for variables in Model 3

Variables	Log of population	Proportion of elderly people (65 years old or more)	Proportion of commuters using non-motorized modes	Proportion of people working at home	Proportion of commuters whose commute time is less than 15 min	Median family income (in 1,000s of US Dollars)
Log of population	1					
Proportion of elderly people (65 years old or more)	0.063	1				
Proportion of commuters using non-motorized modes	-0.113	-0.020	1			
Proportion of people working at home	-0.172	0.111	0.256	1		
Proportion of commuters whose commute time is less than 15 min	-0.224	0.211	0.447	0.236	1	
Median family income (in 1,000s of US Dollars)	0.019	-0.038	-0.040	0.244	-0.054	1

3.5 Discussion of the Result

It is revealed that Model 3 had the best model performance, in terms of DIC. Thus, the variables found significant in Model 3 are mainly discussed. First of all, the intercept term was not significant, which may imply that the number of at-fault drivers can be explained with variables in the model. Furthermore, 'Log of population' is significant at the 5% level It is included as a surrogate of the exposure measure, as stated previously. Unsurprisingly, it has a positive relationship with the number of at-fault drivers.

Besides, 'Proportion of elderly people' is significant at the 5% level. According to Florida Department of Highway Safety and Motor Vehicles (2011), the crash rate (crashes per 10,000 licensed drivers) of elderly people aged 65 years or more is 98.75, which is only half of the total crash rate of 197.05. However, many researchers have found that elderly drivers have functional deficiency which increase crash risks (Owsley, 2003). The result from the models, which shows that the high percentage of elderly people reduces the number of at-fault drivers, may not seem consistent with other studies. This inconsistency can be explained by the degree of exposure. Although elderly drivers are more likely to be involved in traffic crashes, their average trip length is much shorter than those of other age groups. The National Highway Travel Survey (NHTS) (Santos et al., 2011) shows that the average daily vehicle trip length of the working age group (25-64 years old) is 35.3 miles whereas that of elderly people (65 years old or more) is just 22.6 miles. Thus, it is thought that because elderly people are less exposed to traffic, and thus the community with high percentage of elderly people has less number of at-fault drivers. This might not be the case if we consider crash severity or type.

Concerning the commute mode, it was shown that 'Proportion of commuters using non-motorized modes' was one of significant factors at the 5% level. It had a positive association with the number of at-fault drivers. Beck et al. (2007) claimed that the annual fatality rate (fatality per person-trips) of bicyclist or pedestrians is much higher than that of the overall fatality rate. According to FARS and NHTS statistics, the overall fatality rate is 10.4 per 100 million person-trips, whereas the fatality rates of bicyclists and pedestrians are 21.0 and 13.7 per 100 million person-trips, respectively (Beck et al., 2007). Moreover, Pucher and Dijkstsra (2003) showed that pedestrians and bicyclists are 23 and 12 times more likely to be killed than car occupants in traffic crashes, correspondingly, in the United States. These two studies only examined fatalities from traffic crashes and they concluded that both bicyclists and pedestrians have higher crash risks; however, the result from this chapter shows that bicyclists and pedestrians may be not only vulnerable for fatal crashes but also for overall crashes.

In addition, 'Proportion of people working at home' was found significant at the 5% level. It was negatively related with the number of at-fault drivers. Naturally, home workers are less exposed to traffic crashes since they are not required to commute every day. Thus, it is expected that the community with more home workers has less number of at-fault drivers. In a recent study by Pirdavani et al. (2013c), it was shown that increased number of teleworkers reduces the total Vehicle Kilometers Traveled (VKT) in Belgium. Also, it is expected that the number of crashes decreases, correspondingly. Abdel-Aty et al. (2013) found that the number of home workers has a negative relationship with total, severe and pedestrian crashes, which is consistent with the result from this chapter.

Moreover, ‘Proportion of commuters whose commute time is less than 15 min’ was significant at the 5% level and it was negatively related with the number of at-fault drivers. This variable shows the commuters’ exposure to the traffic, which is similar to ‘Proportion of people working at home’. As the commute time is shorter, the community has less number of at-fault drivers since commuters are less exposed to the traffic. This result is in line with the study conducted by Abdel-Aty et al. (2013). Abdel-Aty et al. (2013) included commute time related variables in the crash model such as the number of workers whose travel time is 5-9 minutes and also found that the community with short commute time has less number of severe crashes.

Also, it was found that ‘Median family income’ factor was significant at 10%. It had a negative relationship with the number of at-fault drivers. It is interpreted that drivers from the low-income community are more likely to cause traffic crashes. Martinez and Veloz (1996) asserted that people from lower socioeconomic status might be less likely to purchase newer and safer vehicles or equipment, and also hard to get information regarding traffic safety, thus low-income families are more likely to be involved in traffic crashes. Huang et al. (2010) and Abdel-Aty et al. (2013) also found that economically deprived areas have more traffic crashes, if all other conditions remain constant.

Aside from the traffic safety field ,there have been many efforts to find out the effect of demographic and socioeconomic characteristics of the residence in medical studies (Smith et al., 1996; Sundquist et al., 2004), psychology (Ross, 2000; Cutrona et al., 2006) and criminology (Gruenewald et al., 2006; Gyimah-Brempong; 2006). These studies commonly suggested

deprived economic conditions, such as low-income level, are significant factors for higher rates of mortality, specific diseases including the disease, depression and criminal activities as well. It is interesting to note that there are several socioeconomic factors for traffic crashes that are commonly significant for crimes and diseases in the other field.

3.6 Summary and Conclusion

According to New South Wales Roads and Traffic Authority (1996), human factors contribute to 95% of traffic crashes, whereas 28% and 8% of crashes are due to road environment factors and vehicle factors, respectively. Nevertheless, most of the previous safety studies have focused on physical roadway characteristics of segments, intersections, corridors or zones where traffic crash occurred. In contrast, this chapter aimed at investigating the effect of residence characteristics of drivers, as surrogates for individual human factors. In this chapter, it was assumed that demographic and/or socioeconomic characteristics of the residence of at-fault drivers are more related to traffic crash occurrence than those of the crash location.

In order to find out the relationship between the zonal characteristics and number of at-fault drivers, the Bayesian Poisson lognormal model was applied in this chapter and the result revealed that the exposure measure as 'Log of population', and 'Proportion of commuters using non-motorized modes' of residence zones were positively associated with the number of at-fault drivers. On the other hand, 'Proportion of elderly people', 'Proportion of people working at home', 'Proportion of commuters whose commute time less than 15 min' and 'Median family income' of residence zones had negative relationships with the number of at-fault drivers. It

could be concluded that traffic crashes are a socio-economic problem related to the deprived socio-economic status and specific demographic conditions. The final model in this chapter revealed that there are several demographic, socioeconomic and commute patterns of residence zones that contribute to crash occurrence. The findings could be used to identify residence areas with people who are more likely to be involved in traffic crashes.

Several demographic and socioeconomic variables that are found significant in this chapter are also commonly used in transportation planning. Thus, the results from this chapter can provide guidance to transportation planners and enable them to take traffic safety into account in the long-term planning. For instance, if an area are expected to have more elderly people, the area will be more likely to have lower traffic risk in the future, compared to other areas with less elderly people. Some results can be used for establishing transportation/traffic policies. For example, it was found that 'Proportion of people working at home' reduces the number of at-fault drivers since they would be less exposed to traffic crashes. Pirdavani et al. (2013c) asserted that the increased number of telecommuters reduces not only traffic congestion but also traffic crash risk. A local government with both heavy congestion and many crashes may need to encourage telecommuting by providing effective incentives. Similarly, locating workplace close to workers' residences can be considered at the stage of urban planning to reduce traffic crash risk (Gurstein, 1996), since 'Proportion of commuters whose commute time less than 15 min' is negatively associated with the number of crashes by at-fault drivers.

Also, the results from this chapter are also important for designing and tailoring specific education, engineering and awareness campaigns and stricter enforcement to reduce traffic crashes. Whittam et al. (2006) evaluated the effectiveness of safety campaigns using TV, radio and billboard. A time series analysis revealed that there was a 21.6% decrease in young at-fault drivers. Also, Philbrook et al. (2009) assessed if the safe driving campaign is effective. The Drive Smart Challenge campaign aimed to increase safe driving habits for high school students. The authors found that the campaign increased the seat belt use by15%, in the most improved school. Also we can consider engineering treatments such as providing safe facilities for bicycles for the community with the large number of non-motorized mode users. Moreover, Bates et al. (2012) asserted that traffic law enforcement is effective to reduce the number of DUI (Driving under the Influence), over-speed, and red-light running related crashes. Traffic enforcement officers can be more efficiently dispatched to specific zones with the large number of predicted at-fault drivers using the result from this chapter.

It is important to note that there are several possible extensions to this chapter. First, only residence characteristics were explored in this chapter. Although the residence characteristics of at-fault drivers play a key role in crash occurrence as shown in this chapter, the crash occurrence is also equally affected by the traffic and roadway characteristics of the crash location. A combined analysis for both residence and location characteristics should be addressed. Besides, only total number of at-fault drivers was examined in this chapter. Drivers who caused specific crashes such as crashes by severity levels (property damage only, injury, fatal, etc.),

transportation modes (motor vehicle, bicycle, pedestrian, etc.) or special types (DUI, hit-and-run, etc.) may have different contributing factors.

CHAPTER 4 MULTIVARIATE MODELING FOR MOTOR VEHICLE, BICYCLE, AND PEDESTRIAN CRASH ANALYSIS

4.1 Introduction

As shown in Chapter 2, a large body of literature have investigated traffic safety propensity on the macroscopic level, such as block group (Levine et al., 1995), traffic analysis zone or TAZ (Siddiqui, 2009; Washington et al., 2010; Naderan and Shahi, 2010; Abdel-Aty et al., 2011; Abdel-Aty et al., 2013), census tract (Loukaitou-Sideris et al., 2007; Wier et al., 2009; Cottrill and Thakuriah, 2010; Ukkusuri et al., 2011), county (Aguero-Valverde and Jovanis, 2006; Amoros et al., 2003; Huang et al., 2010; Noland and Oh, 2004), state (Noland, 2003), and others (Noland and Quddus, 2004; MacNab, 2004; Kim et al., 2006). Among these spatial units, TAZ has been widely adopted for traffic safety analysis. A TAZ is a statistical entity delineated by state Departments of Transportation (DOTs) and/or local Metropolitan Planning Organizations (MPOs) officials for tabulating traffic-related census data such as, journey-to-work and place-of-work statistics (U.S. Census Bureau, 2011). Therefore, from a transportation planning perspective, TAZs seem to be preferred spatial entities as compared to other spatial units.

Previous macroscopic safety studies have investigated the occurrence of crashes by various classifications from different perspectives, such as crashes by pedestrian and/or bicycle crashes (Noland and Quddus, 2004; Kim et al., 2006; Abdel-Aty et al., 2011; Siddiqui et al., 2011; Abdel-Aty et al., 2013; Lee et al., 2013), crashes by injury severity levels (Hadayeghi et al., 2003; Noland, 2003; Noland and Oh, 2004; Hadayeghi et al., 2006; Aguero-Valverde and Jovanis, 2006; Hadayeghi et al.; 2010; Naderan and Shashi, 2010; Huang et al., 2010; Abdel-Aty, 2013),

or crashes during specific time periods (Hadayeghi et al., 2003; Hadayeghi et al., 2006; Abdel-Aty et al., 2011). Among these viewpoints, crashes by specific transportation modes such as motor vehicles, bicycles and pedestrians have been investigated by several researchers.

From a methodology perspective, a wide spectrum of modeling approaches has been incorporated in the macro-level safety research so far. The conventional methods include NB models (Hadayeghi et al., 2003, 2006; Siddiqui, 2009, Noland and Quddus, 2004, Karlaftis and Tarko, 1998; Amoros and Laumon, 2003; Noland and Oh, 2004; Aguero-Valverde and Jovanis, 2006), ordinary least square regression model (Wier et al., 2009), log-linear models (Washington, 2006), Bayesian hierarchical models (Quddus, 2008) and Bayesian models accounting for spatial autocorrelation (Huang et al., 2010; Wang et al., 2012).

These models are generally univariate, which assume the occurrence of each type of crashes is independent. However, as has been proven by many research studies, there is strong correlation among crash frequencies of different types (i.e., severity levels) within each site (Tunaru, 2002; Ma and Kockelman, 2006; Aguero-Valverde and Jovanis, 2009; Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009). This correlation is caused by some site specific unobserved factors that may affect traffic safety. Similarly, this type of correlation may also exist in macro-level crash frequencies. Ignoring these correlations may lead to biased parameter estimates and thus the corresponding crash frequency prediction (Ye et al., 2009; Ma et al., 2008). In order to address this problem multivariate models have been proposed and widely used in micro-level crash analysis (Tunaru, 2002; Ma and Kockelman, 2006; Park and Lord, 2007; Ma

et al., 2008; El-Basyouny and Sayed, 2009; Ye et al., 2009). For example, Ma and Kockelman (2006) applied a multivariate Poisson regression model using crash data from roadway segments in Washington State. It was found that there was positive correlation in unobserved factors affecting crash frequencies across severity levels. Also goodness-of-fit measures showed that the multivariate Poisson model is superior to the suite of independent models. Ye et al. (2009) looked into crashes by collision types such as head-on, sideswipe, rear-end, and angle crashes at intersections. The authors developed univariate and multivariate Poisson regression models and compared the two model structures. No significant differences were found in magnitude of coefficients between the two model systems. Nevertheless, with respect to goodness-of-fit measures (adjusted likelihood ratio index), the multivariate model showed better fit than the univariate model. In addition the authors found that the results revealed the presence of common unobserved factors across collision types. Although, many studies have been done in the micro-level crash analysis, few studies have addressed the potential correlations between each type of crashes in the macroscopic safety analysis. Guevara (2004) utilized simultaneous estimation of the models for injury and fatal crashes using TAZ crash data. Although the authors were not successful to show simultaneous models outperformed the independent models in terms of goodness-of fits, they found there is a significant correlation across disturbance terms of injury and fatal models.

At the same time, few researchers have reported that there is a spatial correlation for macro-level crashes. Spatial autocorrelation is a technical term for the fact that spatial data from near locations have higher probabilities to be similar than data from distant location (O'Sullivan and

Unwin, 2002). The existence of spatial autocorrelations in the data may invalidate the assumption of the random distribution (LeSage and Pace, 2004). Several researchers have accounted for spatial effects in the crash model (Levine et al., 1995; Hadayeghi, 2010; Quddus, 2008; Huang et al., 2010; Siddiqui and Abdel-Aty, 2012). . LaScala et al. (2000) discovered that there is a significant spatial relationship between pedestrian injury crashes and specific environmental and demographic characteristics of San Francisco. Huang et al. (2010) showed that the variation accounted for by spatial clustering are essential for crash risk models. The authors discovered that spatial autocorrelations were significant in traffic crashes across adjacent counties in Florida. Siddiqui and Abdel-Aty (2012) developed Bayesian Poisson-lognormal models with spatial error terms for bicycle and pedestrian crashes in Florida. The authors revealed that models accounting for spatial autocorrelations showed better performance compared to models without considering spatial effects.

Previous papers have significantly contributed to addressing the correlations among crash types, as well as the potential spatial correlations among zones. However, limited studies have account for these two issues simultaneously. Song et al. (2006) developed Bayesian multivariate conditional autoregressive (MCAR) models that can account for the spatial effect using county based data of Texas. Authors classified crashes into four types by crash locations such as intersection, intersection-related, drive access and non-intersection crashes and modeled them. Authors revealed that MCAR model performs much better than independent models without spatial effects. Nevertheless, up till now no study has investigated the potential correlations among different crash modes with spatial effect in the macroscopic studies.

Therefore, the objective of this chapter is to develop the model for crashes by transportation modes such as motor vehicle, bicycle and pedestrian crashes, which accounts for potential correlations among spatial effects at the macroscopic level. Moreover, we used not only roadway and traffic variables but also demographic and socioeconomic variables which are generally used for long term transportation plans. Thus Traffic Analysis Zone (TAZ) based multivariate models estimated in this chapter can be more usefully utilized by transportation planners.

4.2 Data Preparation and Methodology

Data from three counties in Central Florida; Orange, Seminole and Osceola Counties were used for the analysis. The three counties are composed of 1,116 TAZs. Crashes occurring between 2008 and 2009 were collected from Florida Department of Transportation and MetroPlan Orlando (the metropolitan planning organization for the Orlando area).

Since data are zone based data, the magnitude of values may be highly affected by the zone size; thus, zonal based data are likely to be highly correlated. For example, 'Roadway lengths with speed limit less than or equal to 20 mph' is not supposed to be related to 'Number of young people', but they could be correlated in the macroscopic analysis since as the zone size is larger both roadway and demographic data also increase. The original demographic, socioeconomic and roadway/traffic data were processed to minimize correlation among predictors. Overall 18 variables were prepared and descriptive statistics were summarized in Table 4-1. Three crash response variables by modes, i.e., motor vehicle, bicycle and pedestrian crashes, were collected from both Florida Department of Transportation and MetroPlan Orlando to ensure the completeness of the data. Besides, five demographic factors were collected from the U.S. Census

Bureau including population density, proportions of African Americans, Hispanics, young people between 15 and 24 years old and elderly people older or equal to 65 years old. Vehicles available by household, hotel, motel, timeshare room, employment and school enrollment data were provided by the MPO. Lastly, the number of intersections, number of traffic signals, roadway length by speed limits, roadway length with poor pavement condition, and vehicle miles traveled (VMT) were collected from Florida Department of Transportation and they were transformed into TAZ based data using ESRI ArcMap 10.0 GIS software. The roadway with poor pavement conditions is defined as the roadway that is virtually impassable or has large potholes and deep cracks.

Before the model estimation, multi-collinearity between variables was investigated using Pearson product-moment correlation. For example, variables 'Log of vehicle-miles-traveled' and 'Log of population density' intuitively seemed to have collinearity since both of these variables can show the urbanization of the area. Nevertheless it was found that the Pearson correlation between these variables was 0.060, which shows there was only weak correlation between them. Meanwhile Pearson correlation between 'Proportion of African Americans' and 'Proportion of household without vehicles' was largest among independent variables and it was 0.373. However, none of correlation coefficients amongst independent variables indicated any obvious correlation.

Table 4-1: Descriptive statistics of the data (N=1116)

Description	Mean	Stdev	Min	Max
Motor vehicle crash	78.384	88.836	0	713
Bicycle crash	0.765	1.249	0	10
Pedestrian crash	1.065	1.650	0	13
Vehicle-miles-traveled	93948	100954	0.0	839660
Log of vehicle-miles-traveled	10.493	2.541	0.000	13.641
Population density (population/mi^2)	1189.1	1437.3	0.0	14904.5
Log of population density	5.307	2.994	0.000	9.609
Proportion of African Americans	0.176	0.225	0.000	1.000
Proportion of Hispanics	0.237	0.188	0.000	1.000
Proportion of young people (15-24 years old)	0.145	0.087	0.000	1.000
Proportion of elderly people (65 years old or older)	0.116	0.106	0.000	1.000
Proportion of household without vehicles	0.069	0.084	0.000	0.557
Number of rooms of hotel, motel and time shares	296.2	1355.1	0	14341
Log of rooms of hotel, motel and time shares	1.155	2.473	0.000	9.571
Employments and school enrollments	1918.1	3364.4	0	54819
Log of employments and school enrollments	6.593	1.707	0.000	10.912
Number of intersections	13.909	11.958	0	78
Number of traffic signals	0.737	1.268	0	9
Proportion of roadway with speed limit less than or equal to 20 mph	0.048	0.083	0.000	1.000
Proportion of roadway with speed limit more than or equal to 55 mph	0.059	0.135	0.000	0.908
Roadway length with poor pavement condition	0.150	0.448	0.000	5.126

The Poisson regression based models play a key role in analyzing crash frequency data. The Poisson regression model has been broadly used since it can cope with non-negative integers. The probability of entity (zone, segment, intersection, etc.) i having y_i crashes per time period is given by:

$$\mathrm{P}(y_i) = {exp(-\lambda_i)\lambda_i}^{y_i} \Big/ y_i! \tag{9}$$

where, $\mathrm{P}(y_i)$ is the probability of entity i having y_i crashes per some time period, and

λ_i is the Poisson parameter, which shows the expected number of crashes per period,

Poisson regression models are estimated by specifying λ_i (Poisson parameter) as a function of explanatory variables, the most widely used functional form is:

$$\lambda_i = exp(x_i\beta) \tag{10}$$

where, x_i is a row vector of explanatory variables of entity i, and

β is a coefficient estimate of model covariates x_i.

Nevertheless Poisson models cannot manage both over- and under-dispersion in the data since it assumes mean and variance are equal. Hence, the Poisson lognormal model was suggested as one of the alternative models to Poisson models to account for the over-dispersion of crash data (Lord and Mannering, 2010).

Multivariate Poisson-lognormal models with the spatial error term in a Bayesian frame were estimated for three response variables by modes, i.e., motor vehicle, bicycle and pedestrian.

Different from the classical regression models, Bayesian models do not depends on the assumption of asymptotic normality. Sampling based methods of Bayesian estimation focus on estimating the entire density of parameters as compared to the traditional classical estimation methods which are intended for finding a single point estimate. Therefore Bayesian statistical models are thought to have advantages compared to the classical likelihood based inference methods and thus have been popular in recent traffic safety research.

Expected crash counts of three modes are formulated as follows:

$$\lambda_{i1} = exp(x_{i1}\beta_1 + \theta_{i1} + \varphi_i) = exp(x_{i1}\beta_1 + \delta_1 u_{i1} + \varphi_i) \quad (11)$$

$$\lambda_{i2} = exp(x_{i2}\beta_2 + \theta_{i2} + \varphi_i) = exp(x_{i2}\beta_2 + \delta_2 u_{i1} + \delta_3 u_{i2} + \varphi_i) \quad (12)$$

$$\lambda_{i3} = exp(x_{i3}\beta_3 + \theta_{i3} + \varphi_i) = exp(x_{i3}\beta_3 + \delta_4 u_{i1} + \delta_5 u_{i2} + \delta_6 u_{i3} + \varphi_i) \quad (13)$$

where, λ_{im} are expected crash frequencies of mode *m* in TAZ *i*

m=1 is motor vehicle crash, *m* =2 is bicycle crash, and *m* =3 is pedestrian crash in the multivariate model for crashes by modes.

x_{im} are row vectors of explanatory variables showing characteristics of TAZ *i*, for mode m,

β_m are coefficient estimates of model covariates x_{im} ,

θ_{im} are zone-mode specific random error terms representing normal heterogeneity of TAZ *i*, for mode *m*,

u_{im} are independent random variables, which follows normal distribution (0, τ_θ) for TAZ *i* and target mode m,

τ_θ is the precision parameter that is the inverse of the variance; it follows prior gamma (0.5, 0.005),

δ_k are coefficients for u_{im}, and

φ_i are spatial autocorrelation error terms.

The model was run considering a non-informative normal (0,10000) prior both for δ_k and β_m. In the case of the univariate model structure, δ_2, δ_4 and δ_5 are set to zero because the univariate model do not account for correlations between heterogeneities of crashes by different modes.

Spatial distribution was implemented by specifying intrinsic Gaussian Conditional Autoregressive (CAR) prior with normal ($\bar{\varphi}_i$, τ_i^2) distribution recommended by Besag (1974). Mean of φ_i is calculated as follows:

$$\bar{\varphi}_i = \left(\sum_{i \neq j} \varphi_j \times w_{ij}\right) \Big/ \left(\sum_{i \neq j} w_{ij}\right) \quad (14)$$

where, w_{ij} is the element of adjacency matrix with a value of 1 if i and j are adjacent or 0 otherwise.

Moreover, Deviance Information Criterion (DIC) was computed in each model for comparison. The following equation is used to calculate DIC (Spiegelhalter et al., 2002). Models with smaller DIC are preferred to models with larger DIC.

$$DIC = 2 \times \overline{D} - \widehat{D} \quad (15)$$

where, $\overline{D}$: posterior mean of deviance, D,

$\widehat{D} = 2 \times (p(y|\theta))$, and

$\bar{\theta}$: posterior mean of θ, respectively.

4.3 Results and Discussion

Overall four Bayesian Poisson-lognormal models for crashes by modes were developed: 1) multivariate model with the spatial error term (MVS); 2) multivariate model without the spatial error term (MV); 3) univariate model with the spatial error term (UVS); and 4) univariate model without the spatial error term (UV). Table 4-2 summarizes DIC of each model. It was found that the goodness-of-fit measure, DIC is the smallest in 1) multivariate model with the spatial error term (12895.3), and DICs of two models, 2) multivariate model without the spatial error terms and 3) univariate model with the spatial error term are comparable (13038.3 and 13027.3, respectively), but 3) univariate model with the spatial error term has a slightly better fit. 4) univariate model without the spatial error term showed the worst model performance (13506.3).

As stated in the methodology part, the error consists of the zone-mode specific random error ($\theta_{i\mathrm{m}}$) and spatial error (φ_i) in this chapter. The zone-mode specific error terms have components that account for unobserved factors across modes while the spatial error components account for spatial autocorrelation among zones. Table 4-3 shows error correlations with spatial error terms. It was found that there is considerably high correlation between crash frequencies by modes (0.881 to 0.994). However, it is also required to calculate error correlations without spatial error terms, in order to compare the effects from the zone-mode specific random error and spatial error. Table 4-4 summarizes error correlations without spatial error terms. The error correlation without spatial error terms (i.e., pure zone-mode specific random error) between motor vehicle and bicycle crashes (0.281) is quite smaller than that in the error correlation with spatial error terms (0.881); it shows that the large proportion of the common error shared by motor vehicle

and bicycle crashes is the spatial error. The zone-mode specific random error correlation between motor vehicle and pedestrian crashes (0.545), it is smaller than that in the error correlation with spatial error terms (0.927) but it still shows the relatively high correlation. To sum up, it is thought that both the zone-mode specific random error and spatial error have meaningful effects. When it comes to bicycle and pedestrian crashes, the error correlations with and without spatial error terms are 0.994 and 0.958, respectively. This shows the effect of the spatial error is limited in bicycle and pedestrian crashes but the effect of the zone-mode specific random error is comparatively large. The significant correlations between errors justify applying multivariate models at the macro-level, because it can account for common unobserved factors across crash types.

The result of MVS, which has a best model fit, is shown in Table 4-5. There are several common significant factors for three target variables, such as 'Log of vehicle-miles-traveled', 'Log of population density', 'Log of employments and school enrollments', 'Number of intersections' and 'Number of traffic signals'. Nevertheless, age related factors (i.e., 'Proportion of young people' and 'Proportion of elderly people') and 'Roadway length with poor pavement condition' are not significant for all response variables. On the other hand, 'Log of vehicle-miles-traveled', 'Number of intersections' and 'Number of traffic signals' ,which represent the traffic volume and complexity of the traffic network in each zone, have positive signs for motor vehicle, bicycle and pedestrian crashes as expected. Thus, traffic crashes occur more frequently as the traffic volume and network complexity increase, regardless of the transportation mode. Demographic variables other than 'Log of population density' are not significant for motor vehicle and bicycle crashes.

Nevertheless, race/ethnicity related variables (i.e., 'Proportion of African Americans' and 'Proportion of Hispanics') are significant and they have positive association with pedestrian crashes.

'Proportion of household without vehicles' is significant for non-motorized mode crashes and it has positive sign for bicycle and pedestrian crashes. People from households with no vehicles are more likely to use public transportation (i.e., bus) or non-motorized modes (i.e., walking, bicycle, etc.). Thus, as households without motor vehicles increase, that zone has more probability to have larger pedestrian and/or bicycle crashes. 'Log of rooms of hotel, motel and time shares' is found statistically significant and it has a positive relationship with motor vehicle and pedestrian crash counts. Since this variable stands for the activity of tourism industry in a zone, it is interpreted that zones with many tourist attractions, which is typical to Central Florida, have more numbers of motor vehicle and pedestrian crashes while 'Log of rooms of hotel, motel and time shares' is not significant for bicycle crashes.

The speed limit of the roadway is also an important factor for the crash occurrence. 'Proportion of roadway with speed limit less than or equal to 20 mph', which stands for the proportion of low-speed roads such as residential roads in a zone, was found negatively associated with bicycle crashes. It implies that bicycle crashes do not frequently occur in zones with many low-speed roadways. In contrast, 'Proportion of roadway with speed limit more than or equal to 55 mph', which represents the proportion of high-speed roadways of a zone, was significant for both motor vehicle and pedestrian crashes but their coefficients have different signs in motor and

pedestrian crashes. 'Proportion of roadway with speed limit more than or equal to 55 mph' was positively associated with motor vehicle crashes; thus zones with more high-speed roads are more likely to have more motor vehicle crashes. Meanwhile, it has a negative relationship with pedestrian crashes, which implies that pedestrian crashes occur frequently in zones with more lower-speed roads.

Furthermore, Figure 4-1 depicts the spatial distribution of crashes by modes. It is commonly observed that all types of crashes are concentrated in the urban areas. However, motor vehicle and pedestrian crashes hotspots are spatially dispersed to suburban peripheries compared to bicycle crashes. On the contrary, most of bicycle crashes hotspots are located in urban areas.

Table 4-2: Summary of the model performance

Models	$\overline{D}$	$\widehat{D}$	DIC
MVS (multivariate model with the spatial error term)	11785.5	10675.6	12895.3
MV (multivariate model without the spatial error term)	11773.9	10509.5	13038.3
UVS (univariate model with the spatial error term)	11793.0	10558.7	13027.3
UV (univariate model without the spatial error term)	11792.0	10077.7	13506.3

Table 4-3: Symmetric matrix of error correlations with spatial error terms ($\boldsymbol{\theta_{im} + \varphi_i}$) of crash models by modes

Crash types	Motor vehicle crash	Bicycle crash	Pedestrian crash
Motor vehicle crash	1	-	-
Bicycle crash	0.881 (p<0.001)	1	-
Pedestrian crash	0.927 (p<0.001)	0.994 (p<0.001)	1

Table 4-4: Symmetric matrix of error correlations without spatial error terms ($\boldsymbol{\theta_{im}}$) of crash model by modes

Crash types	Motor vehicle crash	Bicycle crash	Pedestrian crash
Motor vehicle crash	1	-	-
Bicycle crash	0.281 (p<0.001)	1	-
Pedestrian crash	0.545 (p<0.001)	0.958 (p<0.001)	1

Table 4-5: Multivariate model accounting for the spatial autocorrelation crashes by modes

Variable	Motor vehicle crash				Bicycle crash				Pedestrian crash			
	Mean	Stdev	BCI		Mean	Stdev	BCI		Mean	Stdev	BCI	
			2.5%	97.5%			2.5%	97.5%			2.5%	97.5%
constant	0.366	0.397	-0.112	1.590	-3.899	0.517	-4.824	-2.843	-4.102	0.490	-4.924	-3.059
Log of vehicle-miles-traveled	0.134	0.015	0.105	0.156	0.096	0.031	0.037	0.157	0.132	0.029	0.077	0.188
Log of population density	0.040	0.011	0.016	0.059	0.123	0.021	0.081	0.164	0.046	0.019	0.008	0.082
Proportion of African Americans	-0.008	0.156	-0.314	0.297	-0.317	0.254	-0.825	0.175	0.627	0.231	0.173	1.080
Proportion of Hispanics	0.209	0.195	-0.174	0.576	-0.432	0.309	-1.040	0.175	0.994	0.281	0.439	1.541
Proportion of young people (15-24 years old)	0.065	0.373	-0.696	0.795	0.225	0.523	-0.821	1.237	-0.025	0.529	-1.079	1.005
Proportion of elderly people (65 years old or older)	-0.106	0.293	-0.768	0.415	-0.162	0.535	-1.246	0.847	-0.685	0.543	-1.769	0.356
Proportion of household without vehicles	0.443	0.366	-0.280	1.134	1.519	0.593	0.357	2.683	2.133	0.524	1.115	3.159
Log of rooms of hotel, motel and time shares	0.034	0.011	0.013	0.054	0.022	0.019	-0.015	0.059	0.035	0.017	0.002	0.068
Log of employments and school enrollments	0.171	0.032	0.069	0.213	0.174	0.046	0.072	0.262	0.164	0.043	0.073	0.242
Number of intersections	0.178	0.019	0.140	0.216	0.168	0.032	0.106	0.231	0.152	0.030	0.094	0.210
Number of traffic signals	0.023	0.002	0.019	0.028	0.014	0.004	0.006	0.022	0.021	0.004	0.014	0.029
Proportion of roadway with speed limit less than or equal to 20 mph	0.378	0.333	-0.297	1.000	-2.395	0.915	-4.215	-0.640	-1.290	0.737	-2.755	0.124
Proportion of roadway with speed limit more than or equal to 55 mph	0.598	0.200	0.203	0.991	-0.682	0.493	-1.654	0.280	-1.554	0.450	-2.443	-0.684
Roadway length with poor pavement condition	0.052	0.059	-0.064	0.169	0.035	0.099	-0.164	0.227	0.030	0.095	-0.158	0.216
$\delta_1, \delta_2, \delta_4$	6.702	2.976	2.252	14.120	2.083	2.502	-2.139	6.685	2.823	2.273	-0.673	7.732
δ_3, δ_5	-	-	-	-	8.170	3.418	2.668	15.680	6.367	2.992	1.942	12.79
δ_6	-	-	-	-	-	-	-	-	0.138	5.072	-9.637	9.095
DIC	12895.3											

Non-significant variables in shaded cells, all other variables are significant at 5%.

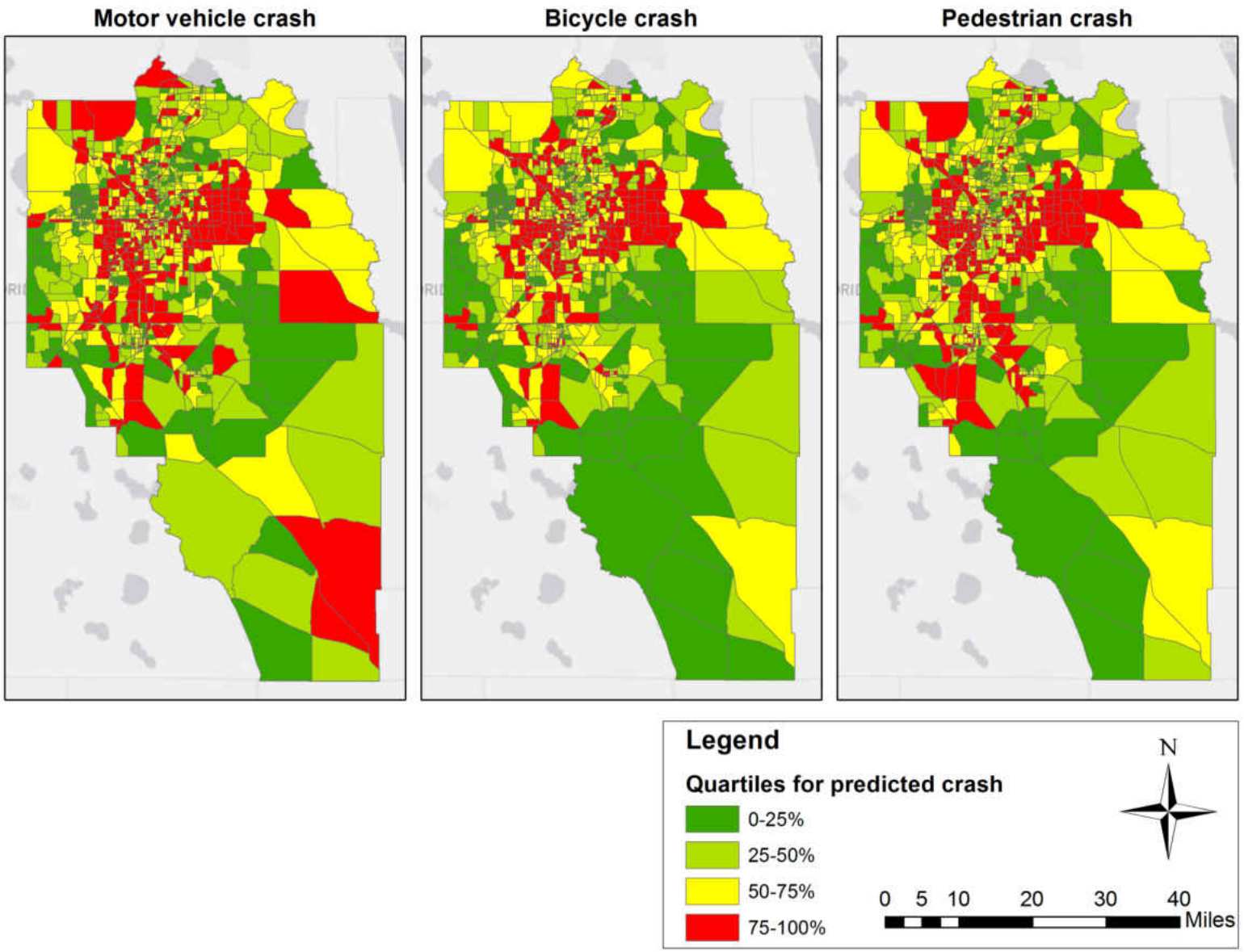


Figure 4-1: Spatial distribution of predicted crashes by modes

4.4 Summary and Conclusions

Generally, macroscopic models developed for transportation safety planning (TSP) deal with multiple crash types by transportation modes, severity levels, et cetera. Many independent univariate models have been developed for these multiple crash types for the macro-level studies, but separate models for various crash types ignore shared unobserved factors across types, which can lead to inefficiency and biases in the model estimation. In order to resolve this issue, MVS, which can account for error correlations among crashes by modes and spatial autocorrelations, were estimated using TAZ-based crash data, and MVS was compared with the MV, UVS and UV. The MVS showed the best model performance, in terms of DIC. Also high correlations in errors between crashes by modes justify the adoption of multivariate model. It is shown that the spatial error component plays a key role in significantly improving the model performance.

Variables found significant in the MVS model are reasonable and explainable. Several traffic/roadway variables represent the traffic volume and complexity of the traffic network commonly significant for all crash response variables and they have positive signs as expected. 'Log of population density', which could be related to trip generation, is significant for all target variables as well. Meanwhile, 'Log of employment and school enrollment', which may represent trip attraction, is found significant exclusively for bicycle and pedestrian crashes. Demographic variables except for 'Log of population density' are not significant for motor vehicle and bicycle crashes but the result showed that zones with more African American and Hispanics are more vulnerable to bicycle and pedestrian crashes. Furthermore, zones with many tourist attractions are more likely to have motor vehicle and pedestrian crashes. It may imply that tourists who are

not familiar with local roadways and rules are exposed more to traffic crashes. It was also revealed that zones with more high-speed roads have higher probabilities to have more number of motor vehicle crashes; but they are likely to have less number of pedestrian crashes. Meanwhile, zones with more low-speed roads are less likely to have many bicycle crashes.

To sum up, four key findings from the results are as follows: First, MVS performs much better than MV, UVS and UV, in terms of DIC. Second, there are significant correlations between zone-mode specific random errors of crashes by each type. This correlation is very strong between bicycle and pedestrian crashes, but it is relatively weak between motor vehicle and bicycle crashes. Third, accounting for spatial autocorrelations is essential to improve the model performance, regardless of univariate or multivariate models. It was also found that the performance of the MV is very comparable to that of the UVS but UVS performs slightly better than MV. Lastly, significant variable sets for crashes are different by transportation modes.

In the aspect of traffic safety, it was shown that there are many factors commonly significant across crashes by transportation modes as stated previously. However, it is also possible that there are additional shared factors that are non-observed or were omitted in the modeling process. Unsurprisingly, the error correlation of bicycle and pedestrian crashes is very high, even without the spatial correlation. It may imply the existence of unobserved shared factors across these two different modes. This relationship may be caused by common inherent characteristics between bicycle and pedestrian crashes. It is interesting that the error correlation between motor vehicle and pedestrian crashes is also found somewhat high; which indicate that there are possible

common omitted factors between these two types of crashes. Thus, it is recommended to consider the unobserved factors shared by different crashes types, while choosing additional variables in future studies.

It is expected that findings from this chapter can contribute to more reliable traffic crash modeling especially when focusing on crashes by different transportation modes in the context of TSP. Also, variables that are found significant for each mode can be used to guide traffic safety policy decision makers to allocate resources more efficiently.

CHAPTER 5 EFFECTS OF GEOGRAPHIC UNITS ON MACROSCOPIC SAFETY MODELING

5.1 Introduction

As shown in the literature review, various geographic units have been explored in macro-level modeling. With the advancement of GIS (Geographic Information System) analysts are able to analyze crashes for various geographical units. However, a clear guideline on which geographic entity should be chosen is not present. Macro level safety analysis is at the core of TSP which in turn is a key in many aspects of policy and decision making of safety investments.

The preference of spatial unit can vary with the dependent variable of the model. Or, for a specific dependent variable, models may be invariant to multiple spatial units by producing a similar goodness-of-fits. In this chapter three different crash models were investigated for TAZs, BGs and CTs of two counties in Florida. The models were developed for the total crashes, severe crashes and pedestrian crashes in this region. The primary objective of this chapter is to explore and investigate the effect of zonal variation on these specific types of crash models.

Based on the aforementioned discussion it can be summarized that a CT is a greater geographic entity than a BG and most likely than a TAZ. However, the size of a block and a TAZ may be comparable. The configuration of the three zones is different by locations. For the study counties, Table 5-1 compares the median area and the number of zones by the urban and rural area. Median areas of TAZs and BGs are 0.370 and 0.366, respectively, which show their sizes are quite comparable (although quite different for the specific urban and rural areas). On the other

hand, the median CT area is 1.508, which is considerably larger than those of TAZs or BGs. It is interesting to note that the BG median size is slightly smaller than TAZ median size in the urban area. Moreover, 82.1% of BGs are located in the urban area whereas 74.8% of TAZs are placed in the urban area. It indicates that BGs are more detailed in the urban area compared to TAZs. This can be also shown in Figure 5-1, which compares urban areas among the three zone systems. The term, urban area used here, is defined as the area with at least 1,000 people per square mile (U.S. Census Bureau, 1994).

Table 5-1: Median area and number of zones of each geographic unit

Location	Median Area (mi^2)			Number of Zones		
	CT	BG	TAZ	CT	BG	TAZ
Urban	1.180	0.288	0.332	357 (78.1%)	1099 (82.1%)	1106 (74.8%)
Rural	8.175	2.940	0.676	100 (21.9%)	239 (17.9%)	373 (25.2%)
Total	1.508	0.366	0.370	457 (100%)	1338 (100%)	1479 (100%)

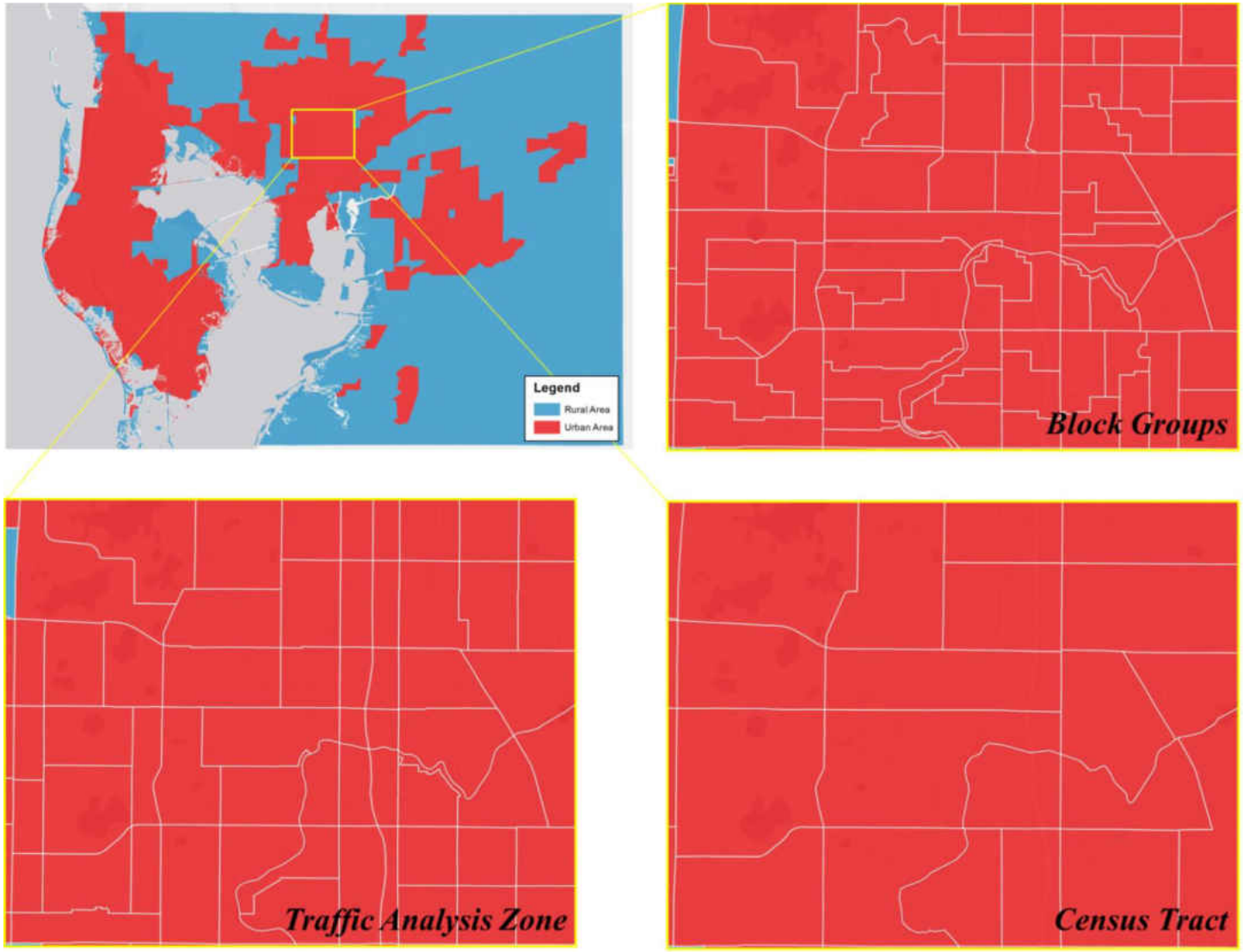


Figure 5-1: Comparison of areas of TAZs, BGs, and CTs in the urbanized area

5.2 Data Preparation

Crashes which occurred in Hillsborough and Pinellas counties in the period of 2005-2006 were analyzed for the study. These two counties are located midway along the west coast and two of the most populous counties of Florida. The geographic region of these two counties are divided into 1479 traffic analysis zones, 1338 block groups and 457 census tracts. A total of 87,718 crashes occurred in these two counties among which 7106 (8.1%) were severe crashes and 1665 (1.9%) pedestrian crashes. Summary statistics of the collected data are presented in Table 5-2. In this chapter, severe crashes were defined as the combined sum of all fatal and incapacitating injury crashes.

GIS shape files (maps) of block groups and census tracts were downloaded from the U.S. Census Bureau website. Shape file of TAZ boundary was collected from Florida DOT (FDOT) District 7 (comprising Hillsborough, Pinellas, Pasco, Citrus and Hernando County) Intermodal Systems Development Unit. Census 2000 data was downloaded from U.S. Census Bureau website using PLANSAFE Census Tool (Washington et al., 2010). The same tool was used to aggregate census information for each geographic entity (BG/CT/TAZ). A total of 75 variables were available for each polygon. A complete list of these variables can be found in pages 107-109 of PLANSAFE manual (Washington et al., 2010). In addition to these variables seven roadway related variables were created for each geographic entity. These variables are- total number of intersections, total roadway length with 15, 25, 35, 45, 55, and 65mph posted speed limit (PSL). The GIS shapefile for roadways provided PSLs for the specific roadway segments.

Many of the census-related variables were thought to be correlated with population and employment. Therefore, collinearities among the variables were specifically investigated before calibrating the statistical models for total, severe and pedestrian crashes. The final estimation of the models ensured the non-inclusion of correlated variables.

Figure 5-2 presents the geographical distribution of total crashes in the three different zonal systems. The five color scales show the zones from the smallest number of crashes (bottom 0-20%) to the largest number of crashes (top 80-100%). In the CT based crash distribution map, it can be roughly observed several zones with many crashes are located in the middle of the map. It is also shown in the BG based map, but it can be identified which specific zones have more crashes more clear.

Moreover, Different from the two previous geographical units, TAZ is delineated based on physical boundaries such as the sea or the river (Baass, 1981). As shown in the TAZ, since the zone size is smaller, more specific hotspots can be identified; however, also due to the small size of the TAZ, the crash distribution in TAZs could be too dispersed to identify large hotspot areas as shown in Figure 5-2. It implies that the identification of crash hotspots can be different by basic geographical units and the aggregation level and the configuration of the zone system also could affect the result of crash modeling at the macroscopic level.

Table 5-2: Summary statistics by geographic entities

Variables	Geographical Units											
	Census Tracts (N=457)				Block Groups (N=1338)				Traffic Analysis Zones (N=1479)			
	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
Total crashes	191.94	166.51	4	1149	65.56	83	0	716	59.31	62.80	0	481
Severe crashes	15.55	14.13	0	74	5.31	7	0	64	4.80	5.87	0	47
Pedestrian crashes	3.64	4.02	0	34	1.24	2	0	16	1.13	1.74	0	24
VMT	2114355	202630.6	0.00	1221197.2	79778.6	115442.9	0.00	968619.4	71519.8	92765.6	0.00	788771.6
Roadway length with 25mph PSL	19.68	10.95	0.92	81.17	6.72	6.10	0.06	57.39	6.38	5.79	0.00	48.12
Roadway length with 35mph PSL	3.67	3.01	0.00	20.46	1.25	1.62	0.00	17.99	1.28	1.33	0.00	15.02
Roadway length with 45mph PSL	0.27	0.73	0.00	7.65	0.09	0.39	0.00	6.84	0.10	0.34	0.00	4.07
Roadway length with 55mph PSL	0.23	0.99	0.00	9.14	0.08	0.49	0.00	7.46	0.08	0.49	0.00	12.37
Roadway length with 65mph PSL	0.57	1.43	0.00	10.25	0.20	0.75	0.00	10.10	0.20	0.64	0.00	9.79
Total number of intersection	38.63	28.93	0.00	233.00	13.20	12.95	0.00	139.00	11.92	10.66	0.00	83.00
Population age 0 to 15	840.33	469.64	0.00	2815.00	287.02	243.27	0.00	2495.00	248.86	264.84	0.00	2672.00
Population age 16 to 64	2645.87	1261.66	127.00	7702.00	903.71	719.65	0.00	6752.00	765.34	711.05	0.00	6398.00
Density of children (K to 12^{th} grade)	0.88	0.66	0.00	5.46	1.06	1.00	0.00	12.55	0.07	0.06	0.00	0.46
Proportion minority population	0.21	0.22	0.00	0.99	0.22	0.25	0.00	1.00	0.21	0.23	0.00	1.00
No of workers: travel time 0-4 min	44.03	37.13	0.00	249.00	15.04	20.13	0.00	249.00	12.46	15.01	0.00	142.00
No of workers: travel time 5-9 min	178.12	115.13	7.00	851.00	60.84	57.60	0.00	571.00	50.97	50.96	0.00	323.00
No of workers: travel time 15-19 min	317.56	186.45	0.00	1155.00	108.46	99.51	0.00	1155.00	91.98	89.38	0.00	599.00
No of workers: travel time ≥ 30 min	648.82	402.67	0.00	2602.00	221.61	224.97	0.00	2383.00	187.80	221.24	0.00	2510.00
No of workers: worked at home	61.87	52.33	0.00	311.00	21.13	27.42	0.00	311.00	16.81	22.89	0.00	236.00
No of workers: commute by public transport	1784.50	932.47	35.00	5719.00	609.50	529.72	0.00	5173.00	515.47	510.36	0.00	4606.00
No of workers: commute by walking	31.70	50.37	0.00	630.00	10.83	24.20	0.00	598.00	9.32	21.11	0.00	568.00
Housing units per acre	35.22	39.07	0.00	246.00	12.03	19.16	0.00	202.00	9.98	14.49	0.00	149.00
Median household income	2.69	1.99	0.01	19.95	3.02	2.68	0.00	42.77	0.21	0.17	0.00	2.21

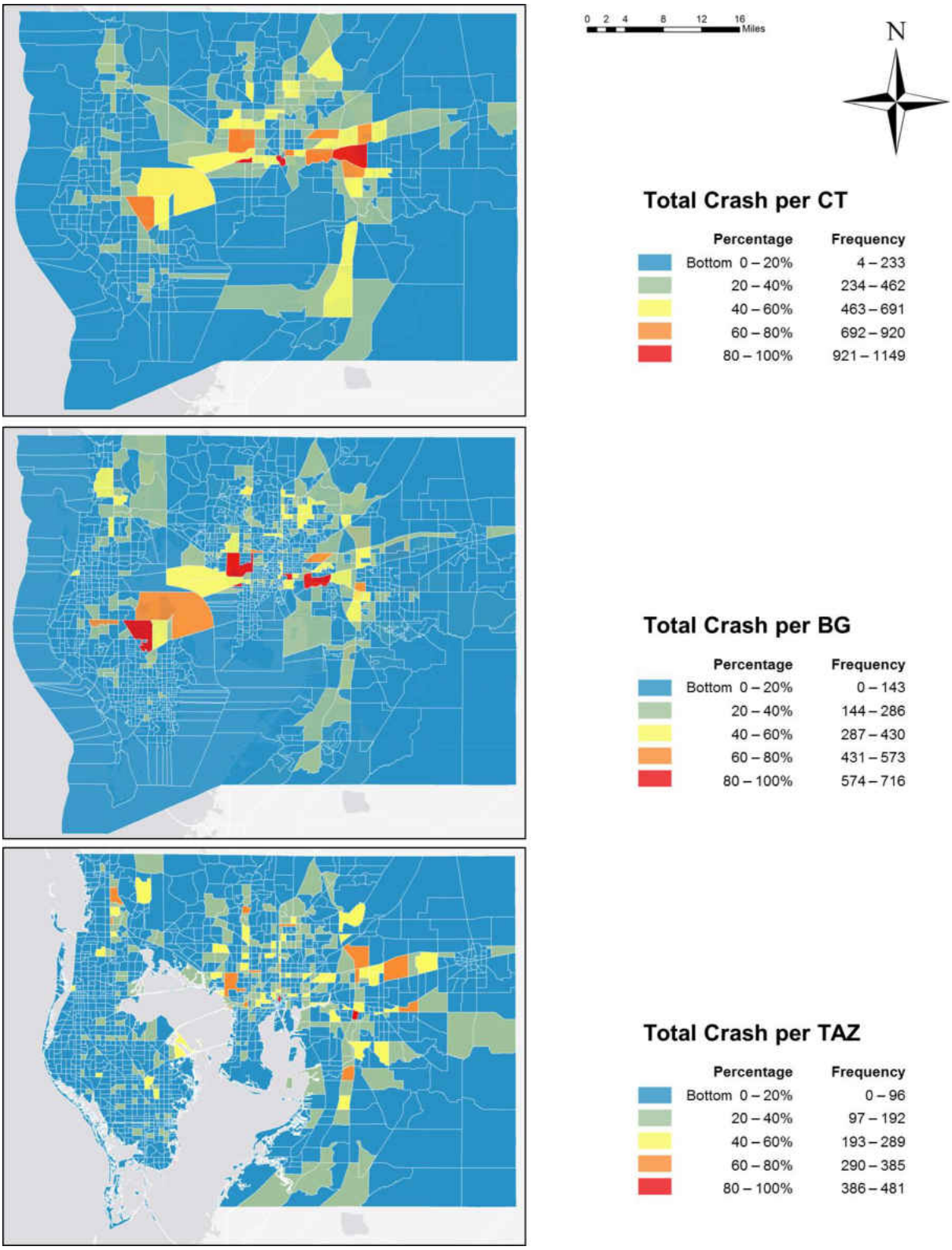


Figure 5-2: Geographical distribution of total crashes per CT, BGs and TAZs

5.3 Statistical Method

Models were developed for three response variables- i) total crashes, ii) severe crashes, and iii) pedestrian crashes for different geographic entities. Any predictors which had no problem of multi-collinearity and were significant at least at the 80% confidence level were retained in the final models. Multi-collinearity between variables was investigated using Pearson product-moment correlation, Spearman's rank order correlation, and bivariate posterior scatter plots derived from Bayesian analysis. As an example, variables 'population aged between 0 and 15', and 'density of children in K to 12^{th} grade' intuitively seemed to have some kind of collinearity.

It was found that for TAZs, the Pearson product-moment correlation and Spearman's rank order correlation between these variables were 0.547 and 0.629, respectively; the same for BGs were 0.286 and 0.321, respectively. None of these numbers indicated any obvious correlation. Similarly correlation between 'median household income' and 'proportion of minority population' was investigated and no correlation was identified.

Poisson lognormal models have been suggested as a substitute for the negative binomial model/Poisson gamma model for the crash frequency data (Park and Lord, 2007; Aguero-Valverde and Jovanis, 2008, 2009; Ma et al., 2008; El-Basyouny and Sayed, 2009, 2010). The Poisson lognormal model is comparable with the negative binomial/Poisson gamma model; however, the Poisson lognormal model provides more flexibility compared to negative binomial model/Poisson gamma model (Lord and Mannering, 2010).

Unlike classical regression methods, Bayesian models do not usually depend on the assumption of asymptotic normality. Sampling based methods of Bayesian estimation focus on estimating the entire density of a parameter as compared to the traditional classical estimation methods which are designed to find a single optimum estimate. Therefore, Bayesian statistical models are thought to have several advantages compared to the 'classical' likelihood based inference methods and have been popular in the recent safety research.

A Poisson-lognormal model can be specified as follows:

$$y[i] \sim \text{Poisson}(\mu[i]) \quad (16)$$

$$\log(\mu[i]) = \beta_0 + \beta X_i + \theta[i] \quad (17)$$

$$\theta[i] \sim \text{Normal}(0, \tau_\theta) \quad (18)$$

where,

β_0 = intercept term,

β's are the coefficient estimates of the model covariates (X_i),

$\theta[i]$ = error component of the model, and

τ_θ = precision parameter which is inverse of the variance; τ_θ follows a prior gamma (0.5, 0.005)

This variance ($1/\tau_\theta$) provides the amount of variation not explained by the Poisson assumption (Lawson et al., 2003). A uniform prior distribution was assumed for β_0. The model was run considering a non-informative Normal (0, 100000) prior for β's.

The model convergence and performance were decided based on Brooks-Gelman-Rubin statistics, overlaps among the Markov chains, autocorrelation plots, and density plots. The significance of the model parameters were judged based on Bayesian credible intervals (BCIs). BCI infers on the true parameter value; for example a 95% BCI will contain the true parameter value with approximately 95% certainty.

Moreover, Deviance Information Criterion (DIC) was calculated in each model. The following equation is used to compute DIC (Spiegelhalter et al., 2002).

$$DIC = 2 * \overline{D} - \widehat{D} \quad (19)$$

where,

$\overline{D}$ = posterior mean of deviance, *D,*

$\widehat{D}$= -2*log $\left(p\left(y|\bar{\theta}\right)\right)$, and

$\bar{\theta}$ = posterior mean of θ, respectively.

DIC is not appropriate to compare models based on different areal units because DIC increases with the deviance which also increases with the sample size. Thus, DIC was used to choose the best variable sets for each model. DIC of each model was shown in Table 5-3, 5-4, and 3-5. As expected, TAZ based models always have the largest DIC (10989.50, 6470.64 and 3849.09 in total, severe and pedestrian crash models, respectively) because TAZ has the biggest number of zones (N=1479). While CT based models have the smallest DIC (4024.68, 2747.92 and 1847.26 in total, severe and pedestrian crash models, respectively), since CT has the smallest number of zones (N=457).

5.4 Results

Significant factors associated with total, severe and pedestrian crashes at 80% Bayesian credible interval were analyzed. Three types of crashes were modeled based on three different geographical units (CT/BG/TAZ), therefore nine models (3 response variables × 3 geographic entities) were developed in total.

Explanatory variables used in the models can be divided into three categories. First, demographic or socioeconomic variables include population by age groups, density of children, minority population, housing unit density, and median household income. Second category is the roadway/traffic related variables. For instance, they are the VMT, the roadway length by each posted speed limit and the number of intersections. Lastly, the third category is the commute characteristics such as workers by commute times and workers by transportation modes. The number of home workers was also considered a special case of commute characteristic variables since they are also workers but do not need to commute. Table 5-3, 5-4, and 5-5 display results from models based on the three geographical units for total, severe and pedestrian crashes, respectively. It was revealed that all 7 roadway/traffic related variables are significant in the TAZ based total and severe models. Also for the pedestrian models, TAZ based models have the largest number of roadway/traffic related significant factors. On the other hand, BG based models have more significant commute related variables compared to models based on other units. Meanwhile, the number of significant demographic/socioeconomic factors is similar in all models.

Table 5-3 summarizes results of three total crash models based on CT/BG/TAZ. These three models have several common significant variables such as the VMT, the roadway length with a speed limit of 35, 55 and 65mph, the number of intersections, the proportion of the minority population, commuters by walking, and the median household income. Among these variables, the VMT and the number of intersections are significant for all nine models. On the other hand, population from 16 to 64 years old and all variables regarding workers by commute times (except for travel time 15-19 min in BG based model) were found not significant for all total crash models, regardless of areal units. Moreover, the population of the youngest age group from 0 to 15 years old, and roadway length with the high speed limit (65mph) were significant solely in the CT based model. Moreover, it is interesting that both workers commuting by public transportation and the density of housing units were negatively associated in the CT/BG based models; however, they were not significant in the TAZ based model.

Modeling results of severe crashes are presented in Table 5-4. Similar to the total crash models, the VMT, the roadway length with 35, 55 and 65mph, the number of intersections and median household income, were found also common significant variables in the severe crash models. It is interesting that the variable, workers with shorter commute time (5-9 min), was also significant and negatively associated with severe crashes in all severe crash models. While the workers longer commute time (30 min or over) variable was positively associated with severe crashes in both the BG and TAZ based models. This result shows that the workers with the longer commute time are more vulnerable to severe crashes compared to workers with shorter commute time. Moreover, the population aged from 0 to 15, the proportion of minority

population and workers commuting by public transportation were not significant in all severe crash models, regardless of areal units. In contrast, the population aged from 16 to 64 is significant in both BG and TAZ based models. It was found that the population from 16 to 64 has a positive relationship with severe crashes. Furthermore, the number of homeworkers is negatively related with severe crashes in BG and TAZ based models. This variable is also significant in the BG based total crashes models and pedestrian models based on BG/TAZ. Negative sign of this coefficient can be explained easily because homeworkers are not exposed to the traffic compared to other workers. However, it is interesting to note that the significance of this variable cannot be found in the larger zone system such as CT. In addition, the housing units per acre are negatively associated with severe crashes in BG/TAZ based models, which is consistent with the total crash models.

Finally, Table 5-5 shows the results of the pedestrian crash modeling. As mentioned previously, the VMT and the number of intersections are also significant factors in pedestrian models. Additionally, the number of workers commuting by public transportation, the workers commuting by walking and the proportion of minority population are significant for all pedestrian models based on the three geographical units. These additional significant variables are well explainable. First, public transportation commuters need to access the public transportation facilities (e.g., bus stop) by walking. Second, it is very clear that walking commuters are more vulnerable to pedestrian crashes because they are more exposed. Lastly, because the minority people have lower vehicle ownership in general (Dawkins et al., 2005), so they have to commute by public transportation or by walking. Thus, minority people have higher

probability to be involved in pedestrian crashes. Regarding the speed related variables, roadway length with relatively low speed limits are positively associated with the pedestrian crashes in BG/TAZ based models. The roadway length with 35mph was found significant both in BG/TAZ based pedestrian models whereas the length of 25 mph speed limit roadway was significant exclusively in the TAZ based pedestrian model. It is noteworthy that 65 mph is significant and negatively associated with pedestrian crashes solely in the CT based model; however, this variable has a positive relationship with the total severe crashes for all three areal units. Concerning the population by age groups, population of the children aged from 0 to 15 are negatively related with pedestrian crashes in the BG/TAZ based models whereas the density of children (K to 12^{th} grade) is positively associated with pedestrian crashes only in the TAZ based model. Variables related to workers by commute time are not significant in the pedestrian models except for the variable of workers with commute time 15 to 19 minutes, significant only in the TAZ based model. Furthermore, the number of home workers has a negative relationship with pedestrian crashes for the BG/TAZ based models.

Key findings in the comparison of models based on different areal units are as follow. First, signs of coefficients are consistent if these variables are significant in models with same response variables, even if the geographical units are different. For instance, signs of the roadway length with the 65mph posted speed limit are positive in total crash models for CT/BG/TAZ. Second, the number of significant variables varies by response variables and also by geographic units. BG based models have the largest number of significant variables for total and severe crashes. On the contrary, CT basis models have the smallest number of significant variables. In addition,

TAZ based pedestrian model has more significant factors compared to the other models based on CT/BG. It was also found that the largest number of roadway/traffic related variables is significant in the TAZ based models for all three response variables. This result seems reasonable because TAZs are created based on traffic-related factors by their definitions.

Table 5-3: Total crash Bayesian Poisson-lognormal models by geographic entities

Variables	Geographical Units											
	Census Tracts (N=457)				Block Groups (N=1338)				Traffic Analysis Zones (N=1479)			
	Mean	Std. Dev.	Bayesian Credible Interval		Mean	Std. Dev.	Bayesian Credible Interval		Mean	Std. Dev.	Bayesian Credible Interval	
			2.50%	97.50%			2.50%	97.50%			2.50%	97.50%
Response variable: Total crashes	191.939	166.512			65.558	82.847			59.309	62.802		
Intercept	5.1300	0.1104	4.8680	5.3160	2.1320	0.4581	1.5080	2.9030	1.7580	0.5655	1.1090	3.2150
Log of VMT	0.1073	0.0512	0.0243	0.2050	0.1037	0.01135	0.0743	0.1211	0.1138	0.0075	0.0985	0.1246
Log of roadway length with 25mph PSL									0.1117	0.0408	0.0422	0.1862
Log of roadway length with 35mph PSL	0.3480	0.0862	0.2107	0.5124	0.5607	0.05005	0.4689	0.6571	0.6155	0.0494	0.5240	0.7176
Log of roadway length with 45mph PSL					0.3099	0.1395	0.0490	0.5827	0.5062	0.1381	0.2682	0.7752
Log of roadway length with 55mph PSL	0.1685	0.0893	0.0199	0.3522	0.1843	0.09937	0.0069	0.3916	0.2661	0.1117	0.0376	0.4781
Log of roadway length with 65mph PSL	0.3609	0.0853	0.2300	0.5440	0.4532	0.09178	0.2737	0.6394	0.3664	0.0997	0.1879	0.5693
Log of total number of intersection	0.2106	0.0925	0.0248	0.3230	0.2976	0.02509	0.2412	0.3382	0.2434	0.0300	0.1524	0.2919
Log of population age 0 to 15	-0.1948	0.0574	-0.2851	-0.1023								
Log of population age 16 to 64												
Density of children (K to 12th grade)	0.3344	0.0698	0.1920	0.4592					2.4160	0.3609	1.7020	3.1420
Proportion minority population					0.5527	0.1277	0.2560	0.7717	0.5878	0.1545	0.1157	0.8364
Log of no of workers: travel time 0-4 min												
Log of no of workers: travel time 5-9 min												
Log of no of workers: travel time 15-19 min					**0.0817	0.0509	-0.0942	0.1346				
Log of no of workers: travel time ≥ 30 min												
Log of no of workers: worked at home					**-0.0282	0.0172	-0.0622	0.0058				
Log of no of workers: commute by public transport	0.0658	0.0324	0.0085	0.1218	0.0387	0.0165	0.0064	0.0701				
Log of no of workers: commute by walking	0.0596	0.0274	0.0177	0.1241	0.0669	0.0150	0.0370	0.0963	0.05586	0.0196	0.0209	0.0961
Housing units per acre	-0.0956	0.0217	-0.1362	-0.0525	-0.0233	0.0108	-0.0466	-0.0039				
Log of median household income	-0.1781	0.0863	-0.3231	-0.0355	-0.1124	0.0343	-0.1638	-0.0666	-0.0894	0.04920	-0.1903	-0.0301
Stdev of θ[i]	0.5592	0.0436	0.4958	0.6568	0.6868	0.0301	0.6578	0.7230	0.7177	0.0338	0.6873	0.7569
DIC	4024.68				9908.52				10989.50			

* significant at 10%, ** significant at 20%, shaded area means explanatory variables were not significant in the models
All other explanatory variables are significantly different from zero at 95% Bayesian credible interval

Table 5-4: Severe crash Bayesian Poisson-lognormal models by geographic entities

Variables	Geographical Units											
	Census Tracts (N=457)				Block Groups (N=1338)				Traffic Analysis Zones (N=1479)			
	Mean	Std. Dev.	Bayesian Credible Interval		Mean	Std. Dev.	Bayesian Credible Interval		Mean	Std. Dev.	Bayesian Credible Interval	
			2.50%	97.50%			2.50%	97.50%			2.50%	97.50%
Response variable: Severe crashes	15.549	14.128			5.311	7.041			4.802	5.872		
Intercept	0.1363	1.0190	-1.3580	2.4510	-1.4800	0.3675	-2.0760	-0.8622	-1.6470	0.4055	-2.3470	-1.0470
Log of VMT	0.3017	0.0726	0.1120	0.4058	0.1464	0.0166	0.1200	0.1806	0.1415	0.0190	0.1114	0.1781
Log of roadway length with 25mph PSL									0.1173	0.0507	0.0251	0.2207
Log of roadway length with 35mph PSL	0.3994	0.0834	0.2278	0.5572	0.5776	0.0647	0.4426	0.6953	0.7149	0.0698	0.5724	0.8457
Log of roadway length with 45mph PSL					0.3283	0.1555	0.0252	0.6345	0.5423	0.1332	0.2729	0.7980
Log of roadway length with 55mph PSL	0.2150	0.0844	0.0561	0.3827	0.3196	0.1077	0.1187	0.5426	0.4294	0.1207	0.1869	0.6647
Log of roadway length with 65mph PSL	0.2152	0.0760	0.0795	0.3751	0.3970	0.1009	0.2060	0.5986	0.3771	0.0899	0.2044	0.5539
Log of total number of intersection	0.1412	0.0382	0.0621	0.2142	0.1792	0.0304	0.1190	0.2397	0.1265	0.0295	0.0697	0.1823
Log of population age 0 to 15												
Log of population age 16 to 64					0.2147	0.0724	0.0400	0.3185	0.1122	0.0628	0.0060	0.2287
Density of children (K to 12th grade)	0.2172	0.0629	0.0964	0.3438					2.4280	0.5005	1.4680	3.4000
Proportion minority population												
Log of no of workers: travel time 0-4 min					*-0.0367	0.0220	-0.0820	0.0067				
Log of no of workers: travel time 5-9 min	**-0.0937	0.0618	-0.1987	0.0608	-0.0961	0.0336	-0.1601	-0.0313	-0.0722	0.0362	-0.1512	-0.0082
Log of no of workers: travel time 15-19 min	*0.1026	0.0548	-0.0019	0.1993								
Log of no of workers: travel time ≥ 30 min					0.1045	0.0496	0.0111	0.2124	*0.1005	0.0575	-0.0038	0.1973
Log of no of workers: worked at home					-0.0749	0.0210	-0.1165	-0.0330	-0.1075	0.0245	-0.1563	-0.0599
Log of no of workers: commute by public transport												
Log of no of workers: commute by walking					0.0740	0.0210	0.0311	0.1152				
Housing units per acre	-0.0548	0.0212	-0.0961	-0.0114	-0.0361	0.0141	-0.0630	-0.0087				
Log of median household income	-0.2425	0.0427	-0.3265	-0.1485	-0.1077	0.0227	-0.1521	-0.0679	-0.0615	0.0298	-0.1225	-0.0077
Stdev of θ[i]	0.5434	0.0420	0.4930	0.6066	0.6984	0.0820	0.6535	0.7587	0.6703	0.0865	0.6273	0.7268
DIC	2747.92				5913.58				6470.64			

* significant at 10%, ** significant at 20%, shaded area means explanatory variables were not significant in the models
All other explanatory variables are significantly different from zero at 95% Bayesian credible interval

Table 5-5: Pedestrian crash Bayesian Poisson-lognormal models by geographic entities

Variables	Geographical Units											
	Census Tracts (N=457)				Block Groups (N=1338)				Traffic Analysis Zones (N=1479)			
	Mean	Std. Dev.	Bayesian Credible Interval		Mean	Std. Dev.	Bayesian Credible Interval		Mean	Std. Dev.	Bayesian Credible Interval	
			2.5%	97.5%			2.5%	97.5%			2.5%	97.5%
Response variable: Pedestrian crashes	3.6433	4.0156			1.2444	1.8799			1.1258	1.7423		
Intercept	3.5050	0.8638	1.5380	5.2000	-1.3890	0.3470	-2.0450	-0.7941	-1.8920	0.2345	-2.3730	-1.4440
Log of VMT	0.2838	0.0796	0.1234	0.4297	0.0730	0.0200	0.0346	0.1127	**0.0261	0.0190	-0.0103	0.0611
Log of roadway length with 25mph PSL									0.2172	0.0814	0.0449	0.3717
Log of roadway length with 35mph PSL					0.4173	0.0765	0.2673	0.5630	0.4202	0.0924	0.2372	0.6037
Log of roadway length with 45mph PSL												
Log of roadway length with 55mph PSL												
Log of roadway length with 65mph PSL	-0.2186	0.0990	-0.4096	-0.0223								
Log of total number of intersection	0.3651	0.0546	0.2656	0.4711	0.3970	0.0461	0.3092	0.4851	0.3783	0.0459	0.2914	0.4659
Log of population age 0 to 15	**-0.0750	0.0594	-0.2093	0.0380					-0.2529	0.0715	-0.3990	-0.1255
Log of population age 16 to 64												
Density of children (K to 12th grade)									4.1870	0.8594	2.5040	5.9160
Proportion minority population	0.5003	0.1746	0.1611	0.8487	0.7802	0.1462	0.4903	1.0640	0.4536	0.1983	0.0650	0.8475
Log of no of workers: travel time 0-4 min												
Log of no of workers: travel time 5-9 min												
Log of no of workers: travel time 15-19 min									**0.1126	0.0760	-0.0350	0.2646
Log of no of workers: travel time ≥ 30 min												
Log of no of workers: worked at home					-0.0602	0.0245	-0.1084	-0.0122	-0.1161	0.0350	-0.1852	-0.0450
Log of no of workers: commute by public transport	0.0727	0.0332	0.0088	0.1369	0.0596	0.0243	0.0119	0.1075	*0.0607	0.0342	-0.0071	0.1246
Log of no of workers: commute by walking	0.1098	0.0363	0.0384	0.1795	0.1355	0.0239	0.0896	0.1825	0.1321	0.0336	0.0665	0.1975
Housing units per acre	0.0602	0.0230	0.0121	0.1045	0.0704	0.0144	0.0424	0.0972	0.6148	0.2907	0.0194	1.1700
Log of median household income	-0.7019	0.1369	-0.9269	-0.2620	-0.1294	0.0255	-0.1749	-0.0750				
Stdev of θ[i]	0.4727	0.0609	0.3901	0.5566	0.6376	0.1142	0.0930	0.7312	0.6799	0.1466	0.0561	0.7914
DIC	1847.26				3535.33				3849.09			

* significant at 10%, ** significant at 20%, shaded area means explanatory variables were not significant in the models
All other explanatory variables are significantly different from zero at 95% Bayesian credible interval

5.5 Summary and Conclusion

The study of this chapter was carried out particularly to understand the zoning and scale effects in modeling three types of crashes (total, severe and pedestrian crashes) in three types of geographic entities (CT, BG, and TAZ). These models were developed based on various roadway characteristics and census variables. As shown in the results, the significant variables are not consistent in the same response models across the geographic units. However, it was observed that TAZ based models have more roadway/traffic related explanatory factors whereas BG based models include more of the commute related variables. It is hoped that the relative comparison of macro-level crash models developed in this chapter for different spatial units will be a contribution to the growing body of the macro analysis and safety planning literature in regard to the scale and zoning effects. Coefficient estimates from all models based on the three geographic units were discussed.

For many of the variables, their causal effects onto the response variables (total, severe or pedestrian crashes) were explainable. But there were a few variables for which a satisfactory elucidation in terms of their association with particular crash type was difficult to find. This is most likely due to the information loss which occurs during the process of aggregation of data for a spatial unit. This limitation can be compensated while models are to be used for prediction only which is also one of the important objectives in TSP. So far, TAZs have been the base spatial units of analyses for developing travel demand models. MPOs widely use TAZs in developing their long range transportation plans (LRTPs).

You et al. (1997) cited that the most important criteria used to define TAZs include spatial contiguity, homogeneity, and compactness. Additionally, TAZs are one of the Census Transportation Planning Products (CTPPs). Note that CTPP is a set of special tabulations from decennial census demographic surveys designed for transportation planners (FHWA, CTPP). Therefore, TAZs seem to be preferred spatial units compared to BGs or CTs. TAZ has a relative ascendancy over block group and census tract in terms of crash prediction models and integration with LRTPs.

Admittedly, TAZs are now the only traffic related zone system, thus TAZs are being widely used for the macroscopic crash analysis in practice. This chapter showed that TAZs and BGs are equally desirable, however TAZs are superior in utilizing more transportation related factors and the ability to be integrated easily with the transportation planning process. Nevertheless, considering that TAZs are not delineated for traffic crash analysis but they were designed for the long range transportation plans, TAZs might not be the optimal zone system for the traffic crash modeling at the macroscopic level. Thus, creating a new zonal system for macroscopic safety analysis is proposed for the dissertation.

CHAPTER 6 DEVELOPMENT OF ZONE SYSTEM FOR MACRO-LEVEL TRAFFIC SAFETY ANALYSIS

6.1 Introduction

Modifiable Areal Unit Problem (MAUP) has been one major topic in the spatial data analysis. MAUP is presented when artificial boundaries are imposed on continuous geographical surfaces and the aggregation of geographic data cause the variation in statistical results (Openshaw, 1984). Assuming that areal units in a particular study were specified differently, it is possible that very different patterns and relationships are shown up (O'Sullivan and Unwin, 2003). MAUP was investigated by Gehlke and Biehl (1934) for the first time. Authors found that the correlation coefficient increases as the unit area enlarges. According to Openshaw (1984), MAUP is composed of two effects: scale effects and zoning effect. Scale effects result from the different level of spatial aggregation. For example, traffic crash patterns are differently described in lower aggregation spatial units such as TAZs and higher aggregation units such as counties or states. Meanwhile zoning effects are from the different zoning configurations at a same level of the spatial aggregation. This phenomenon occurs due to the scale effect of MAUP causing inconsistent statistical results based on the different spatial units. Several transportation planning studies have addressed MAUP (Ding, 1998; Chang et al., 2002; Zhang and Kukadia, 2005; Viegas et al., 2009).

Quite a few studies have been conducted regarding the MAUP on macroscopic traffic crash analysis. For example, the comparison between census block and block group, TAZ, county, etc. In a recent study, Abdel-Aty et al. (2013) compared three models based on different areal units;

census tracts, census block groups and TAZs. The authors concluded that TAZs are practically most adequate units.

Previous studies have contributed significantly to the exploration of the MAUP issue in macroscopic traffic safety analysis. However, these studies are limited to the comparison of existing spatial units. The objective of this chapter is to develop a new zone system based on the current TAZ system. The proposed zone system aggregates TAZs of identical traffic patterns using a state-of-art regionalization method. The new zone system would alleviate the boundary issues while not sacrificing the advantages of TAZs.

6.2 Data Preparation

Crashes that occurred in Orange, Seminole and Osceola Counties in Central Florida in the period of 2008-2009 were obtained for the analysis. Three counties are divided into 1116 TAZs, on which 86,828 crashes occurred and among them, 2,470 crashes were severe crashes with incapacitating injuries or fatalities. Crash data are collected both from Florida Department of Transportation (FDOT) and MetroPlan Orlando (MPO). MPO is the metropolitan planning organization for Orange, Seminole and Osceola Counties. Two forms of crash report are used in the State of Florida, short form and long form crash reports. A short form is usually used to report property damage only (PDO) crashes, whereas a long form is used when if the crash involves injuries or fatalities. Criminal related (e.g., hit-and-run, driving under influence, etc.) PDO crashes are also reported on long forms. Since only long form crashes have been coded and archived in FDOT crash analysis reporting (CAR) database so far, previous researchers could only get access to long form crashes for the crash analysis in Florida. Thus, safety analysts had

been missing many crash data in Florida, especially PDO crashes. Fortunately, a new system started to code short crashes recently, and therefore we could acquire short form crash data of three counties. Thus we were able to use the most complete data in this study.

Overall eleven explanatory variables were prepared. In the macroscopic traffic safety analysis, it is attempted to find out the relationships between the number of crashes and demographic, socioeconomic, and traffic/roadway characteristics at the zonal-level. The variables used in this study were carefully selected considering previous literature and the expected influence on traffic crashes. Regarding demographic factors, population density, proportion of African Americans, and proportion of Hispanics were chosen. The population density represents the number of people residing per square mile at the zone. Guevara et al. (2004) found that the population density is a contributing factor for PDO, injury, and fatal crashes. The authors explained that the population density reflects the degree of interaction among people. Thus, zones with higher population densities may imply greater interaction and possible conflicts. Moreover, there have been several studies that some minority race/ethnic groups are vulnerable to traffic crashes (Baker et al., 1998; Harper et al., 2000). The authors commonly asserted that African Americans and Hispanics are more exposed to traffic crashes and fatalities. Thus, both the proportion of African Americans and proportion of Hispanics variables were chosen in this study. Besides, the proportion of young people aged 15-24 years old and proportion of elderly people aged 65 or older were added in this study. It is known that young drivers tend to be involved in more crashes because of their lack of driving experience. On the other hand, elderly drivers have decreased sensory and motor abilities but they drive less frequently than other age groups, and it is expected that elderly drivers have different crash patterns. Huang et al. (2010) found that the

number of young drivers increases the crash risk whereas elderly population decreases it, if all other conditions remain the same. Demographic data of Census 2010 were obtained from the US Census Bureau. Although these demographic data were provided based on census blocks, they could be aggregated to TAZ based data because TAZs are combinations of one or multiple census tracts.

Concerning the socioeconomic data, the proportion of households without vehicle and number of hotel, motel, and timeshare rooms were selected as explanatory variables. The proportion of household without vehicle is related to traffic safety in two aspects. First, household members with no vehicle would use alternative transportation (i.e., public transportation, bicycle, etc.) modes for activities. It may result in different crash patterns. Second, household vehicle ownership is related to economic status which might affect crash severity. People from low-income households are more vulnerable to severe crashes because they are less likely to purchase newer and safer vehicles or equipment (Martinez and Veloz, 1996). The number of hotel, motel, and timeshare rooms is linked with tourism industry. This variable roughly represents the number of tourists in a zone, and it was included because it is possible that tourists have different driving behavior and less familiarity compared to local drivers. Ng et al. (2002) revealed that hotel rooms were negatively associated with fatal crashes. These socioeconomic data were attained from the MPO based on TAZs.

About roadway/traffic data, the proportion of roadway length with low speed limit (20 mph or lower), proportion of length with high speed limit (55 mph or higher), roadway length with poor pavement condition, and VMT (Vehicle-Miles-Traveled) were selected as explanatory variables.

The variables related to speed limits were included because it was already discovered that the speed limits contribute not only crash counts but also increased severity levels by many researchers (Johansson, 1996; Clark, 2003; Hadayeghi et al., 2003, 2006; Quddus, 2008; Siddiqui et al., 2012; Siddiqui and Abdel-Aty, 2012; Abdel-Aty et al., 2013). The roadway with poor pavement conditions is defined as the roadway that is virtually impassable or has large potholes and deep cracks. The roadway length with poor condition was chosen because it is expected the poor pavement conditions might affect traffic safety (Tighe et al., 2000). Obviously, VMT was included as the indicator of exposure. Roadway and traffic data were obtained from FDOT and MPO. The descriptive statistics of the collected data based on TAZs are summarized in Table 6-1.

Table 6-1: Descriptive statistics of collected data

Variables	Mean	S.D.	Max
Total crash frequency	80.214	90.379	729
Severe crash frequency	2.278	2.812	21
Population density (Population per mi^2)	1189.1	1437.3	14904.4
Proportion of African Americans	0.176	0.202	1
Proportion of Hispanics	0.237	0.168	1
Proportion of young people (15-24)	0.145	0.078	1
Proportion of elderly people (65 or more)	0.118	0.095	1
Proportion of households without vehicle	0.069	0.080	0.557
Hotel, motel, and timeshare rooms	296.2	1355.1	14341.0
Proportion of roadway length with low speed limit (20 mph or lower)	0.048	0.083	1
Proportion of roadway length with high speed limit (55 mph or higher)	0.059	0.134	0.908
Roadway length with poor pavement condition	0.150	0.448	5.126
Vehicle-miles traveled	93948.5	100954.5	839699.0

* The minimum values for all variables are zero.

6.3 Optimal Zone Scale for Traffic Safety Analysis Zones

6.3.1 Regionalization

Many researchers used TAZs as the basic spatial units for their macroscopic safety studies (Ng et al., 2002; Hadayeghi et al., 2003, 2006, 2010a, 2010b; Guevara et al., 2004; Naderan and Shashi, 2010; Abdel-Aty et al., 2011; Siddiqui et al., 2012; Siddiqui and Abdel-Aty, 2012; Abdel-Aty et al., 2013). TAZs may be reasonable for the traffic safety research because they are transportation based zonal system. However, it is essential that we investigate whether TAZs are appropriate spatial units for macroscopic traffic safety modeling. General zoning criteria for TAZs are as follow (Baass, 1980; Meyer and Miller, 2001).

1) Homogeneous socioeconomic characteristics for each zone's population.
2) Minimizing the number of intrazonal trips.
3) Recognizing physical, political, and historical boundaries.
4) Generating only connected zones and avoiding zones that are completely contained within another zone.
5) Devising a zonal system in which the number of households, population, area, or trips generated and attracted are nearly equal in each zone.
6) Basing zonal boundaries on census zones.

Criteria 1), 4), 5) and 6) are also reasonable for the macroscopic modeling. Nevertheless, possible limitation of TAZs for the crash analysis arise from criteria 2) and 3). Basically, TAZs were designed to find out origin-destination pairs of trips generated from each zone. Thus, transportation planners need to minimize trips which start and end in the same zone. However, it is thought that minimizing intrazonal trips may end up with the small size of TAZs for traffic safety analysts. Especially in the downtown area, TAZs consist of few buildings. In the context

of transportation planning, these small zones in downtown can be separated since they attract a large number of trips, regardless of their sizes. However, in the point of view of traffic safety, these small zones have only short and low-speed access roadways. Thus, it is difficult to analyze traffic crashes with these small zones at the macroscopic level. Moreover, the small size of zones makes many zones with zero crash frequencies, especially for rarely occurring crashes such as severe, fatal or pedestrian crashes. Criterion 3) indicates that TAZs are usually divided based on physical boundaries, mostly arterial roadways. Considering that many crashes occur on arterial roads, between zones, inaccurate results will be made from relating traffic crashes on the boundary of the zone to only the characteristics of that zone (Siddiqui and Abdel-Aty, 2012). A simple way to overcome these two issues while using TAZs for safety analysis is to aggregate TAZs into sufficiently large and homogenous traffic crash patterns. It can be achieved by a regionalization process. Regionalization is a process of aggregating a large number of units into a smaller number of regions while optimizing an objective function (Guo, 2008). The objective function (i.e., sum of squared differences) can be expressed as follows:

$$\text{Minimize SSD} = \sum_{r=1}^{k}\sum_{i=1}^{n_r}(x_i - \bar{x})^2 \quad (20)$$

where, k is the number of regions, n_r is the number of data objects in region r, x_i is a variable value at observation i and $\bar{x}$ is the regional mean for the variable. The constraint of the objective function is that only regions that are adjacent can be aggregated.

1,116 TAZs in Orange, Seminole and Orange Counties were combined into 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1,000 zones in keeping with homogenous total crash rates (crashes per miles) in each zone. The roadway length was used as the exposure measure instead of VMT. Although using VMT is more reasonable for the exposure measure since it includes both traffic

and roadway length information, VMT is only available for arterial roads. Accordingly, it was found that there are 23% of TAZs with zero VMT, which means these TAZs do not include any arterial roads but only contain local/residential roads. In contrast, the roadway length data from MPO is complete and it includes not only arterial roads but also local/residential roads. Also, there are no TAZs with zero roadway length. Therefore, the roadway length was used as the exposure measure for the regionalization process.

6.3.2 Brown-Forsythe test for homogeneity of variance

It is essential to set reasonable criteria for the optimal zone scale. As the zone aggregation level becomes higher, it is expected that the number of boundary crashes decreases; however, it will end up with a too large zone system. In this case, we may lose a lot of the local features of traffic crash patterns. On the other hand, as the zone aggregation level remains lower, the number of boundary crashes may still be large, and zone sizes may not be large enough for the macroscopic crash analysis. Thus, this is a trade-off between the local and global features of traffic safety. On one hand, there is a need to merge multiple TAZs with homogenous traffic crash patterns into single TSAZ to capture the global patterns of traffic safety, and on the other hand, it is necessary to guarantee that the new zone system can capture the safety features of the original TAZs as much as possible.

The optimal zone scale for TSAZs was determined using the Brown-Forsythe (F_{BF}) test. F_{BF} test evaluates whether the variance of variables of interest, such as crash rates, is equal when the scales of zone systems change. The underlying assumption of the F_{BF} test is that there is a greater variance in crash rates among smaller zones and lower variance among larger zones. A high

variance value means that the crash risks are local, whereas a low variance means that they capture more global characteristics. The optimal zone scale ensures that the variance of crash rate is somewhere in between. Root et al. (2011) and Root (2012) used F_{BF} test in medical assumption of F_{BF} test is that there is greater variance in crash rates among smaller zones and a lower variance among larger zones. A high variance value means that the crash risks are local, whereas a low variance means studies for the disease analysis. F_{BF} statistics is calculated using the following formula:

$$F_{BF} = \left[\sum_{i=1}^{t}(\overline{D}_i - \overline{D})^2/(t-1)\right] \Big/ \left[\sum_{i=1}^{t}\sum_{j=1}^{n_1}\left(\overline{D}_{ij} - \overline{D}_1\right)^2/(N-t)\right] \quad (21)$$

where, n_i is the number of samples in the *i*th zone system,

N is the total number of samples for all zone systems,

t is the number of neighborhood groups,

y_{ij} is the crash rates of the jth sample from the *i*th zone system,

$\bar{y}_i$ is the median of crash rate from the *i*th zone system,

$D_{ij}=\left|y_{ij} - \bar{y}_i\right|$ is the absolute deviation of the *j*th observation from the *i*th zone system median,

$\overline{D}_i$ is the mean of D_{ij} for zone system *i*, and

$\overline{D}$ is the mean of all D_{ij}.

The test assumes that the variances of different zones are equal under the null hypothesis. The calculated value was obtained using an *F* distribution with (*t* - 1, *N* - *t*) degrees of freedom and α=0.01 was used to test for statistical significance.

There are two steps involved in the F_{BF} test. First, the variance between each zone system from N200 to N1000 and the largest zone system (N100) is compared for a total of 9 separate calculations of F_{BF}, as shown in the F_{BF1} column of Table 6-2. Second, the variance between each neighborhood group from N900 to N100 and the smallest zone system (N1000) is compared (F_{BF2}). N1000 was used as the smallest zone system instead of TAZs (N1116) since the variance of crash rates based on TAZs is quite large (var=103.10), which shows the crash rates are not relevant to TAZs. A significant value of F_{BF1} implies that the zone system does not reflect the global pattern of crash data; in essence each zone is so small that it only captures local crash patterns. On the contrary, the significant value of F_{BF2} indicates that the zone data are not local; they are so large that local level crash patterns are undetectable. The zone systems between lower and upper limits identify a spatial scale at which local level variation is still detectable but also captures larger zonal level crash characteristics.

The F_{BF} test results for homogeneity of variance for crash rates under various zone scales are presented in Table 6-2. The F_{BF1} test statistics shows that zone systems smaller than N700 (i.e., N800, N900 and N1000) have significantly different variance from that of N100. Thus, zone systems smaller than N700 are too small to capture global crash patterns. On the other hand, F_{BF2} test statistics indicates that zone systems larger than N500 (i.e., N400, N300, N200 and N100) are so large that they cannot capture local crash characteristics. Given the result, systems with 500-700 zones are considered optimal for macro-level crash analysis. Finally, the zone system with 500 zones were chosen for TSAZ since, it can minimize boundary crashes and zones without certain types of crashes.

Table 6-2: Brown-Forsythe test for determining TSAZ scale

Zone systems	Crashes per miles		Brown-Forsythe test			
	Mean	Var	F_{BF1}	CV*	F_{BF2}	CV
TAZs	7.94	103.10	**-**	**-**	**-**	**-**
N1000	6.98	63.02	5.32	2.407	-	-
N900	6.59	55.09	4.38	2.511	0.54	6.635
N800	6.27	44.94	3.53	2.639	1.31	4.605
N700	**5.99**	**40.05**	**2.92**	**2.802**	**1.77**	**3.782**
N600	**5.65**	**35.18**	**2.02**	**3.017**	**2.6**	**3.319**
N500	**5.32**	**29.99**	**1.45**	**3.319**	**3.61**	**3.017**
N400	4.71	24.99	1.31	3.782	4.2	2.802
N300	3.91	18.76	0.84	4.605	4.76	2.639
N200	3.18	12.53	0.4	6.635	5.23	2.511
N100	2.67	9.06	-	-	5.32	2.407

*Critical value of the *F* distribution with (t - 1, N - t) degrees of freedom (α=0.01)

Before regionalization, there were 1039 TAZ in the urban area and 77 TAZ in the rural area. In TSAZ, there are 428 and 72 zones in the urban and the rural area, respectively. Thus, 58.8% of urban zones were reduced but only 6.5% of rural zones decreased after the regionalization. It shows that regionalization aggregation was conducted mostly in urban areas (Table 6-3). Moreover, it was observed that the smallest TAZ is the urban area only is only 0.021m^2, which is considered extremely small for the macroscopic safety analysis. In this case, one TAZ contains only a couple of buildings in the downtown area. After the regionalization, the minimum zone size of TSAZs became 1.051 m^2, which is approximately 50 times larger than that of TAZs (Table 6-3).

Table 6-3: Areas of TAZ and TSAZ

Zone system		Area (mi^2)				
		No of zones	**Average**	**S.D.**	**Min**	**Max**
TAZs	Total	1116	2.544	10.710	0.021	170.664
	Urban	1039	0.964	1.377	0.021	12.403
	Rural	77	23.870	33.890	0.889	170.664
TSAZs	Total	500	5.678	15.493	1.051	170.664
	Urban	428	2.337	1.624	1.051	12.403
	Rural	72	25.541	34.502	1.265	170.664

Table 6-4 summarizes crash rates of TAZs and TSAZs. As expected, both crash rates and their standard deviations are reduced after the regionalization. The overall average crash rate of TAZs (7.942) is higher than that of TSAZs (5.318). The average crash rate was largely influenced by several zones with extremely high crash rates, and it was lowered after the regionalization because the number of small TAZs with high crash frequency and short roadway were decreased. Especially in the urban area, the average crash rate considerably decreased from 8.431 to 6.015 after the regionalization. On the contrary, no big difference was observed in crash rates between TAZs and TSAZs in the rural area. Concerning the standard deviation, the overall standard deviation of TAZs was 10.154, and it became 5.476 in TSAZs. As explained in the regionalization process, it is because zones with homogenous crash rates have been merged and the standard deviation decreased extensively. The standard deviation decreased slightly in the rural area while it was reduced largely in the urban area, since the regionalization aggregation was conducted mostly in urban areas.

Table 6-4: Crash rates of TAZ and TSAZ

Zone system		Crash rates (crash per mile)				
		No of zones	**Average**	**S.D.**	**Min**	**Max**
TAZs	Total	1116	7.942	10.154	0.000	96.946
	Urban	1039	8.431	10.352	0.000	96.946
	Rural	77	1.345	1.127	0.000	4.493
TSAZs	Total	500	5.318	5.476	0.000	44.127
	Urban	428	6.015	5.621	0.000	44.127
	Rural	72	1.306	1.204	0.000	4.493

Zones without total crashes were reduced slightly from 1.5% to 0.8% after the regionalization. However, zones without severe crashes were significantly lessened from 30.6% to 14.2% (Table 6-5). Admittedly, the number of zones without a certain crash types itself is not a critical problem because it can be handled by statistical modeling. Nevertheless, it may imply that zones are too small for the crash analysis since more than 30% of zones do not have any severe crashes. Considering Tables 6-3 and 6-5, it is thought that zone sizes became enlarged enough for the macroscopic safety analysis after the regionalization. Furthermore, before the regionalization, 76.5% of crashes occurred near boundaries of TAZs and it was decreased to 61.9% after the regionalization (Table 6-6).

Table 6-5: Zones without crashes in TAZs and TSAZs

Zone system	Zones without total crashes		Zones without severe crashes	
	Zones	Percentage	Zones	Percentage
TAZs	17	1.5%	341	30.6%
TSAZs	4	0.8%	71	14.2%

Table 6-6: Crashes occurred near boundaries of TAZs and TSAZs

Zone system	Boundary crashes	Total crashes	Percentage
TAZs	68,451	89,527	76.5%
TSAZs	55,411	89,527	61.9%

Figure 6-1 shows TAZs in the overall study area and Downtown Orlando. In comparison with TSAZ shown in Figure 6-2, TAZs in the downtown area are much smaller than TSAZs whereas there is no big difference in the rural area as mentioned earlier.

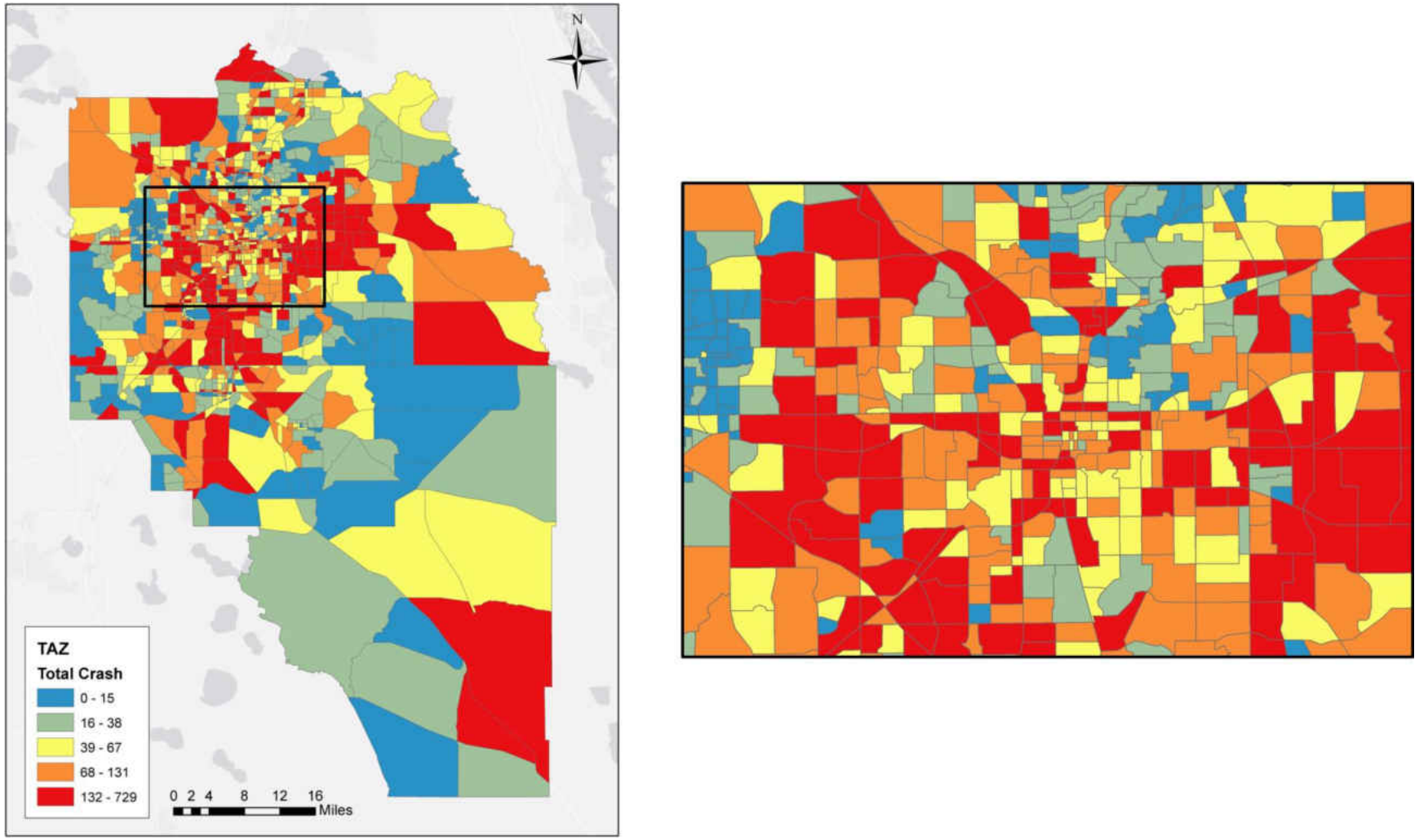


Figure 6-1: Total crashes based on TAZ in the overall study area (left) and TAZ in Downtown Orlando (right)

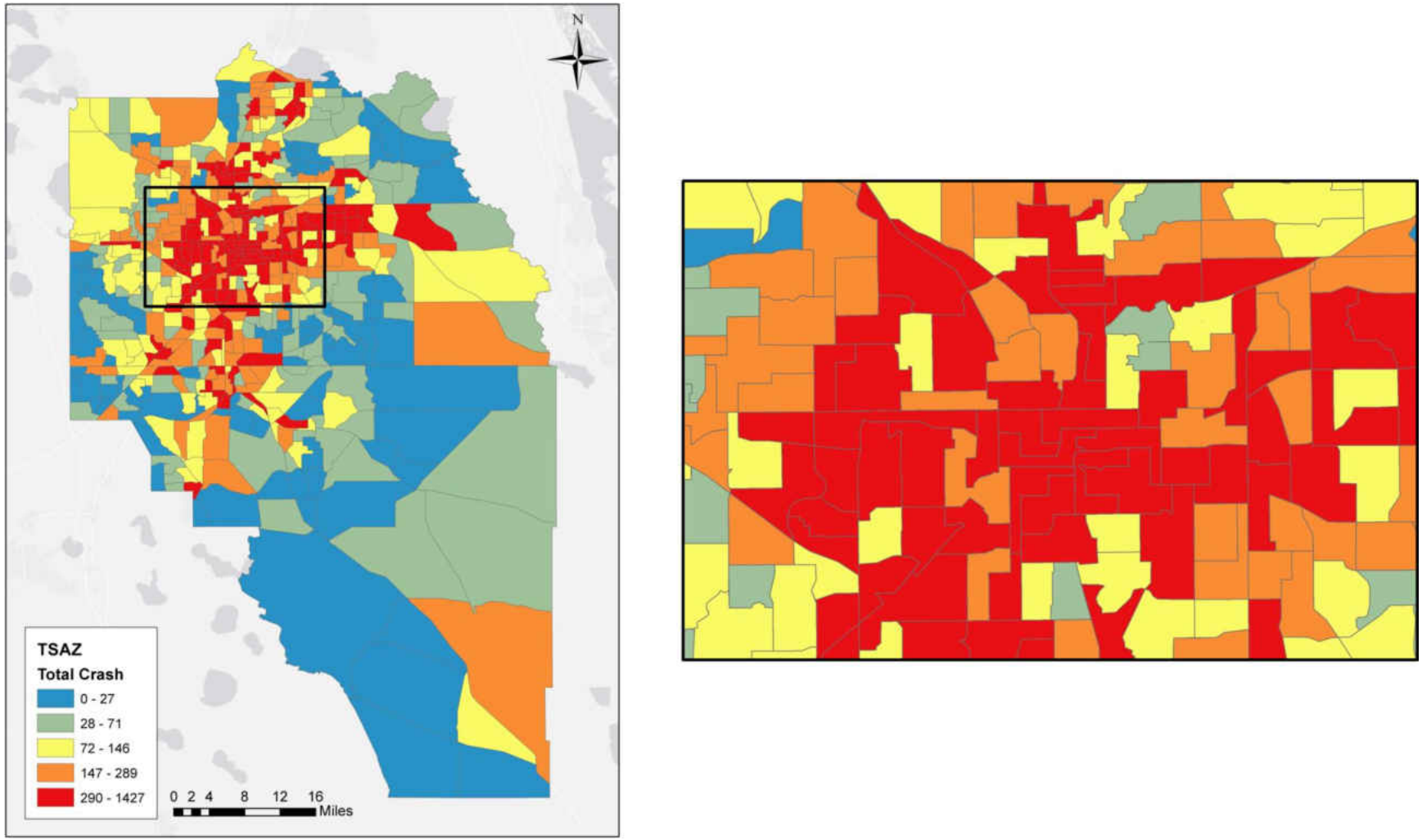


Figure 6-2: Total crashes based on TSAZ in the overall study area (left) and TSAZ in Downtown Orlando (right)

6.4 Crash Model Estimation

6.4.1 Bayesian Poisson-Lognormal Model

The Bayesian Poisson-lognormal model was adopted in this chapter. Poisson-lognormal models have been proposed as a substitute of the negative binomial (or Poisson-gamma) model for the frequency data in traffic safety modeling (Park and Lord, 2007; Aguero-Valverde and Jovanis, 2008, 2009; Ma et al., 2008; El-Basyouny et al., 2009, 2010; Abdel-Aty et al., 2013). The Poisson-lognormal model is comparable with the negative binomial model; however, the Poisson-lognormal model provides more flexibility compared to the negative binomial model (Lord and Mannering, 2010).

A Poisson-lognormal model in a Bayesian framework was fitted for the response variable (i.e., crashes per TSAZ). Unlike classical regression methods, Bayesian models do not usually depend on the assumption of asymptotic normality. Sample based methods of Bayesian estimation focus on estimating the entire density of a parameter as compared to the traditional classical estimation methods which are aimed at finding a single optimum estimate. Thus, Bayesian models are thought to have several advantages compared to the classical likelihood based inference methods and have been popular in recent traffic safety research. A Poisson-lognormal model is specified as follows:

$$y[i] \sim Poisson(\mu[i]) \tag{22}$$

$$log(\mu[i]) = \beta_0 + \beta X[i] + \theta[i] \tag{23}$$

$$\theta[i] \sim Normal\ (0, \tau_\theta) \tag{24}$$

where,

β_0 = intercept term,

β's are the coefficient estimates of model covariates $X[i]$,

$\theta[i]$ = error component of the model, and

τ_θ = precision parameter which is inverse of the variance; τ_θ follows a prior gamma (0.5, 0.005)

This variance ($1/\tau_\theta$) provides the amount of variation not explained by the Poisson assumption (Lawson et al., 2003).

In the modeling process, VMT was used as the exposure measure, instead of the roadway length. Some may think using roadway length as the exposure measure is more consistent because the roadway length was also used for the regionalization process. However, VMT contains both traffic and roadway length information, and thus it is expected that VMT better explains the variation in the crash data.

6.4.2 Comparison of Two Models based on TAZ and TSAZ

Table 6-7 presents the result of model estimations for total crashes based on TAZs and TSAZs. It was revealed that significant variable sets for total crash models based on the two zone systems are exactly same. Five factors were found significant for total crash models. First of all, the exposure measure 'Vehicle-miles-traveled' was significant and positively related to total crashes. For demographic variables, 'Log of population' and 'Proportion of Hispanics' were found commonly significant and positively associated to total crashes. 'Proportion of households without vehicle', which represents not only lower car-ownership but also deprived economic

status, had a negative relationship with total crash frequency. 'Log of hotel, motel and timeshare rooms', which stands for the activity of tourism industry in a zone, also had a positive association with total crashes.

On the other hand, significant variable sets for severe crash models based on the two zone systems are quite different (Table 6-8). Five variables were significant for severe crashes estimated based on TAZs whereas the severe model based on TSAZs had 7 significant variables. Only 3 variables including 'Vehicle-miles-traveled', 'Proportion of Hispanics' and 'Log of hotel, motel and timeshare rooms' were commonly significant in the two models. 'Log of population density' and 'Roadway length with poor pavement condition' were solely significant for the severe crash model based on TAZs. Meanwhile, 'Proportion of African Americans', 'Proportion of young people (15-24)', 'Proportion of households without vehicle' and 'Proportion of roadway length with high speed limit (55 mph or higher)' were solely significant factors for the severe crash model based on TSAZs.

One of the possible reasons for the difference in contributing factors between TAZs and TSAZs based severe crash model is that TSAZs were not developed with the consideration of severe crashes. As explained in the previous chapters, the regionalization process was conducted solely using total crash rates. Also, the optimal scale was determined by Brown-Forsythe test only using the variance of the total crash rates. Thus, different significant variable sets in the TAZs and TSAZs based severe crash models might be caused by the regionalization based on total

crash rates. It may imply that the zone system designed based on total crash pattern is not necessarily optimal for other crash types.

Although DIC[1] (Deviance Information Criterion) values for the four models were presented in both Table 6-7 and 6-8, DIC is not appropriate for comparing models based on different zone systems since they were estimated based on different zone systems and so they have different sample sizes (Abdel-Aty et al., 2013). Instead, other goodness-of-fit measures were suggested for the comparison in Table 6-9. MAD (Mean Absolute Deviation) calculates sum of absolute deviations divided by the number of observations as follows:

$$MAD = \sum_{i=1}^{N} |y[i] - \hat{y}[i]| / N \quad (25)$$

where,

N=number of observations, and

$y[i]$and $\hat{y}[i]$are observed and predicted values for i, respectively.

RMSE (Root Mean Squared Errors) computes the square root of the sum of the squared errors divided by the number of observations as the following formula:

[1] $DIC = 2 \times \overline{D} - \widehat{D}$

where, $\overline{D}$ =posterior mean of deviance, D, $\widehat{D} = -2 \times p(y|\bar{\theta})$, and $\bar{\theta}$ is posterior mean of θ (Spiegelhalter et al., 2002).

$$RMSE = \sqrt{\sum_{i=1}^{N}(y[i] - \hat{y}[i])^2 / N} \quad (26)$$

SAD (Sum of Absolute Deviation) calculates sum of absolute deviations as follows:

$$SAD = \sum_{i=1}^{N}|y[i] - \hat{y}[i]| \quad (27)$$

Lastly, PMAD (Percent Mean Absolute Deviation) computes sum of absolute deviations divided by the sum of absolute observed values.

$$PMAD = \sum_{i=1}^{N}|y[i] - \hat{y}[i]| \Big/ \sum_{i=1}^{N}|y[i]| \quad (28)$$

Although MAD and MSPE are widely used for comparing model performance, MAD and RMSE are not appropriate for comparing models with different sample sizes because they are largely influenced by the number of observations, as seen in the formulae above. On the contrary, SAD and PMAD do not rely on the number of observations and they are not influenced by sample sizes. Thus, SAD and PMAD were used to compare models from different zone systems. As shown in Table 6-9, MAD and RMSE are smaller in TAZ based models compare to those in TSAZ based models. It was as expected because both MAD and RMSE depend largely on the number of observation. In contrast, SAD and PMAD, which are not influenced by the number of observations, are smaller in TSAZ based models. Therefore, it is concluded that TSAZs based models outperforms TAZs based models for both total and severe crash models.

Table 6-7: Bayesian Poisson lognormal model for total crashes based on TSAZ and TAZ

Zone systems	**TAZs (N=1,116)**				**TSAZs (N=500)**			
Variables	**Mean**	**S.D.**	**Bayesian Credible Interval**		**Mean**	**S.D.**	**Bayesian Credible Interval**	
			2.5%	**97.5%**			**2.5%**	**97.5%**
Intercept	-0.211	1.156	-5.213	0.416	-0.784	1.294	-5.757	0.126
Vehicle-miles-traveled	***0.264**	**0.100**	**0.214**	**0.658**	***0.295**	**0.142**	**0.215**	**0.852**
Log of population density	***0.108**	**0.025**	**0.080**	**0.208**	***0.171**	**0.027**	**0.097**	**0.214**
Proportion of African Americans	0.124	0.184	-0.313	0.451	0.690	0.363	-0.210	1.301
Proportion of Hispanics	***0.594**	**0.207**	**0.170**	**0.998**	***0.604**	**0.266**	**0.040**	**1.110**
Proportion of young people (15-24)	0.889	0.446	-0.086	1.729	1.250	0.887	-1.143	2.582
Proportion of elderly people (65 or older)	0.050	0.356	-0.671	0.751	0.816	1.209	-2.826	2.736
Proportion of households without vehicle	***2.086**	**0.549**	**1.276**	**3.679**	***2.581**	**0.970**	**0.487**	**4.416**
Log of hotel, motel, and timeshare rooms	***0.113**	**0.013**	**0.088**	**0.140**	***0.111**	**0.015**	**0.083**	**0.142**
Proportion of roadway length with low speed limit (20 mph or lower)	0.567	0.624	-0.285	2.439	0.731	0.863	-1.397	2.268
Proportion of roadway length with high speed limit (55 mph or higher)	0.071	0.376	-0.791	0.591	-0.079	0.415	-1.123	0.590
Roadway length with poor pavement condition	0.155	0.078	-0.002	0.290	0.115	0.061	-0.031	0.220
DIC	8427.12				4122.53			

*** significant at 5%**

Table 6-8: Bayesian Poisson lognormal model for severe crashes based on TSAZ and TAZ

Zone systems	**TAZs (N=1,116)**				**TSAZs (N=500)**			
Variables	**Mean**	**S.D.**	**Bayesian Credible Interval**		**Mean**	**S.D.**	**Bayesian Credible Interval**	
			2.5%	**97.5%**			**2.5%**	**97.5%**
Intercept	-5.090	0.408	-5.836	-4.367	-4.953	0.509	-5.955	-4.069
Vehicle-miles-traveled	***0.447**	**0.037**	**0.383**	**0.514**	***0.460**	**0.045**	**0.383**	**0.547**
Log of population density	***0.041**	**0.012**	**0.017**	**0.066**	0.025	0.018	-0.009	0.060
Proportion of African Americans	0.120	0.167	-0.209	0.442	***0.606**	**0.233**	**0.154**	**1.067**
Proportion of Hispanics	***0.633**	**0.190**	**0.257**	**1.005**	***0.734**	**0.224**	**0.300**	**1.178**
Proportion of young people (15-24)	0.585	0.388	-0.187	1.346	***0.959**	**0.461**	**0.047**	**1.863**
Proportion of elderly people (65 or older)	0.549	0.361	-0.159	1.252	1.119	0.632	-0.119	2.365
Proportion of households without vehicle	0.707	0.414	-0.108	1.516	***1.882**	**0.758**	**0.389**	**3.354**
Log of hotel, motel, and timeshare rooms	***0.050**	**0.012**	**0.026**	**0.074**	***0.037**	**0.013**	**0.012**	**0.061**
Proportion of roadway length with low speed limit (20 mph or lower)	0.154	0.442	-0.721	1.006	-0.131	0.677	-1.477	1.190
Proportion of roadway length with high speed limit (55 mph or higher)	-0.200	0.261	-0.717	0.305	***-0.696**	**0.326**	**-1.339**	**-0.059**
Roadway length with poor pavement condition	***0.125**	**0.062**	**0.004**	**0.246**	0.053	0.041	-0.028	0.134
DIC	3919.9				2270.81			

*** significant at 5%**

Table 6-9: Comparison of goodness-of-fits between TAZs and TSAZs based models

Crash types	Zones	MAD	RMSE	SAD	PMAD
Total crash	TAZs	46.834	1351.771	52266.667	0.584
	TSAZs	90.316	2337.436	45158.046	0.504
Severe crash	TAZs	1.560	52.127	1741.393	0.685
	TSAZs	2.638	58.982	1318.867	0.519

Furthermore, Figure 6-3 compares the probability distribution of predicted and observed crashes from two models based on TAZs and TSAZs. Upper two graphs show predicted and observed probability distributions of total crashes from TAZs and TSAZs based models. TAZ based total crash model underestimates the proportion of values ranging from 0 to 25 but overestimates the proportion of values ranging from 25 to 200. Although TSAZ based total crash model overestimates the proportion of values range from 50 to 150, it seems TSAZ based total crash model has a better fit compared to TAZ based the total crash model. Regarding distributions of severe crashes, both models have tendencies to underestimate the proportion of zero. TAZ based severe crash model overestimate the proportion of values range from 1 to 2 and underestimates the proportion of values from 4. In contrast, TSAZ based severe crash model overestimates the proportion of values from 4 to 8, but quite accurate at the larger values. It is concluded that TSAZ based severe crash model fits the data better than the TAZ based model. Although better fits were observed from TSAZ based models, it was not a surprising result because the regionalization process always reduces the number of zones and their variances.

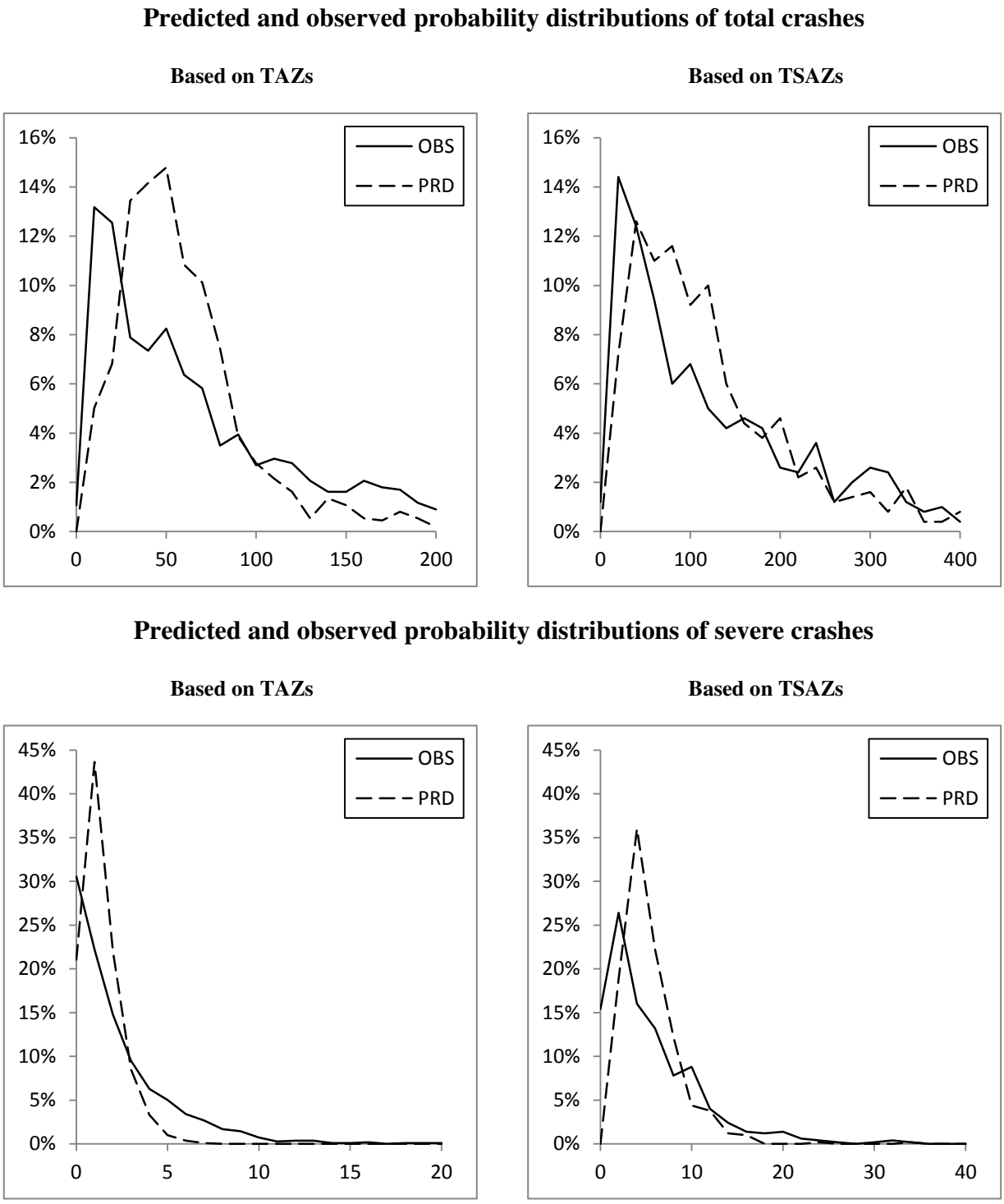


Figure 6-3: Predicted and observed probability distributions of crashes based on two zone systems

6.5 Summary and Conclusion

There have been many efforts to overcome limitations of macroscopic traffic analysis and improve the safety model performance. Generally, it is assumed that traffic crashes are hypothesized to be influenced solely by the zone within which they are spatially located at the macroscopic level. However, this assumption causes traffic safety models to heavily rely on the zone systems and its configurations. Some researchers considered the spatial autocorrelation in crash modeling to overcome this issue. The models accounting for the spatial autocorrelation enable us to use information from adjacent zones to estimate the safety model. It was discovered that considering spatial autocorrelation in crash modeling significantly improves model performance (Levine et al., 1995; Siddiqui and Abdel-Aty, 2012). Nevertheless, just accounting for the spatial autocorrelation cannot fundamentally solve the problem if the zone system itself has limitations. This chapter aims at developing a new zone system for macro-level crash analysis. It was discussed that there are several possible limitations of TAZs for macroscopic crash modeling. First, TAZs might be too small for crash analysis. Second, since TAZs are often delineated by arterial roads, many crashes occur at or near the boundary of TAZs. The existence of many boundary crashes can result in inaccurate modeling results. One simple way to overcome these two possible limitations of TAZs is to combine small TAZs with similar traffic crash patterns (i.e., crash rates) into sufficiently large zones, which is called the regionalization process.

After the regionalization, ten new zones systems were created by different scales of zones. Among these zone systems, the zone system with the optimal scale was determined using

Brown-Forsythe homogeneity of variance test. According to the result, zone systems have 500 to 700 zones were considered optimal for macroscopic crash analysis. In other words, zone systems with 500-700 zones can capture both local and global traffic crash characteristics. Finally, the zone system with 500 zones was chosen as TSAZs since it can minimize boundary crashes as well as zones without rare crashes (i.e., severe crashes). In the result, the proportion of severe crashes and boundary crashes decreased by 16.4% and 14.6% after the regionalization, respectively.

Both total and severe crash models based on TSAZs were estimated and compared with models based on TAZs. It was found that significant variable sets were exactly identical for both TSAZ and TAZ based total crash models. Five variables including an exposure measure, demographic factors, vehicle ownership and scale of accommodation facilities were commonly significant. On the other hand, significant variable sets for severe crash models based on the two different zone systems were quite different. It may imply that TSAZs are not the optimal zone system for severe crashes because TSAZs were not developed with the consideration of severe crashes. Furthermore, TSAZ based models had better fit compared to TAZ based models, in terms of SAD and PMAD but it was as expected because the regionalization process reduced the number of zones with smaller variances compared to TAZs.

The main contribution of this chapter is the adoption of the Brown-Forsythe homogeneity test for variance for developing new zone systems for macroscopic safety modeling. From a traffic safety management aspect, zonal crash hotspots can be identified easily with TSAZs since the

new zone system is more clustered with similar crash patterns. Although TSAZs have less percentage of near boundary crashes compared to TAZs, still more than half of crashes occur on or near the boundaries (61.9%) in TSAZs. Thus, there is a need to account for boundary crashes in the modeling process in follow-up studies. Furthermore, only total crash patterns were used to develop TSAZs. However, it might be useful to also investigate other types of crashes. For instance, if bicycle and pedestrian related crashes have very different spatial distributions from total crashes, it is also essential to develop a different zone system exclusively for the non-motorized crash analysis.

CHAPTER 7 COMPARISON OF CONCEPTUALIZATION METHODS OF SPATIAL AUTOCORRELATIONS IN THE CRASH MODELING

7.1 Introduction

Spatial autocorrelation is the term for the tendency for data from locations near one another in space to be more similar than data from locations remote to each other (O'Sullivan and Unwin, 2002). Most statistical models assume that the values of observations in each sample are independent or randomly distributed. A positive spatial autocorrelation between zones, however, may violate this assumption if the samples were collected from nearby areas (Lai et al., 2008). In this chapter, the spatial autocorrelation effects in the residuals were explored using Moran's *I*. After the existence of spatial autocorrelations in the residuals was identified, the spatial autocorrelation term was included in the SPF to account for the spatial autocorrelations and then compared the corrected SPF using several spatial error terms from different conceptualization methods to determine the method with the best performance.

7.2 Detection of the Spatial Autocorrelation

In order to identify the existence of spatial autocorrelations in the data, Moran's *I* was used. Moran's *I* is one of the measures of spatial autocorrelation developed by Moran (1950) and is calculated from the formula as follows:

$$I = \frac{n}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \quad (29)$$

where, *n* is the number of areal units indexed by *i* and *j*,

y is the value of interest (i.e., the number of crashes),

$\bar{y}$ is the mean of y, and

w_{ij} is an element of the matrix of spatial weights.

Spatial weights calculated using three different conceptualization methods were examined as follows:

1) Inverse distance: w_{ij} is the inverse distance between zones i and j;

2) Inverse distance squared: w_{ij} is the inverse distance squared between zones i and j; and

3) First order polygon contiguity: $w_{ij} = 1$ if zones i and j are adjacent based on the 1st order contiguity, otherwise $w_{ij} = 0$.

A positive value of Moran's I index stands for a positive spatial autocorrelation, whereas a negative value indicates a negative autocorrelation. The value ranges from -1 to +1, where -1 means that regions are perfectly dispersed and +1 indicates that the regions are perfectly correlated. On the contrary, if the index is close to zero, it indicates a random pattern. Moran's I index can be converted to a z-score for the statistical test, in which values greater than 1.96 or smaller than -1.96 show that there is a statistically significant spatial autocorrelation in the regions.

Table 7-1 presents the Moran's I calculated from the residuals of SPF and the corresponding z-values' p-value for each conceptualization method. All of the values of Moran's I showed positive spatial autocorrelations, and they were all statistically significant.

Table 7-1: Moran's *I* of residuals by spatial autocorrelation conceptualization methods

Model	**Conceptualization**	**Moran's *I***	***z***	***p***
Total crashes	Inverse distance	0.075	11.376	<0.001
	Inverse distance squared	0.126	8.045	<0.001
	First order rook polygon contiguity	0.178	6.681	<0.001
Severe crashes	Inverse distance	0.033	5.089	<0.001
	Inverse distance squared	0.060	3.909	<0.001
	First order rook polygon contiguity	0.134	5.032	<0.001

Since statistically significant spatial autocorrelations in residuals both in total and severe crash SPFs were detected, spatial autocorrelations should be accounted in the estimations of SPF. One possible solution to account for spatial autocorrelation is to include a spatial random effect component in the model formulation, which will be discussed in the next section.

7.3 Comparison of Spatial Effect Conceptualization Methods

As mentioned in the previous section, the spatial error term (φ_i) was included in the SPF to account for the spatial autocorrelation using the following equation:

$$y_i \sim Poisson(\mu_i) \tag{30}$$

$$\lambda_i = \exp(\beta_0 + \beta X_i + \theta_i + \varphi_i) \tag{31}$$

$$\theta_i = Normal\ (0, \tau_\theta) \tag{32}$$

where,

y_i is the number of aggregated total crashes of the i^{th} TSAZ,

β_0 is the intercept,

β's are the coefficient estimates of covariates (X_i),

θ_i is the random effect term,

φ_i is the spatial effect term, and

τ_θ is the precision parameter, which is the inverse of the variance and a given prior gamma distribution (0.5, 0.005).

The Bayesian model was fit with non-informative prior distributions, Normal (0, 10^{-6}) for β's.

The spatial distribution was implemented by specifying an intrinsic Gaussian Conditional Autoregressive (CAR) prior with a $Normal\ (0, \tau_\varphi)$ distribution. The mean of φ_i is defined by

$$\bar{\varphi}_i = \sum_{i \neq j} \varphi_j \times w_{ij} \Big/ \sum_{i \neq j} w_{ij} \tag{33}$$

where, values for w_{ij} are defined in Table 7-2 by the different spatial autocorrelation conceptualization methods.

Table 7-2: Definition of w_{ij} by different spatial autocorrelation conceptualization methods

Conceptualization	$\boldsymbol{w_{ij}}$
No spatial error term	$\varphi_i = 0$
First order rook polygon contiguity	$w_{ij} = 1$, if zone *i* and *j* are adjacent; $w_{ij} = 0$, otherwise
Inverse distance	$w_{ij} = 1/d_{ij}$
Inverse distance squared	$w_{ij} = 1/d_{ij}^2$

The Deviance Information Criterion (DIC) was computed in each model for comparison. The following equation is used to calculate DIC (Spiegelhalter et al., 2002):

$$DIC = 2 \times \overline{D} - \widehat{D} \quad (34)$$

where, $\overline{D}$ is the posterior mean of deviance, D,

$\widehat{D} = 2 \times (p(y|\theta))$, and

$\bar{\theta}$ is the posterior mean of θ.

Models with a smaller DIC are preferred to models with a larger DIC. (Spiegelhalter et al., 2003). Table 7-3 summarizes DIC from total and severe crash models with the different spatial autocorrelation conceptualization methods. It was found that only the spatial error term conceptualized by first order rook polygon slightly improves the model performance both in total and severe crash models.

Table 7-3: Comparison of DIC by different spatial autocorrelation conceptualization methods

Conceptualization	DIC	
	Total crash model	Severe crash model
No spatial error term	4122.53	2270.81
First order rook polygon contiguity	4122.32	2247.51
Inverse distance	4121.07	2270.79
Inverse distance squared	4121.66	2272.65

7.4 Summary and Conclusion

In this chapter, it was shown that there are spatial autocorrelations in the data set using Moran's *I* statistics. Overall three conceptualization methods were applied including 1) first order rook polygon contiguity, 2) inverse distance, and 3) inverse distance squared. It was shown that spatial autocorrelations are significant, regardless of conceptualization methods. Two SPFs for total and severe crashes were estimated and each model includes a spatial error term based on different conceptualization methods. It was revealed that there is no big difference in DIC among models using different conceptualization methods. Nevertheless, only the spatial error term conceptualized by first order rook polygon slightly improves the model performance both in total and severe crash models. Thus, final models will adopt a spatial error component based on first order rook polygon contiguity in the following chapters.

CHAPTER 8 NESTED STRUCTURE AND VARIABLE TRANSFORMATION TO ACCOUNT FOR BOUNDARY CRASHES

8.1 Nested Modeling Structure

In this chapter, a nested structure was adopted which allows different contributing factors for different crash types (such as boundary and interior crashes or crashes located on different roadway types). It was expected that more accurate and predictable results are shown from this nested structure than those that could be obtained from a single model.

The nested structure includes six sub-models which are named based on their locations (i.e., near the zone boundary or the interior) and roadway types (i.e., freeway-and-expressway, other state roads, and non-state roads). The nested structure is presented in Figure 8-1. Both total crashes and fatal-and-injury crashes were modeled using the nested structure. In addition, a Bayesian Poisson Lognormal Spatial Error Model (BPLSEM) was adopted for the SPF analysis in this nested structure. This model has a disturbance term for handling the over-dispersion problem, and its spatial error term could control for the spatial autocorrelation of crash data. See Appendix A for more details of the model formulation. It was assumed that factors contributing to crashes on freeway-and-expressway are different than those contributing to crashes on other state roads. Thus, a nested structure with 6 models as shown in Figure 8-1 was constructed.

The six types of crashes in each model are varied based on their locations (boundary or interior) and roadways (freeway-and-expressway, other state roads or non-state roads). They are FSB (Freeway-and-expressway State road Boundary crashes), FSI (Freeway-and-expressway State

road Interior crashes), OSB (Other State road Boundary crashes), OSI (Other State road Interior crashes), NSB (Non-state road Boundary crashes) and NSI (Non-state road Interior crashes). For better illustration, Figure 8-2 illustrates examples of these six crash types.

Meanwhile, some zones have zero probability to have specific types of crashes. For instance, zone #1 in Figure 8-2 does not have any freeway-and-expressway or other state roads. The expected numbers of FSB, FSI, OSB or OSI in zone #1 should be zero, regardless of zonal characteristics. It is meaningless to include these zones without freeway-and-expressway or state roads in the estimation of FSB, FSI, OSB or OSI models. Therefore, the zones without specific types of roads were excluded when the models for crashes that occurred on those types of roads were estimated.

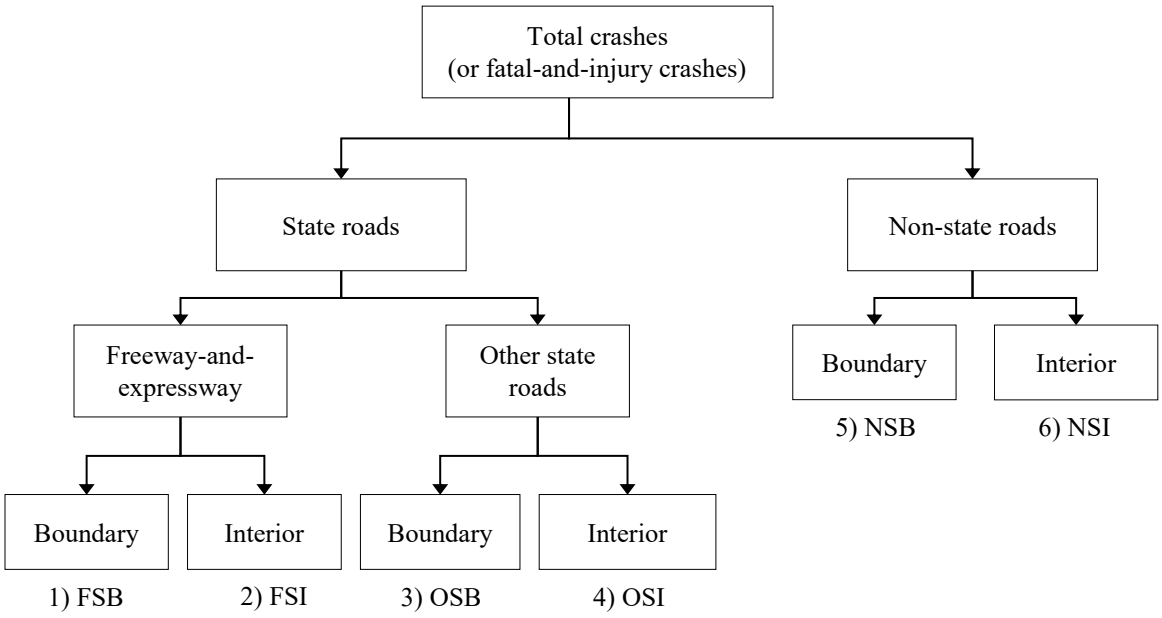


Figure 8-1: Nested structure for macroscopic crash modeling (with six sub-models)

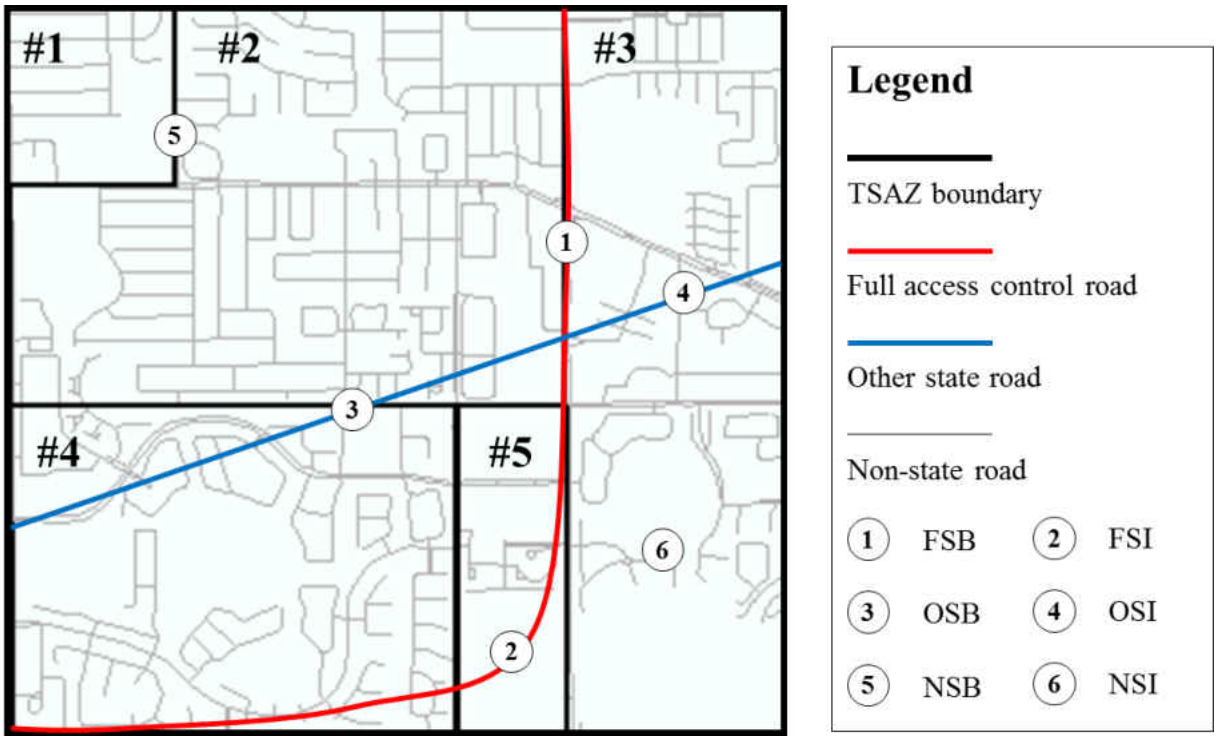


Figure 8-2: Examples of crashes by locations used in the nested structure

8.2 Variable Transformation for Boundary Crashes

It was assumed that interior crashes were influenced only by the characteristics of the zone in which they were located. Thus, the models for the interior crashes were developed using individual zonal characteristics. In contrast, crashes occurring near or on zone boundaries, known as boundary crashes, were hypothesized to be influenced not only by the zonal characteristics of where the crashes occurred, but also by the characteristics of the adjacent zones. Therefore, the models for boundary crashes were estimated using 'transformed' variables possessing information for both the crash zone and any adjacent zones.

Let any TSAZ *i* share its boundary with adjacent zones *j* = 1, 2, … *k*, as shown in Figure 8-3. An original variable *x* will be transformed to x_{ABC} using the following Equation (35).

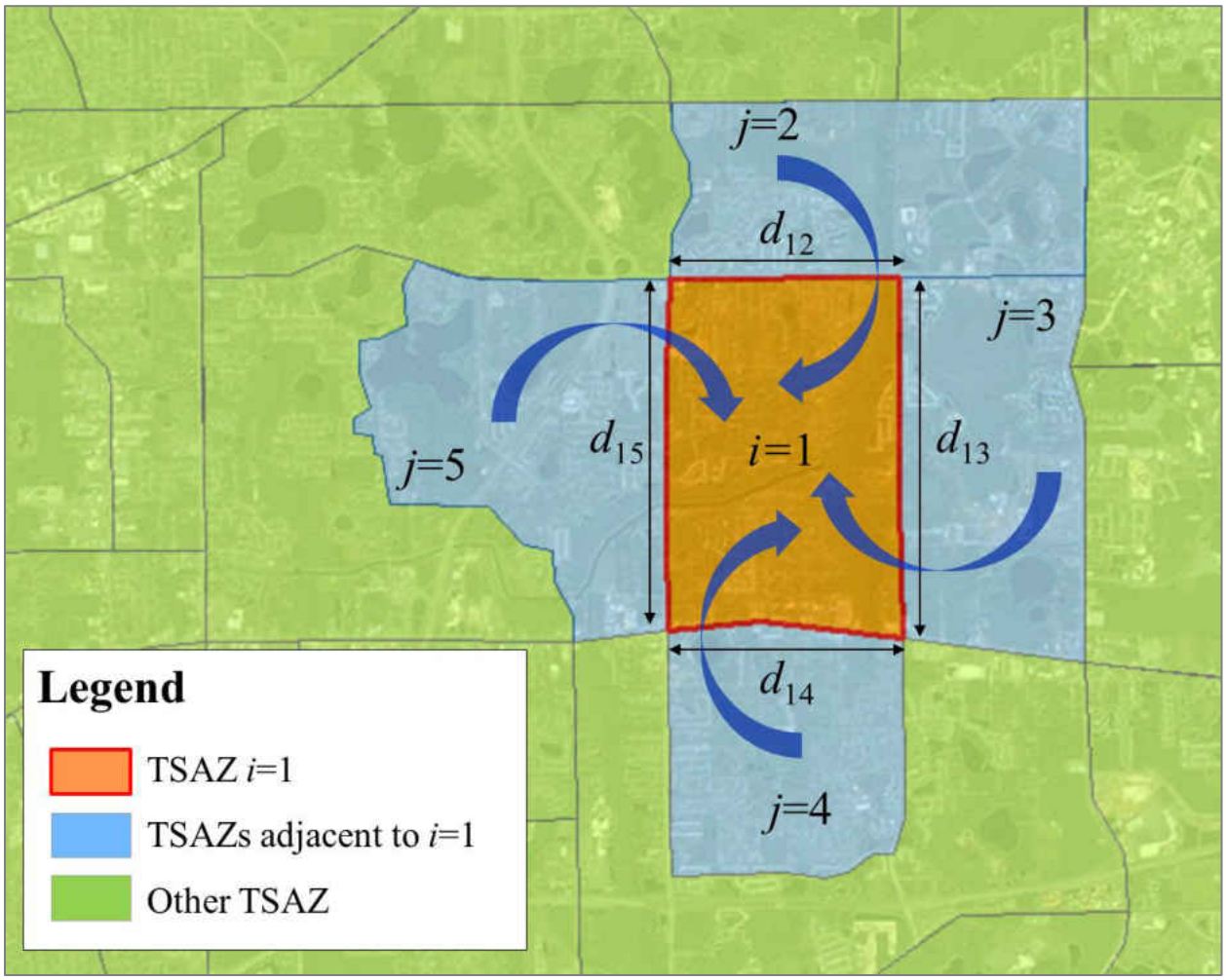


Figure 8-3: Illustration of adjacent zones for crash zone *i*

$$x_{ABCi} = wx_i + (1 - \mathrm{w})\left[(d_{i1}x_1 + d_{i2}x_2 + \cdots + d_{ik}x_k) \Big/ (d_{i1} + d_{i2} + \cdots + d_{ik})\right] \quad (35)$$

where,

x_{ABCi} = transformed variable *x* for i^{th} zone,

x_i = variable x for i^{th} zone,

x_j = variable x for the zones adjacent to i^{th} zone (j=1, 2, … k),

d_{ij} = length of the shared boundary between zones i and j,

$d_{i1} + d_{i2} + ... + d_{ik}$ = perimeter of zone i, and

w= weight to balance effects from the crash zone and its adjacent zones.

The first term in Equation (35) represents the characteristics of the i^{th} zone; the second term denotes the weight-averaged features of the adjacent zones. The weighted average is based on the length of the shared boundary between the adjacent zones.

The weight component (w) reflects the actual influence on boundary crashes from zone i and its adjacent zones. For instance, if the boundary crash was affected by the features of the crash zone (i^{th} zone) and the adjacent zones uniformly, the weight is 0.5. Meanwhile, if the boundary crash was solely influenced by the crash zone, the weight is 1.0.

Overall 120 Negative Binomial (NB) were developed with ten weights (from 1.0 to 0.1 by 0.1) for twelve sub-models (6 sub-models×2 crash types), in order to find out optimal weights. Although, the Poisson-lognormal model was adopted for the final model in this dissertation, the NB model was applied because there are too many models to estimate and NB model estimation takes less time and efforts compared to Poisson-lognormal model. Also, the results of NB model are expected to be comparable to Poisson-lognormal models from the experience. The models with the lowest Akaike Information Criterion (AIC) values were selected. The AIC was developed by Akaike (1974), and is calculated as follows:

$$AIC = 2k - 2\ln(L) \quad (36)$$

where,

k is the number of parameters in the model, and

L is the maximum likelihood for the model.

The AIC is an index which compares the relative qualities among various models; it is widely used for model selection. The AIC copes with the tradeoff between the goodness-of-fit and the complexity of the model. Among the candidate models, the model with the minimum AIC was chosen. Tables 8-1 and 8-2 present the AICs of candidate models for total and fatal-and-injury crashes, respectively.

The optimal weights for the FSB, OSB, and NSB are 0.7, 0.9, and 0.7 for the total crash model, and 0.7, 0.8, and 0.8 for the fatal-and-injury model. These optimal weights were used to estimate the final SPF. However, future studies could use 0.8 as an optimal weight for boundary crashes and 1.0 for interior crashes because no significant difference was observed between the models with weights equal to 0.7, 0.8, or 0.9.

Table 8-1: AIC table of candidate total crash models

Weights	1) FSB	2) FSI	3) OSB	4) OSI	5) NSB	6) NSI
1.0	1835.41	**1083.89**	3353.83	**1514.34**	4114.86	**4134.00**
0.9	1833.67	1084.92	**3351.47**	1516.03	4102.09	4137.71
0.8	1833.30	1087.79	3352.41	1520.01	4099.41	4150.41
0.7	**1833.23**	1093.23	3355.75	1526.46	**4099.20**	4170.31
0.6	1833.75	1102.28	3362.91	1535.29	4101.13	4199.80
0.5	1835.28	1115.46	3374.43	1545.71	4105.22	4236.48
0.4	1838.38	1132.19	3389.40	1556.25	4111.41	4273.68
0.3	1843.35	1150.42	3405.48	1565.51	4119.36	4306.42
0.2	1849.75	1167.25	3420.14	1572.85	4128.59	4334.83
0.1	1856.54	1180.64	3432.07	1578.46	4138.74	4361.82

Table 8-2: AIC table of candidate fatal-and-injury crash models

Weights	1) FSB	2) FSI	3) OSB	4) OSI	5) NSB	6) NSI
1.0	1373.05	**827.93**	2466.38	**1020.80**	2951.42	**2861.46**
0.9	1373.39	833.08	2463.98	1024.11	2947.82	2867.83
0.8	1372.71	837.80	**2462.60**	1027.26	**2947.23**	2882.72
0.7	**1372.33**	844.98	2462.81	1032.17	2947.50	2905.83
0.6	1372.59	855.25	2466.17	1038.96	2948.88	2937.01
0.5	1374.10	868.68	2473.68	1047.12	2951.56	2971.64
0.4	1377.64	884.32	2485.10	1055.58	2955.51	3002.94
0.3	1383.48	900.12	2498.58	1063.18	2960.42	3027.24
0.2	1390.60	913.69	2511.61	1069.26	2965.87	3045.49
0.1	1397.36	923.51	2522.59	1073.77	2971.52	3060.60

In summary, boundary crash types were greatly affected by the crash zone (70-90%) and rarely influenced by adjacent zones (10-30%). Moreover, it was proven that interior crashes were affected only by the characteristics of the zone wherein the crash occurred, because the optimal

weight for all interior crash models is 1.0. However, these optimal weights are solely applicable to crash modeling with TSAZ-based data. The optimal weights may be different if the model is developed based on different zone systems (i.e., census tract, traffic analysis district, block groups, traffic analysis zone, etc.).

8.3 Summary and Conclusion

In order to account for boundary crashes, two methods were suggested. First, a complex nested structure was constructed to estimated individual six sub-models for boundary and interior crashes by roadway types. It enables estimating boundary and interior crash models individually. Second, a variable transformation was proposed to relate boundary crashes with adjacent multiple zones. It allows developing boundary crash models with zonal factors from neighboring zones. The variable transformation formula includes a weight to balance effects from the crash zone and its adjacent zones. The optimal weights were calculated for six sub-models. Optimal weights for FSB, OSB, and NSB total crash models were 0.7, 0.9, and 0.7, respectively. It implies that boundary crash types were greatly affected by the crash zone (70-90%) and rarely influenced by adjacent zones (10-30%). On the other hand, optimal weights for all interior crash models were 1.0. It shows that interior crashes are influenced solely by the zone within which they are spatially located. It is expected that the model based on the complex nested structure with the optimal variable transformation performs better than other models without nested structures.

CHAPTER 9 MACROSCOPIC SAFETY MODELING

9.1 Model Specification

As shown in the previous chapter, NBPLSEM (Nested Bayesian Poisson-Lognormal Spatial Error Model) was proposed to separate boundary and interior crashes, and roadway types. Also the variable transformation was suggested to account for boundary crashes. Each sub-model of NBPLSEM was formulated as follows:

$$y_i \sim Poisson(\mu_i) \tag{37}$$

$$\lambda_i = \exp(\beta_0 + \beta X_i + \theta_i + \varphi_i) \tag{38}$$

$$\theta_i = Normal\ (0, \tau_\theta) \tag{39}$$

where,

y_i is the number of aggregated total crashes of the i^{th} TSAZ,

β_0 is the intercept,

β's are the coefficient estimates of covariates (X_i),

θ_i is the random effect term,

φ_i is the spatial effect term, and

τ_θ is the precision parameter, which is the inverse of the variance and a given prior gamma distribution (0.5, 0.005).

The Bayesian model was fit with non-informative prior distributions, Normal (0, 10^{-6}) for β's.

The spatial distribution was implemented by specifying an intrinsic Gaussian Conditional Autoregressive (CAR) prior with a $Normal\ (0, \tau_\varphi)$ distribution. The mean of φ_i is defined by

$$\bar{\varphi}_i = \sum_{i \neq j} \varphi_j \times w_{ij} \Big/ \sum_{i \neq j} w_{ij} \quad (40)$$

where, $w_{ij} = 1$, if zone i and j are adjacent, and

$w_{ij} = 0$, otherwise

w_{ij} was conceptualized by first order rook polygon contiguity since the spatial autocorrelation at TSAZ level can be effectively explained by adjacency.

9.2 Modeling Results

A series SPFs in the nested structure at the macroscopic level (i.e., NBPLSEM) was estimated. For the boundary crashes, the variable transformation was conducted using optimal weights. The modeling results for total and fatal-and-injury crashes are suggested in Tables 9-1 and 9-2, respectively.

As seen in these tables, each sub-model had different sample sizes. This is because some zones had zero probability of having specific types of crashes. For instance, one zone had no state roads. In this case, the expected numbers for the FSB, FSI, OSB, or OSI in this zone should be zero, regardless of zonal characteristics. It is not reasonable to include this type of zone in crash prediction models. Therefore, the zones without specific types of roads were taken out when the models for crashes that occurred on those types of roads were developed.

In addition, the total crash model and the fatal-and-injury model both show that each sub-model has its own variable set. All models seem to have reasonable and explainable coefficients.

Taking the FSB model as an example, the exposure variable (vehicle-miles-traveled) was positively associated with crash counts. Also, the coefficients of the proportion of young people (15-24 years old), the natural logarithm of employment and school enrollment, and the proportion of roads with 20 mph or lower max speed were positive. The first two variables are self-explanatory. The third variable, the proportion of roads with low speed limit, refers to the proportion of residential roads. It is interpreted that if a zone has higher proportion of residential roads, then there are more local drivers enter the freeway-and-expressway from these residential roads. Thus, the zone may have more crashes on freeway-and-expressway. Also, the results show that the spatial effects were significant. They reveal the existence of spatial autocorrelations among the explanatory variables with associated total and fatal-and-injury crashes.

Interestingly, Ψ (the apportionment of the variability in the error component due to spatial autocorrelation) is always larger in the boundary models than in the interior models of the same roadway type. For example, the Ψ of the FSB and FSI in the total crash model are 0.505 and 0.228, respectively. From this, it was concluded that the unobserved heterogeneity in the error component from the spatial effects in the FSB is 50.5%, whereas it is only 22.8% in the FSI. In other words, boundary crashes are more significantly influenced by spatial autocorrelation, because they are close to other adjacent zones.

Table 9-1: Nested Poisson Lognormal Spatial Error Model Accounting for Boundary Crashes: Total Crashes

Variable	1) FSB N=213		2) FSI N=155		3) OSB N=325		4) OSI N=174		5) NSB N=439		6) NSI N=465	
	mean	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**
Intercept	-0.482#	0.791	-1.690**	0.537	-0.602#	0.559	-2.538**	0.747	0.806**	0.485	-0.804**	0.243
Ln of population density			0.096**	0.043	0.122**	0.050					0.100**	0.020
Proportion African Americans											0.911**	0.243
Proportion of Hispanics					1.292**	0.552						
Proportion of young people (15-24 yr)	3.850**	1.645										
Proportion of old people (65yr+)												
Proportion households without car					3.098**	1.439						
Ln of hotel, motel and timeshare rooms	-0.057	0.040			0.092**	0.034	0.514**	0.091			0.049**	0.014
Ln of employment and school enrollment	0.240**	0.108	0.245**	0.066	0.227**	0.079			0.259**	0.064	0.326**	0.033
Proportion of roads with 20 mph or lower max speed	5.943**	1.673										
Proportion of roads with 55 mph or higher max speed												
Roads with poor pavement conditions											0.210**	0.047
Ln of VMT at FSB	0.097**	0.020										
Ln of VMT at FSI			0.173**	0.018								
Ln of VMT at OSB					0.137**	0.016						
Ln of VMT at OSI							0.145**	0.025				
Ln of VMT at NSB									0.051**	0.017		
Ln of VMT at NSI											0.094**	0.010
s.d. of θ_i	0.850	0.108	0.886	0.081	1.213	0.070	1.335	0.091	0.639	0.117	0.527	0.060
s.d. of φ_i	0.890	0.222	0.271	0.128	0.402	0.236	0.180	0.146	1.567	0.222	0.873	0.150
Ψ	0.505	0.087	0.228	0.090	0.236	0.109	0.113	0.076	0.707	0.065	0.620	0.065
DIC	1342.46		846.222		2349.98		1085.5		3007.29		3214.22	

not significant at 20%, ** significant at 5%, * significant at 10%, and all other variables are significant at 20%

Table 9-2: Nested Poisson Lognormal Spatial Error Model Accounting for Boundary Crashes: Fatal-and-Injury Crashes

Variable	**1) FSB N=213**		**2) FSI N=155**		**3) OSB N=325**		**4) OSI N=174**		**5) NSB N=439**		**6) NSI N=465**	
	mean	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**	**mean**	**s.d.**
Intercept	-3.153**	1.043	-4.240**	0.800	-2.525**	1.046	-10.260**	2.632	-0.137#	0.451	-1.796**	0.286
Ln of population density					0.133**	0.067					0.063**	0.022
Proportion African Americans									1.279**	0.425	0.925**	0.261
Proportion of Hispanics					2.146**	0.788						
Proportion of young people (15-24 yr)	4.813**	2.129							-1.558*	0.908		
Proportion of old people (65yr+)												
Proportion households without car					5.968**	2.002						
Ln of hotel, motel and timeshare rooms	-0.091*	0.050	-0.057	0.037	0.113**	0.039	0.132**	0.059			0.042**	0.015
Ln of employment and school enrollment	0.244**	0.121	0.506**	0.104			0.234	0.153	0.226**	0.061	0.282**	0.037
Proportion of roads with 20 mph or lower max speed	3.968*	2.098							-2.690**	1.264		
Proportion of roads with 55 mph or higher max speed					-2.881**	1.165	-5.308**	2.031				
Roads with poor pavement conditions									0.109	0.078	0.193**	0.049
Ln of VMT at FSB	0.166**	0.026										
Ln of VMT at FSI			0.203**	0.022								
Ln of VMT at OSB					0.179**	0.022						
Ln of VMT at OSI							0.271**	0.038				
Ln of VMT at NSB									0.041**	0.015		
Ln of VMT at NSI											0.112**	0.011
s.d. of θ_i	1.173	0.105	0.903	0.108	1.364	0.135	1.558	0.145	0.401	0.183	0.480	0.087
s.d. of φ_i	0.614	0.188	0.303	0.169	0.705	0.451	0.236	0.208	1.789	0.225	0.897	0.180
Ψ	0.339	0.079	0.243	0.112	0.315	0.156	0.124	0.089	0.816	0.087	0.646	0.084
DIC	1037.090		672.768		1788.360		716.974		2337.760		2450.26	

not significant at 20%, ** significant at 5%, * significant at 10%, and all other variables are significant at 20%

9.3 Model Comparison

Although NBPLSEM was suggested as the best model for the macroscopic safety analysis, it is still necessary to compare it with other models with different nested structures or weights. Overall, five different models with different structures were compared. Structure 1 is an aggregated single model without the nested structure (Figure 9-1).

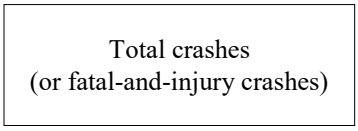


Figure 9-1: Structure 1-aggregated single model (1 sub-model)

Structure 2 separates boundary and interior crashes and estimates two sub-models as seen in Figure 9-2. However, it does not split crashes by roadway types.

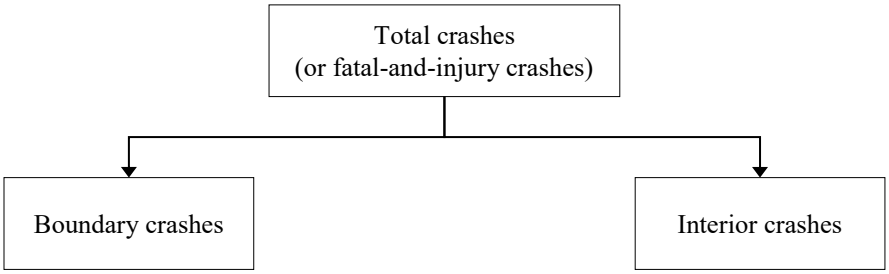


Figure 9-2: Structure 2-boundary and interior models (2 sub-models)

Structure 3 has 3 sub-models, it splits crashes by roadway types but it does not separate boundary and interior crashes (Figure 9-3).

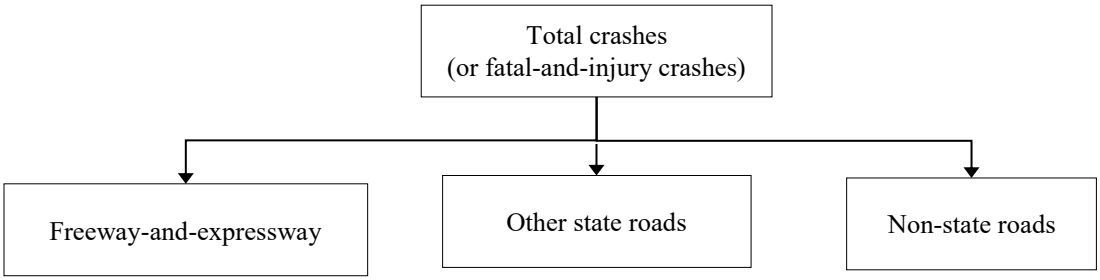


Figure 9-3: Structure 3-models by roadway types (3 sub-models)

As presented in Figure 9-4, Structure 4 has the most complicated nested structure. It splits crashes by roadway types and also separates boundary and interior crashes. Basically, Structures 4A and 4B has the same nested composition. Nevertheless, Structure 4A uses variables without variables transformation, in other words, weights for the boundary variable transformation are zero. On the other hand, Structure 4B applies transformed variables with optimal weights for boundary crashes as explained in the previous chapter.

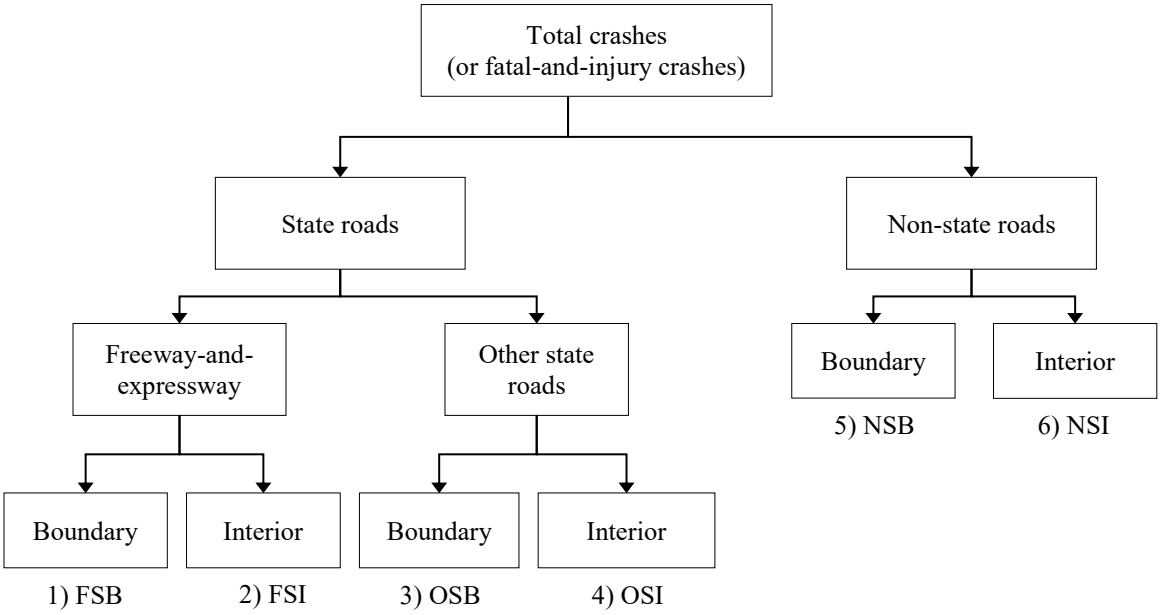


Figure 9-4: Structure 4-models by roadway types, and boundary and interior crashes (6 sub-models)

For the structure comparison, four goodness-of-fit measure, MAD, RMSE, PMAD and R^2_{FT}, were used. MAD (Mean Absolute Deviation) calculates sum of absolute deviations divided by the number of observations as follows:

$$MAD = \sum_{i=1}^{N} |y[i] - \hat{y}[i]| / N \quad (41)$$

where,

N=number of observations, and

$y[i]$and $\hat{y}[i]$are observed and predicted values for i, respectively.

RMSE (Root Mean Squared Errors) computes the square root of the sum of the squared errors divided by the number of observations as the following formula:

$$RMSE = \sqrt{\sum_{i=1}^{N}(y[i] - \hat{y}[i])^2 / N} \quad (42)$$

SAD (Sum of Absolute Deviation) calculates sum of absolute deviations as follows:

$$SAD = \sum_{i=1}^{N}|y[i] - \hat{y}[i]| \quad (43)$$

PMAD (Percent Mean Absolute Deviation) computes sum of absolute deviations divided by the sum of absolute observed values.

$$PMAD = \sum_{i=1}^{N}|y[i] - \hat{y}[i]| \Big/ \sum_{i=1}^{N}|y[i]| \quad (44)$$

Lastly, R_{FT}^2 was suggested to approximate the Poisson to a normal distribution by Freeman and Tukey (1950) (Fridstrom et al., 1994). The variance stabilizing transformation of a Poisson variable $y_i$with mean $\lambda_i$was suggested as follows:

$$f_i = \sqrt{y_i} + \sqrt{y_i + 1} \quad (45)$$

This statistics is approximately normally distribute with mean ϕ_i and unit variance e_i

$$\phi_i = \sqrt{4\lambda_i + 1} \quad (46)$$

$$e_i = f_i - \phi_i \quad (47)$$

The Freeman-Tukey deviates can be estimated by the corresponding residuals as follows:

$$\widehat{e_i} = \sqrt{y_i} + \sqrt{y_i + 1} - \sqrt{4\widehat{y_i} + 1} \tag{48}$$

An R^2goodness-of-fit measure for the Freeman-Tukey transformed variable (R^2_{FT}) is

$$R^2_{FT} = 1 - {\sum_i \widehat{e_i}^2}\Big/{\sum_i (f_i - \bar{f})^2} \tag{49}$$

The goodness-of-fit measures by structures are shown in Table 9-3. Structure 4B, the nested structure splitting boundary and interior crashes by roadway types with the variable transformation performs the best, in terms of MAD, PMAD and R^2_{FT}. Only RMSE is slightly smaller in Structure 4A but it is thought that overall goodness-of-fit measures are superior in Structure 4B.

Table 9-3: Comparison of goodness-of-fit measure by structures

Nested structure	**Description**	**MAD**	**RMSE**	**PMAD**	$\boldsymbol{R^2_{FT}}$
Structure 1	Aggregated single model *1 sub-model*	96.565	156.239	0.785	0.480
Structure 2	Interior & Boundary *2 sub-models*	90.141	165.232	0.504	0.539
Structure 3	By roadway types *3 sub-models*	96.201	173.264	0.537	0.467
Structure 4A	By roadways and I&B *6 sub-models* (w/o variable transformation)	83.889	154.461	0.469	0.605
Structure 4B	**By roadways and I&B** ***6 sub-models*** **(w/ variable transformation)**	**83.012**	**154.720**	**0.464**	**0.613**

9.4 Summary and Conclusion

In this chapter, NBPLSEM was adopted to estimate both total and fatal-and-injury crashes at the macroscopic level. NBPLSEM contains a spatial error term based on first rook contiguity that controls for spatial autocorrelations. NBPLSEM with six sub-models was estimated and the result shows that each sub-model has different significant variable sets. It also justifies that applying the nested structure is reasonable. Several goodness-of-fit measures such as MAD, RMSE, PMAD, and R_{FT}^2 were used for the comparison of models with different nested structures. The result revealed that the most complicated nested structure with the variable transformation outperforms all other models, despite of the complex structure.

CHAPTER 10 MACROSCOPIC AND MICROSCOPIC SCREENING

10.1 Performance Measure for the Screening

In the previous chapter, macroscopic safety models for total and fatal-and-injury crashes in the nested structure were estimated. In this chapter, hot zones in three counties in Central Florida (Orange, Seminole, and Osceola Counties) are identified using the estimated models. PSI (Potential for Safety Improvements) was chosen as a performance measure for both macroscopic and microscopic screening procedures. PSI, or excess crash frequency, is a measure of how many crashes can be effectively reduced. The PSI for each zone is the difference between the expected crash count and the predicted crash count as displayed in Figure 10-1.

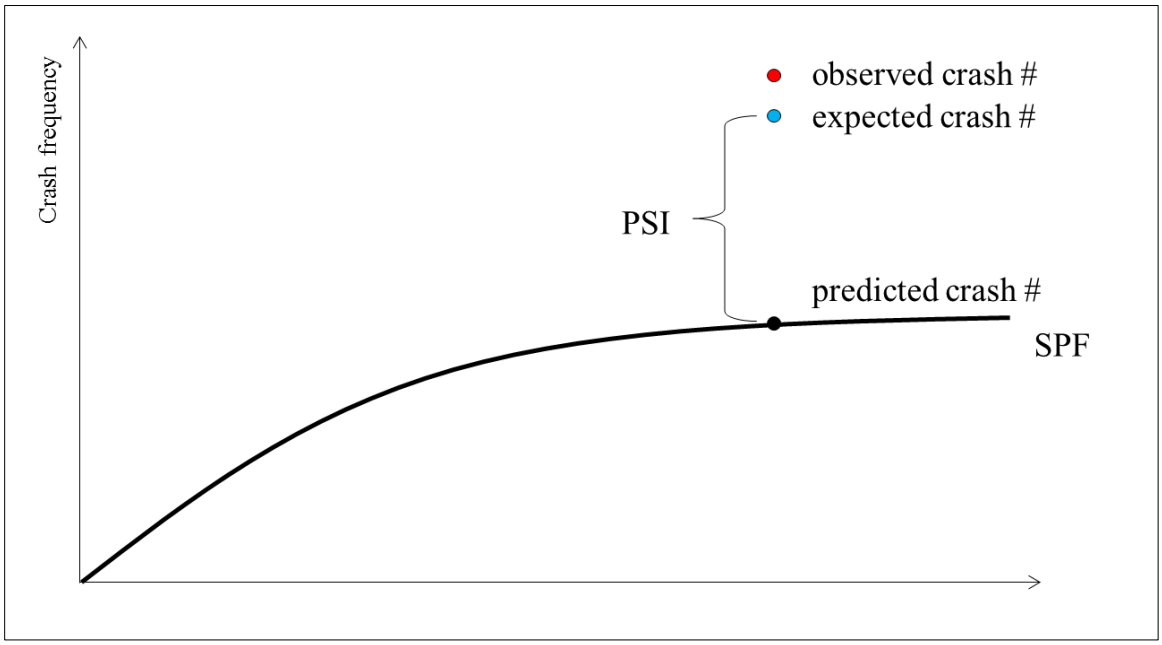


Figure 10-1: Schematic showing definition of PSI

Predicted crash counts were estimated using six sub-models in Structure 4B, as shown earlier, and the PSIs were calculated by following equations proposed by Aguero-

Valverde and Jovanis (2009). As suggested in Equation (50), the PSI is the gap between the expected crash counts and the predicted crash counts. Equation (51) and (52) were derived from Equation (50), for convenience of calculation.

$$PSI = N_{expected} - N_{prdeicted} \quad (50)$$

$$= \exp(\beta_0 + \beta X_i + \theta_i + \varphi_i) - \exp(\beta_0 + \beta X_i) \quad (51)$$

$$= \exp(\beta_0 + \beta X_i)(\exp(\theta_i + \varphi_i) - 1) \quad (52)$$

10.2 Macroscopic Screening

The PSIs at the macroscopic level were calculated and the TSAZs were ranked separately for the urban and rural area. Tables 10-1 and 10-2 present the TSAZ with the top 10% PSIs in rural and urban areas, correspondingly.

Table 10-1 Ranking TSAZ with the top 10% PSIs (rural areas)

Rank	Rank percentile	Total crashes		Fatal-and-injury crashes	
		TSAZ ID	PSI	TSAZ ID	PSI
1	1.4%	367	215.548	367	79.229
2	2.8%	337	152.669	337	70.096
3	4.2%	347	145.548	347	51.083
4	5.6%	406	130.475	281	48.928
5	6.9%	281	118.346	406	45.225
6	8.3%	49	103.374	464	31.660
7	9.7%	361	70.069	49	31.319
8	11.1%	247	61.156	394	26.761

Table 10-2 Ranking TSAZ with the top 10% PSI (urban area)

Rank	Rank percentile	Total crash		Fatal-and-injury crash	
		TSAZ ID	PSI	TSAZ ID	PSI
1	0.2%	56	1127.880	202	334.644
2	0.5%	15	971.440	8	272.738
3	0.7%	202	791.730	196	255.596
4	0.9%	8	651.180	2	234.250
5	1.2%	9	648.000	56	233.255
6	1.4%	196	625.459	15	204.740
7	1.6%	192	620.349	89	188.698
8	1.9%	89	595.207	207	179.557
9	2.1%	69	549.530	5	178.469
10	2.3%	104	510.150	43	175.275
11	2.6%	382	498.175	69	171.608
12	2.8%	130	492.320	3	156.202
13	3.0%	224	470.914	12	151.874
14	3.3%	0	433.720	192	150.154
15	3.5%	92	429.485	67	138.363
16	3.7%	67	428.796	62	137.494
17	4.0%	62	413.550	130	134.979
18	4.2%	6	411.870	18	133.330
19	4.4%	43	402.370	104	131.018
20	4.7%	66	385.870	9	129.910
21	4.9%	146	384.350	0	125.090
22	5.1%	178	381.803	66	124.134
23	5.4%	18	376.160	58	118.026
24	5.6%	42	361.726	101	111.759
25	5.8%	212	354.540	65	111.366
26	6.1%	195	350.338	93	110.636
27	6.3%	29	345.127	212	110.133
28	6.5%	35	330.897	16	109.178
29	6.8%	180	327.315	180	104.254
30	7.0%	19	318.380	86	96.362
31	7.2%	207	318.163	57	96.124
32	7.5%	60	302.947	6	94.408
33	7.7%	14	293.278	224	94.395
34	7.9%	2	287.020	38	91.203
35	8.2%	28	280.342	105	87.838
36	8.4%	57	268.780	382	87.799
37	8.6%	3	257.644	195	86.882
38	8.9%	250	253.911	250	82.205
39	9.1%	98	252.000	19	80.906
40	9.3%	5	250.610	233	79.481
41	9.6%	38	248.007	345	79.408
42	9.8%	22	247.428	42	78.342
43	10.0%	93	235.027	333	76.341

Figures 10-2 and 10-3 show the spatial distributions of hot zones with top 10% PSIs in Orange, Seminole, and Osceola counties for both total crashes and fatal-and-injury crashes, respectively. As for total crashes, Figure 10-2 indicates that most of the hot zones in the rural areas were close to the fringe of the urban areas, or they contained major arterials (i.e., SR50) or full access control roads (i.e., SR528, SR91, etc.). With regards to urban areas, the hot zones mostly were located in downtown Orlando to eastern Orlando along SR50 (E. Colonial Drive). As for fatal-and-injury crashes, the hot zones had patterns very close to those of the total crash hot zones (slight differences are indicated in Figure 10-3). These hot zones were closer to high speed roads (freeways or expressways) in both urban and rural areas. For example, some fatal-and-injury crash hot zones contained I-4 (urban) and SR91 (rural).

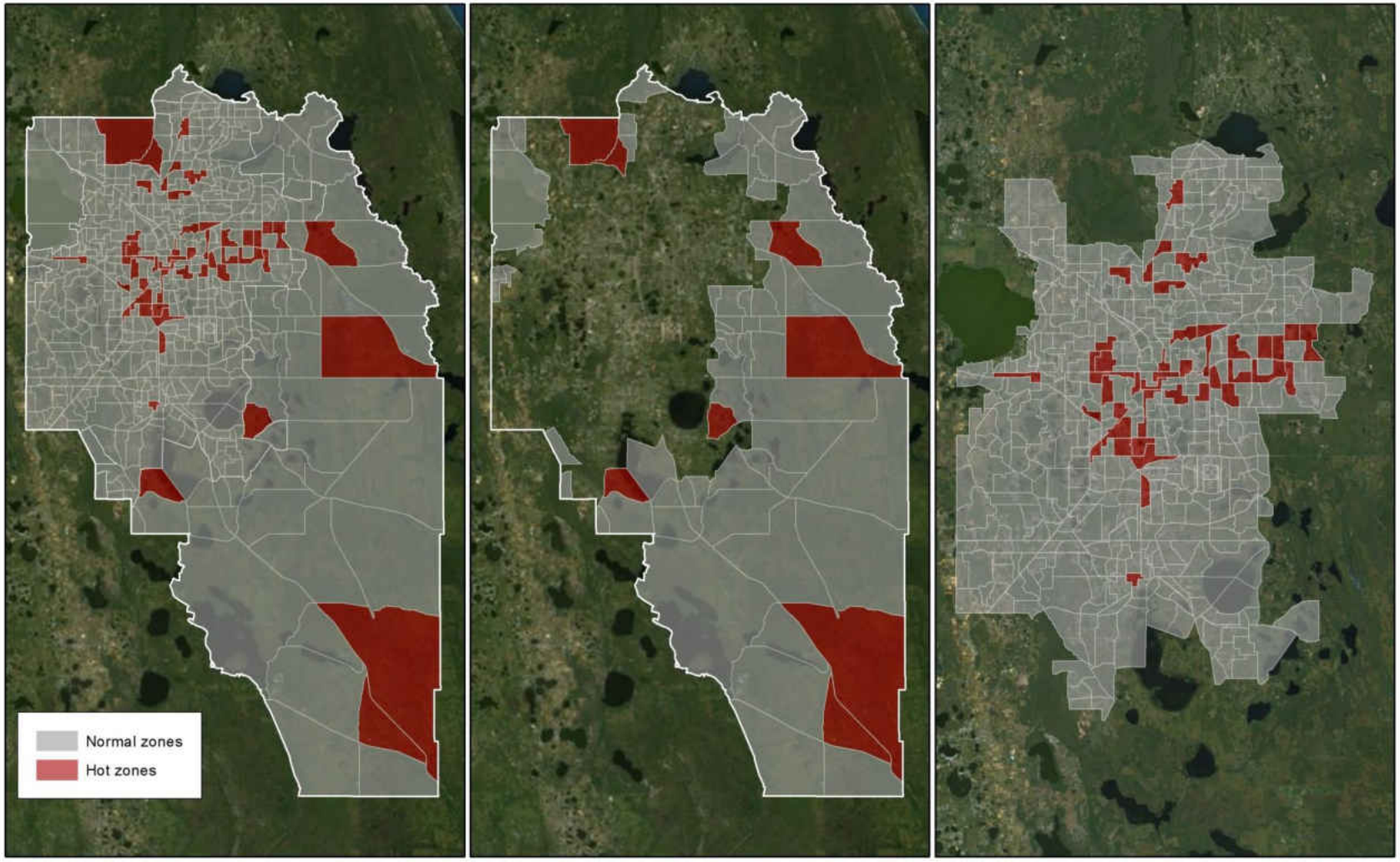


Figure 10-2: Top 10% hot zones for total crashes in both urban and rural areas, rural area, and urban area (left to right respectively)

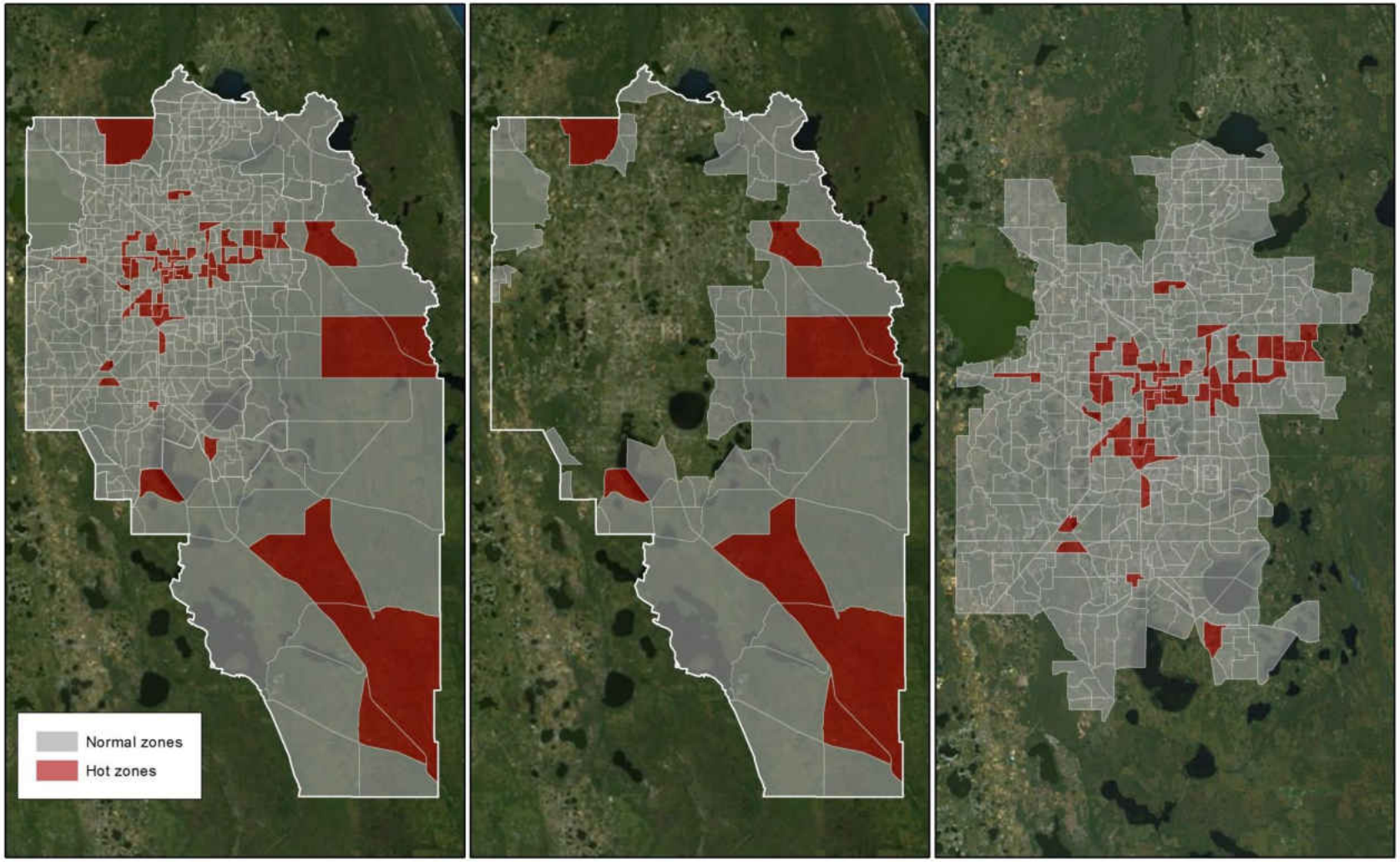


Figure 10-3: Top 10% hot zones for fatal-and-injury crashes in both urban and rural areas, rural area, and urban area (left to right respectively)

10.3 Microscopic Screening

As in the macroscopic screening process, the microscopic screenings were conducted using PSIs. Intersections and segments which have AADT (Average Annual Daily Traffic) information in Orange, Seminole, and Osceola Counties were screened using estimated own SPFs. The total crash and fatal-and-injury screening maps for intersections are presented in Figures 10-4 and 10-5, respectively. The intersection screening maps cover only urban area because there are no major intersections (i.e., intersections of major roads) in the rural area. Besides, the total crash and fatal-and-injury screening maps for segments are displayed in Figures 10-6 and 10-7, respectively.

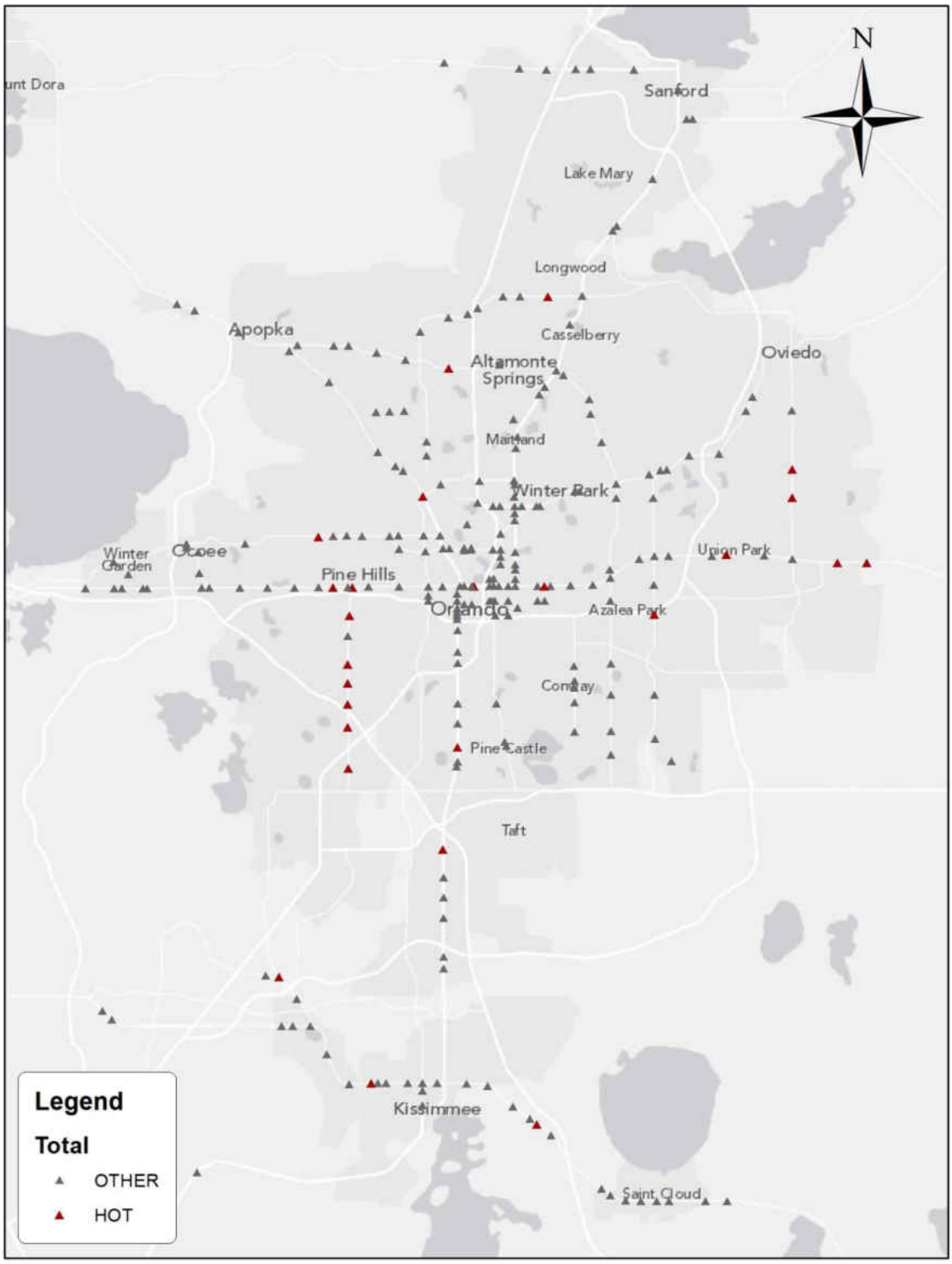


Figure 10-4: Intersection screening map for total crashes

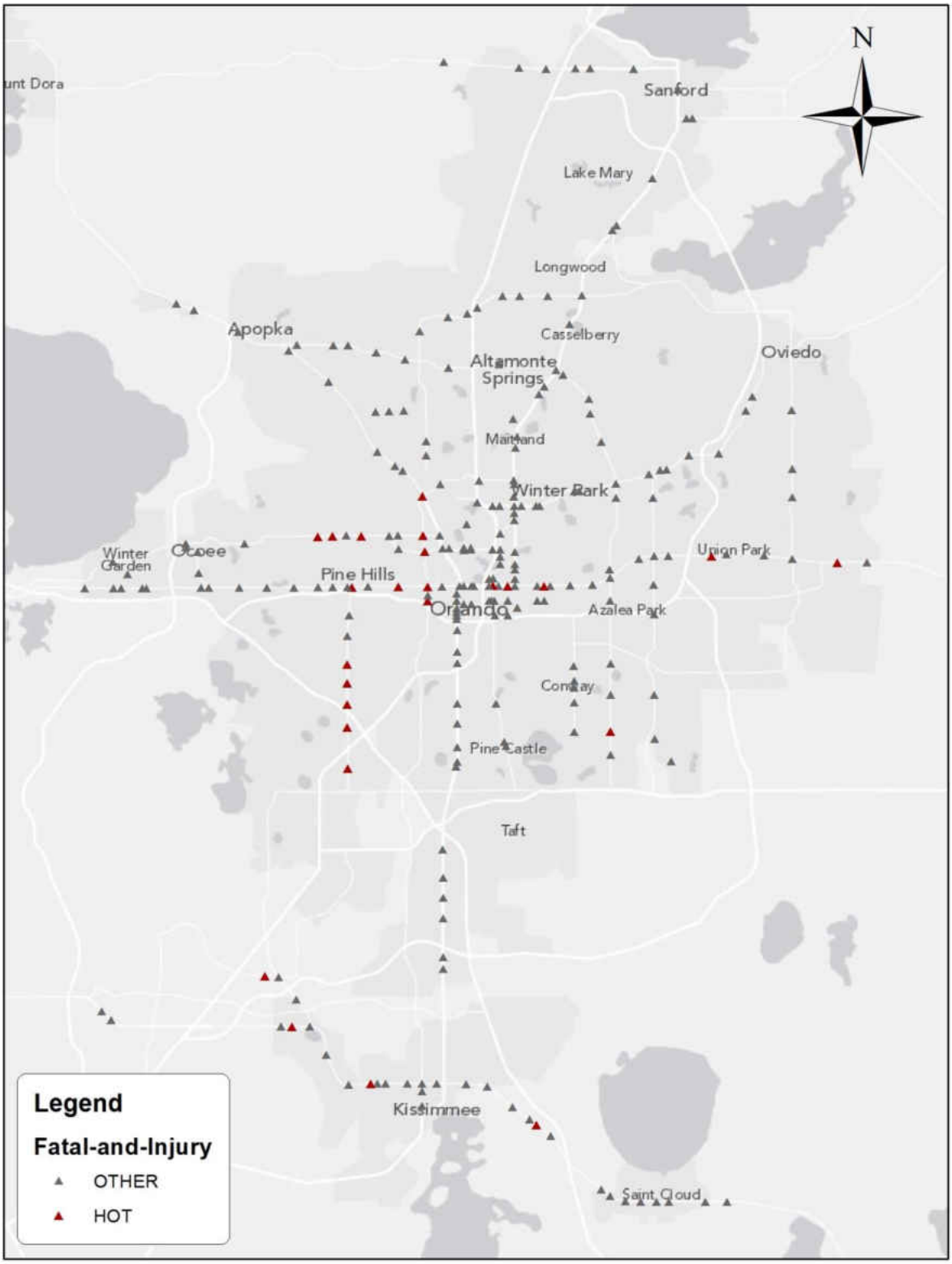


Figure 10-5: Intersection screening map for fatal-and-injury crashes

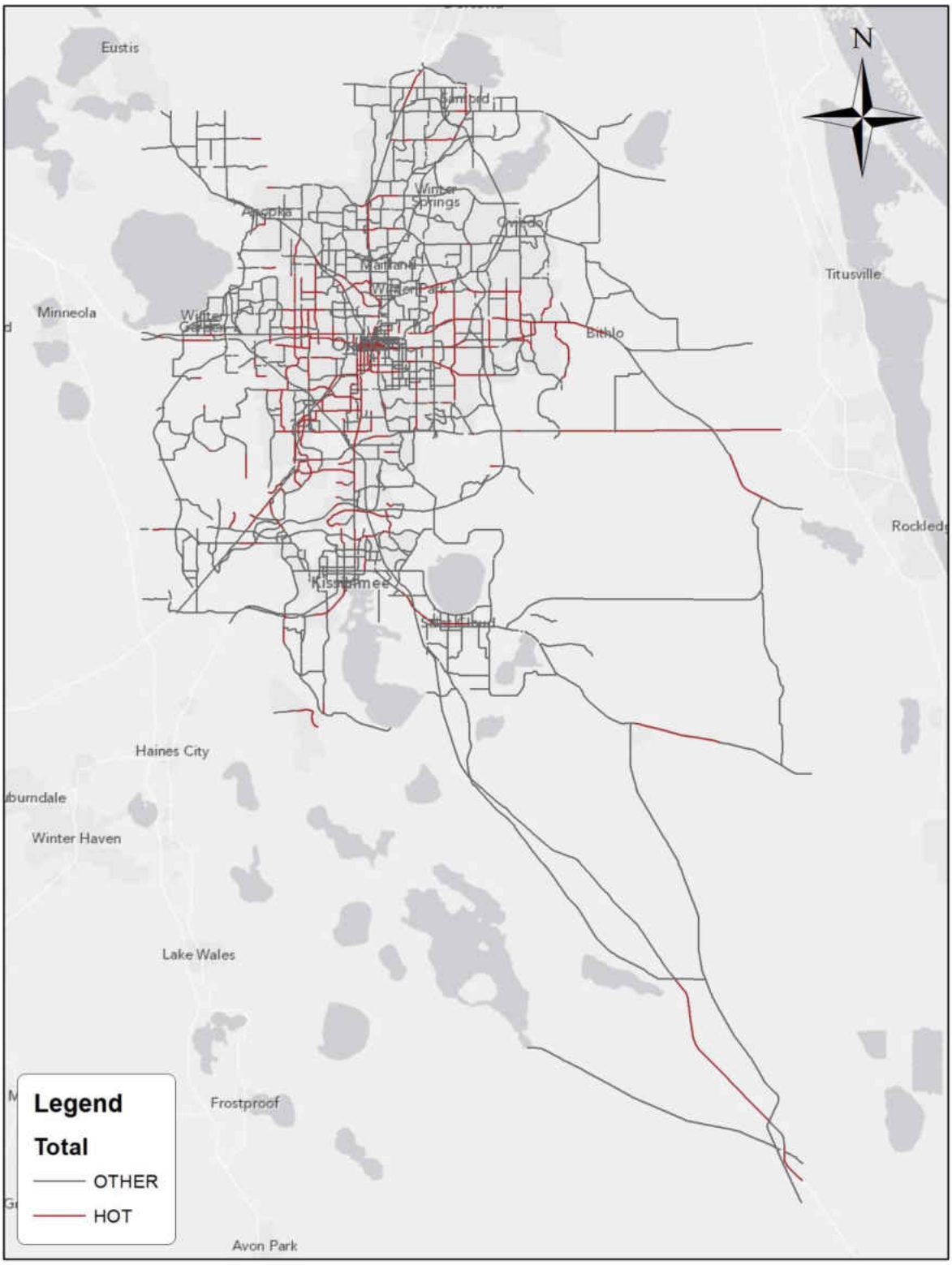


Figure 10-6: Segment screening map for total crashes

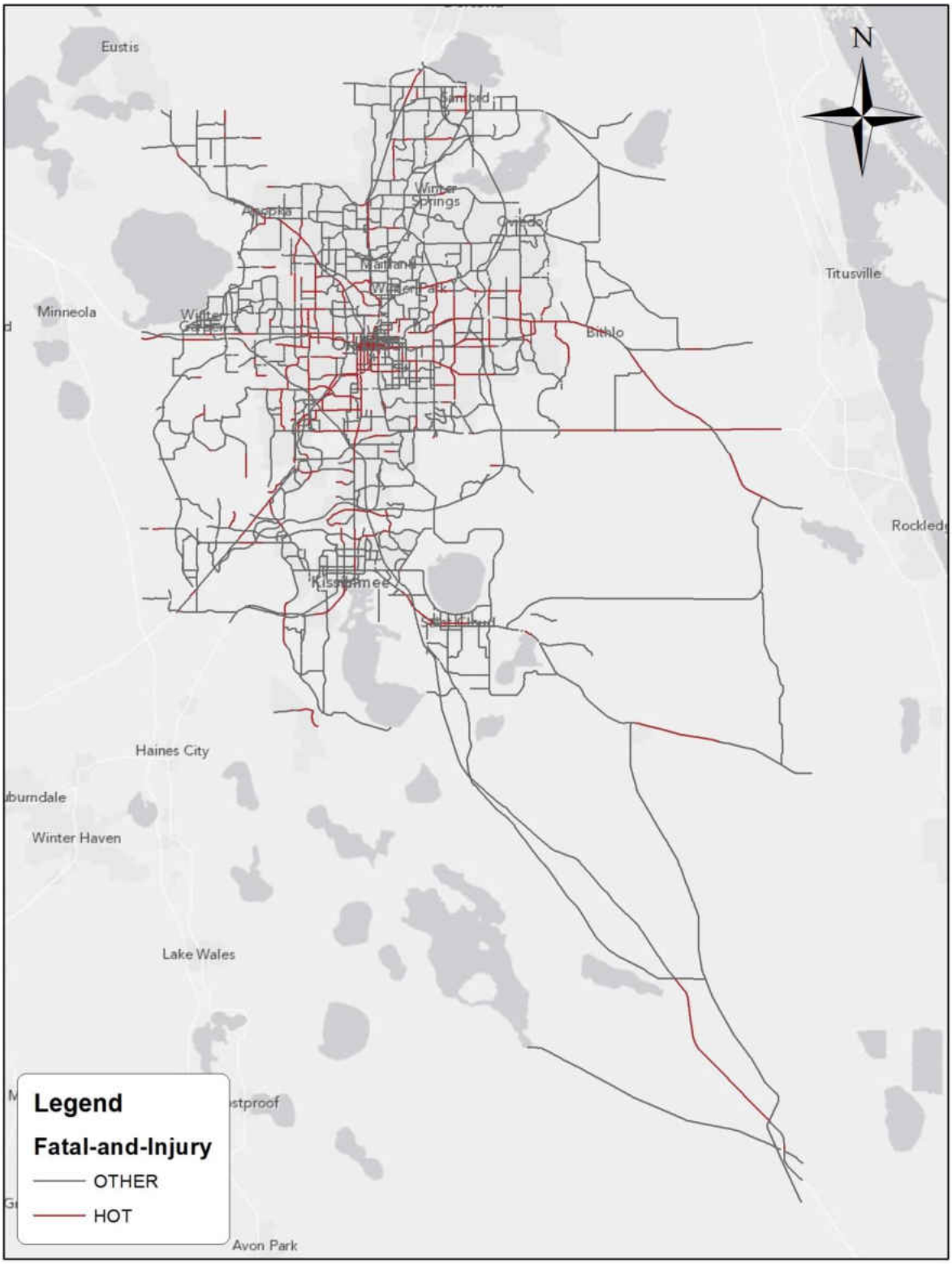


Figure 10-7: Segment screening map for fatal-and-injury crashes

10.4 Summary and Conclusion

In this chapter, hot zones/spots in Orange, Seminole, and Osceola Counties were identified both at the macroscopic and microscopic levels. PSI, which is a measure of how many crashes can be effectively reduced, was chosen as a performance measure for the screening. Each TSAZ was ranked for both total and fatal-and-injury crashes. It was shown that most of the hot zones in the rural areas were close to the fringe of the urban areas, or they contained major arterials or full access control roads. With regards to urban areas, the hot zones mostly were located in downtown Orlando to eastern Orlando along SR50. As for fatal-and-injury crashes, the hot zones had patterns very close to those of the total crash hot zones but the hot zones for fatal-and-injury crashes were closer to high speed roads in both urban and rural areas The microscopic screenings were also conducted for roadway segments and intersections which have AADT information for both total and fatal-and-injury crashes. The screening results from this chapter will be used for the integrated screening in Chapter 11.

CHAPTER 11 INTEGRATED MACROSCOPIC AND MICROSCOPIC SAFETY DATA ANALYTICS

11.1 Introduction

In the previous chapters, hot zones and hot spots at the macro- and micro-level were identified, respectively. In this chapter, both macroscopic and microscopic screening results are combined to provide a comprehensive, strategic, and effective traffic safety improvement planning. The integration strategy to integrate these two-level screening results is suggested.

11.2 Integrated Screening Process

Numerous studies have been done to analyze locations/sites with high traffic safety risks at the microscopic level, including the HSM Part B (Hauer, 1996; Heydecker et al., 1991; Kononov et al., 2003; Chung et al., 2007; Ragland et al., 2007; HSM, 2010). Recently, several studies started to focus on the zonal screening at the macroscopic level (Abdel-Aty et al., 20313; Pirdavani et al., 2013). In comparison with the microscopic analysis, macroscopic analysis can use zonal-level socio-demographic features into crash prediction models and identify hot zones. Thus, it allows transportation planners/traffic safety engineers to identify and predict area-wide traffic safety issues at the macroscopic level and provide appropriate policy-based treatments.

Nevertheless, the macroscopic analysis may neglect specific engineering problems from roadway segments and/or intersections. Therefore, it is necessary to develop a new integrated screening approach to overcome the shortcomings of both macroscopic and microscopic level screening methods.

In order to identify whether a zone has safety issues at macro-level and/or micro-level, all TSAZs are classified into twelve categories: two scale group (macro or micro) and four safety levels (hot, normal, cold, and no data) They are HH, HN, HC, HO, NH, NN, NC, NO,CH, CN, CC, and CO (Table 11-1). The former character of the classification represents a macroscopic safety risk, and the latter character symbolizes the microscopic safety risk. Thus, HH zones have both macro- and micro-level safety problems, and HN zones are risky at the macro-level but micro-level safety is moderate. Also, HC zones have safety problems only at the macroscopic level. NH zones have moderate crash risks at the zonal-level, but their microscopic crash risk is quite high. NN zones are intermediate for traffic safety both at macro- and micro-level. Likewise, NC zones have moderate risks at the macro-level but their safety risk at the micro-level is low. In the case of CH zones, they have high safety risks only at the microscopic level such as intersections and segments, while CN zones have low crash risks at the macroscopic level but intermediate crash risks at the microscopic level. CC zones are safe at both macro- and micro-level. HO, NO, and CO zones are dangerous, moderate, and safe, respectively, at the macro-level but they do not have segments or intersections data at all.

Table 11-1: Hot zone classification

	Scope	**Micro-level**			
Scope	**Category**	**Hot**	**Normal**	**Cold**	**No Data**
Macro-level	**Hot**	HH	HN	HC	HO
	Normal	NH	NN	NC	NO
	Cold	CH	CN	CC	CO

The integration process was conducted using macroscopic and microscopic screening results from the previous chapter (Figure 11-1). The overall integration process is presented in Figure 11-2. At the macro-level, TSAZs were ranked by their zonal PSIs, and TSAZs with top 10% macro-level PSIs were classified as "Hot" zones. While TSAZs with bottom 10% zonal PSIs were classified as "Cold" zones, and other TSAZs which were neither "Hot" nor "Cold" were categorized as "Normal". Likewise, at the micro-level, the calculation of average PSI is more complicated because each TSAZ has several intersections and segments. PSIs of intersections in each TSAZ were averaged by the number of intersections, and zones were ranked by their averaged intersection PSI. Simultaneously, PSIs of segments in each zone were averaged by the total length of segments in the zone, and zones were ranked by their averaged segment PSI. After that, both intersection and segment PSI ranks were averaged and TSAZs were ranked by the final averaged intersection and segment PSIs. Same as at the macro-level, TSAZs with top 10% micro-level PSIs were categorized as "Hot" zone at the microscopic level. Finally, TSAZs were classified into twelve categories based on macro- and micro-level screening results. It should be noted that we used the total length of segments to normalize the segment PSIs because the length of the segments may vary. Also, the percentile rank of PSIs was used in the integration (instead of the original PSIs) since the units of PSIs of intersection and segments are different.

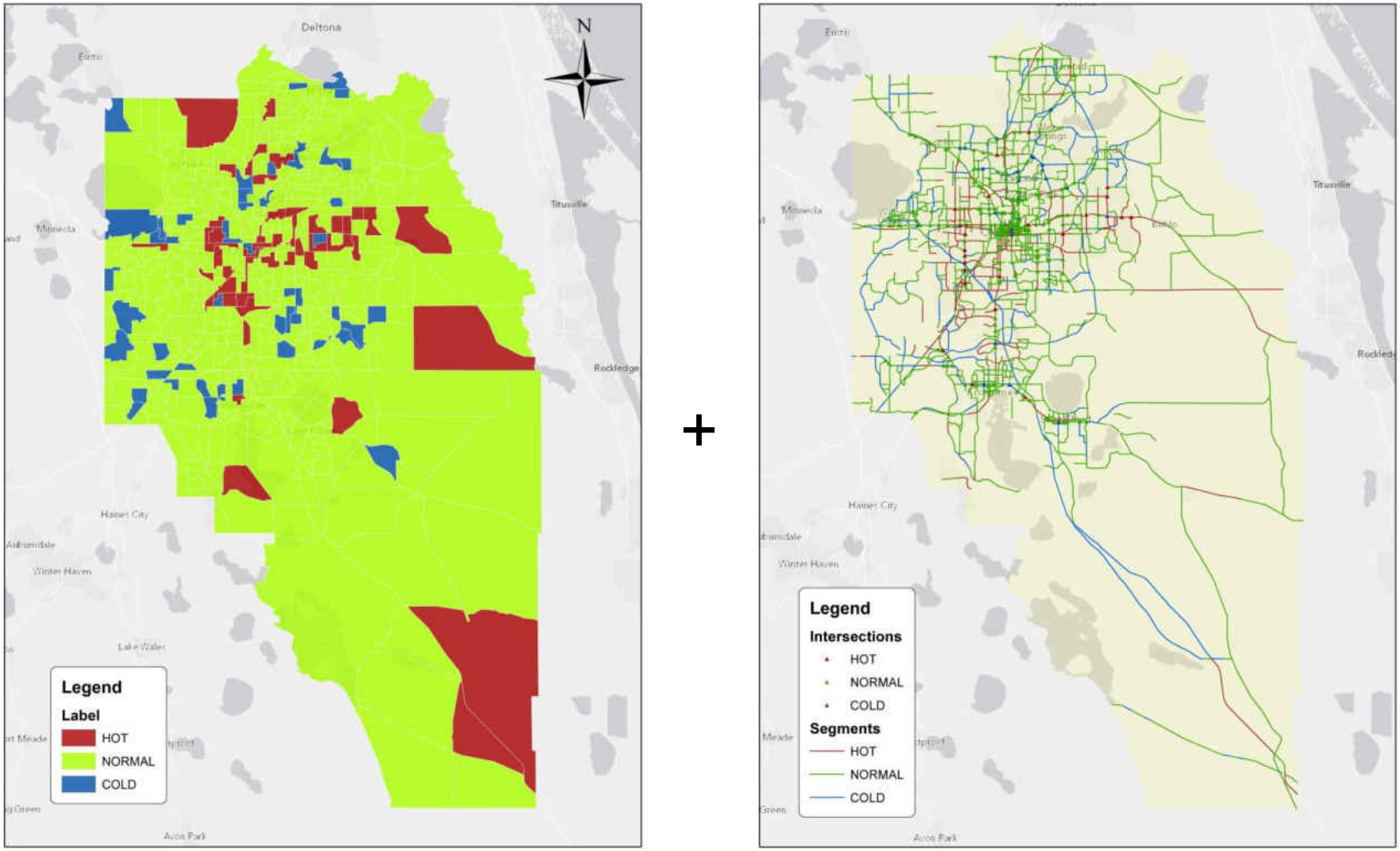


Figure 11-1: Results of macroscopic hot zone screening (left) and microscopic hot spot screening (right)

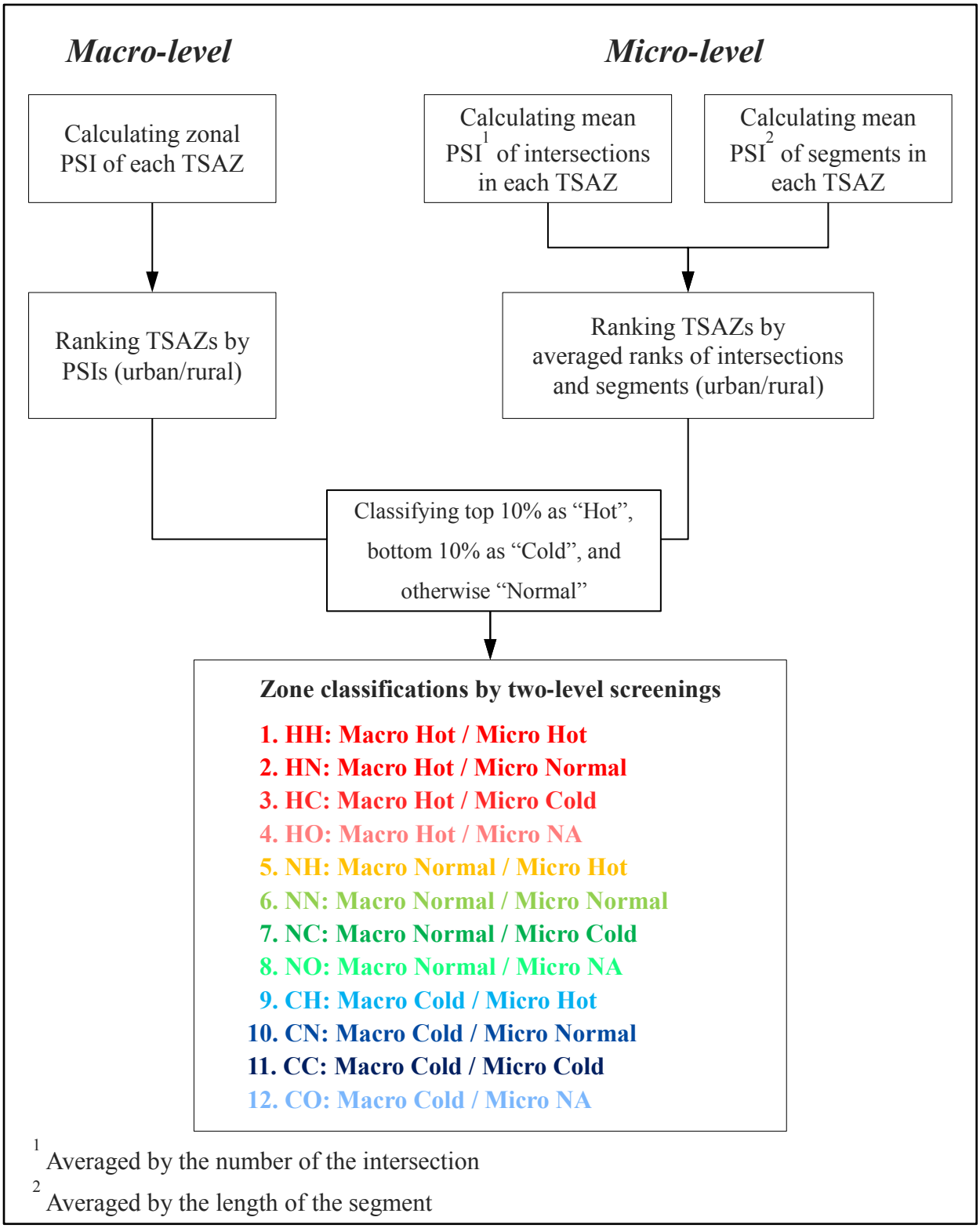


Figure 11-2: Integration process

11.3 Integrated Screening Results

11.3.1 Total crash

Table 11-2 shows the number of zones by hot zone classification for total crashes. Overall 26 HH zones are identified, which are the top priority for safety treatments since they have a higher crash risk at both macroscopic and microscopic levels. Moreover, there are 20 HN zones and 21 NH zones, which are the next priority for the treatment. HN zones have serious safety problems at the macro-level and intermediate risks at the micro-level whereas NH zones have high traffic crash risks at micro-level and intermediate risks at macro-level. Moreover, it is necessary to pay attention to HC and CH zones. Both HC and CH zones have contradicting hot zone identification at different levels. There are three HC zones, which are exceedingly risky only at the macroscopic level but safe at the microscopic level. Overall two CH zones were identified, which are very dangerous at micro-level but safe at macro-level. Besides, eight zones were identified as CC, which means they are safe both at macroscopic and microscopic levels. There is no significant difference in the hot zone identification between the urban and rural area, except the NO zones. NO/CO zones, which have no micro-level components, have a higher percentage in the rural area (25%) than in the urban area (7%). It is because the density of major roadway network in the rural area is much lower than those in the urban area, even though zones in the rural area are much larger.

Figure 11-4 presents the spatial distribution of TSAZs by hot zone classification for total crashes in the urban area. It was observed that many HH/HN/NH zones are located along State Road 50 (Colonial Drive), State Road 435 (Kirkman Road), State Road 408 (East-West Expressway), US

Route 17/92/441 (Orange Blossom Trail), and Interstate 4. Concerning HH zones, there are two large clusters containing multiple HH zones. The first HH cluster is located in the center of the map, which is adjacent to Interstate 4, State Road 435 (Kirkman Road), and US 17/92/441 (Orange Blossom Trail). The second cluster is in East Orlando area. It is shown that several principle arterial roadways such as State Road 408 and State Road 50 cross the second HH cluster. On the other hand, it seems that CC zones do not form any clusters. Some CC zones are located in the downtown area whereas some other zones are located in the suburban area.

Table 11-2: Number of zones by hot zone classification (total crash)

Classification	Urban		Rural		Sum	
	Zones	%	Zones	%	Zones	%
HH	22	5.1%	4	5.6%	26	5.2%
HN	18	4.2%	2	2.8%	20	4.0%
HC	2	0.5%	1	1.4%	3	0.6%
HO	0	0.0%	0	0.0%	0	0.0%
NH	19	4.4%	2	2.8%	21	4.2%
NN	261	61.0%	32	44.4%	293	58.6%
NC	34	7.9%	6	8.3%	40	8.0%
NO	29	6.8%	17	23.6%	46	9.2%
CH	1	0.2%	1	1.4%	2	0.4%
CN	34	7.9%	5	6.9%	39	7.8%
CC	7	1.6%	1	1.4%	8	1.6%
CO	1	0.2%	1	1.4%	2	0.4%
Sum	428	100.0%	72	100.0%	500	100.0%

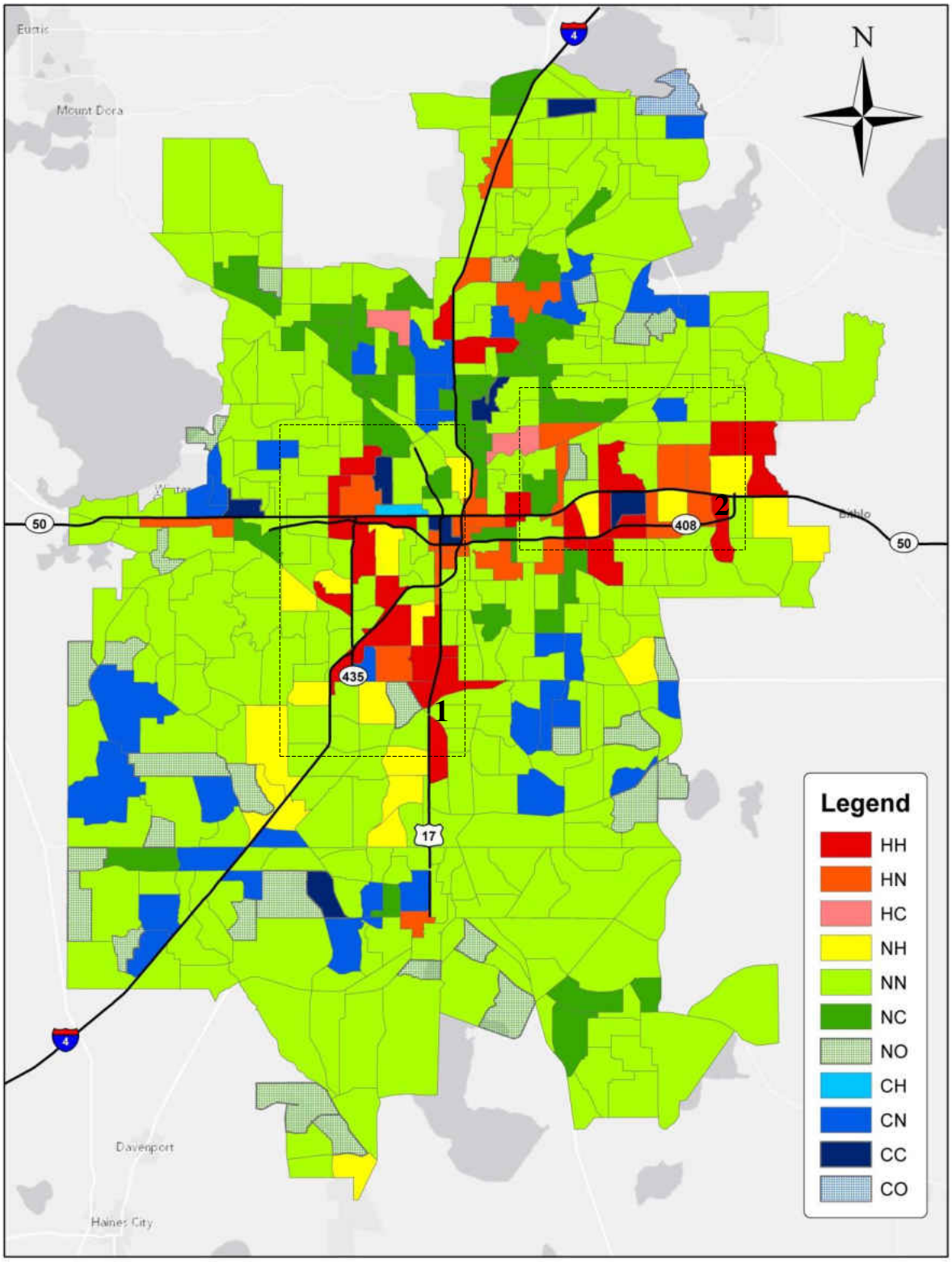


Figure 11-3: Distribution of zones by hot zone classification in the urban area (total crashes)

Figure 11-4 displays the spatial distribution of TSAZs by hot zone classification for total crashes in the rural area. It was found that HH/HN/NH zones are placed near principle arterial roadways including State Road 528 (Beachline Expressway), State Road 520 and State Road 91 (Florida's Turn Pike), or they are adjacent to the urban area. However, compared to urban area, HH zones for total crashes in the rural area form no clusters and all of them are spatially isolated. Two zones are located in the east (near State Road 520). The first HH zone in the northwest has a mixed land use of residential and commercial area, and it has a collector road crossing the zone (County Road 435). The other HH zone is placed in the southwest and its land use is also a mixture of residential and commercial area. County Road 531 is a boundary of this zone, which functions as a collector. Only one zone in the rural area is classified as CC zone for total crashes. This zone is mainly an agriculture area with some residential buildings.

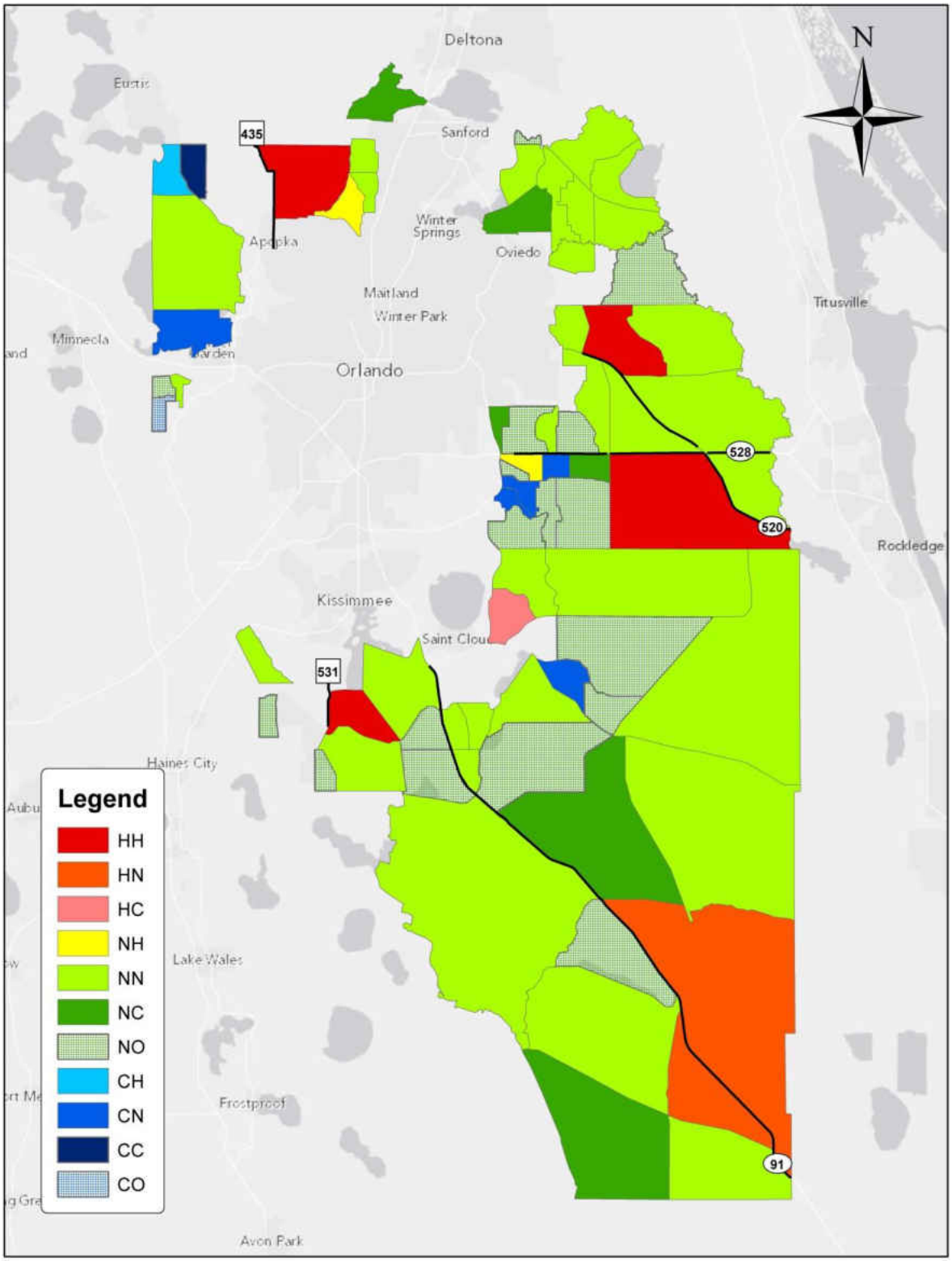


Figure 11-4: Distribution of zones by hot zone classification in the rural area (total crashes)

Table 11-3 compares the local features of HH and CC zones in the urban and rural area. For HH zones in the urban area, it was shown that 'Population density' in HH zones is more than three times larger than that in the entire urban area. Both 'Proportion of Hispanics' and 'Number of hotel, motel, and timeshare rooms per square mile' in HH zones are larger compared to those in the overall urban area. Moreover, 'Proportion of roadways with 55 mph or higher speed limits' is higher in HH zones as well. It suggests that zones containing more high speed roadways are more vulnerable for traffic crash occurrence.

Concerning CC zones in the urban area, they also have larger 'Population density' compared to the average. However, 'Proportion of Hispanics' in CC zones is slightly lower than those of the averages. Likewise, it was revealed that 'Number of hotel, motel, and timeshare rooms per square mile' and 'Proportion of roadways with 55 mph or higher speed limits' in CC zones are also smaller compared to the averages.

In comparison of zonal features in Average, HH, and CC zones in the rural area (Table 11-3), it was revealed that 'Population density' in HH zones is much larger than the average whereas that in CC zones is only half of the average. 'Proportion of Hispanics' in HH zones is also higher than the average whereas that in CC zones is slightly lower.

Table 11-3: Comparison of zonal features between the average, HH, and CC zones (total crash)

Zonal factors	**Urban**			**Rural**		
	Average	**HH**	**CC**	**Average**	**HH**	**CC**
Population density	410.0	1258.0	1297.5	124.4	551.9	62.4
Proportion of Hispanics	0.274	0.340	0.240	0.279	0.399	0.238
Number of hotel, motel, and timeshare rooms per square mile	139.7	590.9	65.86	51.05	1.138	0.000
Proportion of roadways with 55 mph or higher speed limits	0.052	0.077	0.045	0.075	0.000	0.000

11.3.2 Fatal-and-injury crash

The hot zone screening for total crashes represents general crash hot zone distributions as shown in the previous section. Nevertheless, it is also necessary to examine where severer crashes occur and the corresponding features. Thus, the result of fatal-and-injury crash hot zone identification was conducted and compared with the results of the total crash hot zone identification.

Table 11-4 summarizes the number of zones by hot zone classification for fatal-and-injury crashes. In the first, we start from the zones which are consistent in macro-and micro levels. It was shown that there are only 12 HH and 2 CC zones identified. Considering that there are 26 HH and 8 CC zones for total crashes, the number of HH/CC zones for fatal-and-injury crashes is quite smaller than those in total crash case. It seems that the consistency in hot zone/cold zone classification between macro- and micro-level is reduced in the fatal-and-injury crash case. It

might be interpreted that fatal-and-injury crashes are more influenced by network level characteristics than zonal factors, in comparison with total crashes.

Furthermore, it was observed that there is a little difference in the percentage of each category in between the urban and rural area. The proportion of HH zones in the urban area is 2.1% whereas that in the rural area is 4.2%. Similarly, the proportion of CC zones in the urban area is only 0.2% but that in the rural area is 1.4%. It shows the hot zone classification from two levels is more consistent in the rural area than the urban area.

Table 11-4: Number of zones by hot zone classification (fatal-and-injury crash)

Classification	Urban		Rural		Sum	
	Zones	%	Zones	%	Zones	%
HH	9	2.1%	3	4.2%	12	2.4%
HN	26	6.1%	4	5.6%	30	6.0%
HC	7	1.6%	0	0.0%	7	1.4%
HO	0	0.0%	0	0.0%	0	0.0%
NH	31	7.2%	4	5.6%	35	7.0%
NN	253	59.1%	30	41.7%	283	56.6%
NC	35	8.2%	7	9.7%	42	8.4%
NO	24	5.6%	16	22.2%	40	8.0%
CH	2	0.5%	0	0.0%	2	0.4%
CN	34	7.9%	5	6.9%	39	7.8%
CC	1	0.2%	1	1.4%	2	0.4%
CO	6	1.4%	2	2.8%	8	1.6%
Sum	428	100.0%	72	100.0%	500	100.0%

As seen in Figure 11-5, majority of HH/HC zones in the urban area are located along State Road 50 and State Road 408, however, HH/HN zones near Interstate 4 were considerably reduced compared to total crash hot zones. As mentioned earlier, NH zones for total crashes are concentrated in the downtown Orlando, however, NH zones for FI crashes are dispersed from the center of Orlando and most of them are located in suburban area. It implies that severer crashes are more vulnerable in the suburban area than in the urban area. It is explained that total crash risks are higher in the urban area since it has more largely exposed to traffic but driving speed in the urban area is slower than that in the suburban area.

It was also observed that HH zones form two clusters. The first cluster is placed between State Road 435 (Kirkman Road) and US 17/92/441 (Orange Blossom Trail) near Interstate 4 in the center of Orlando. The second cluster is located in East Orlando along State Road 50 (Colonial Drive) and surrounds the University of Central Florida.

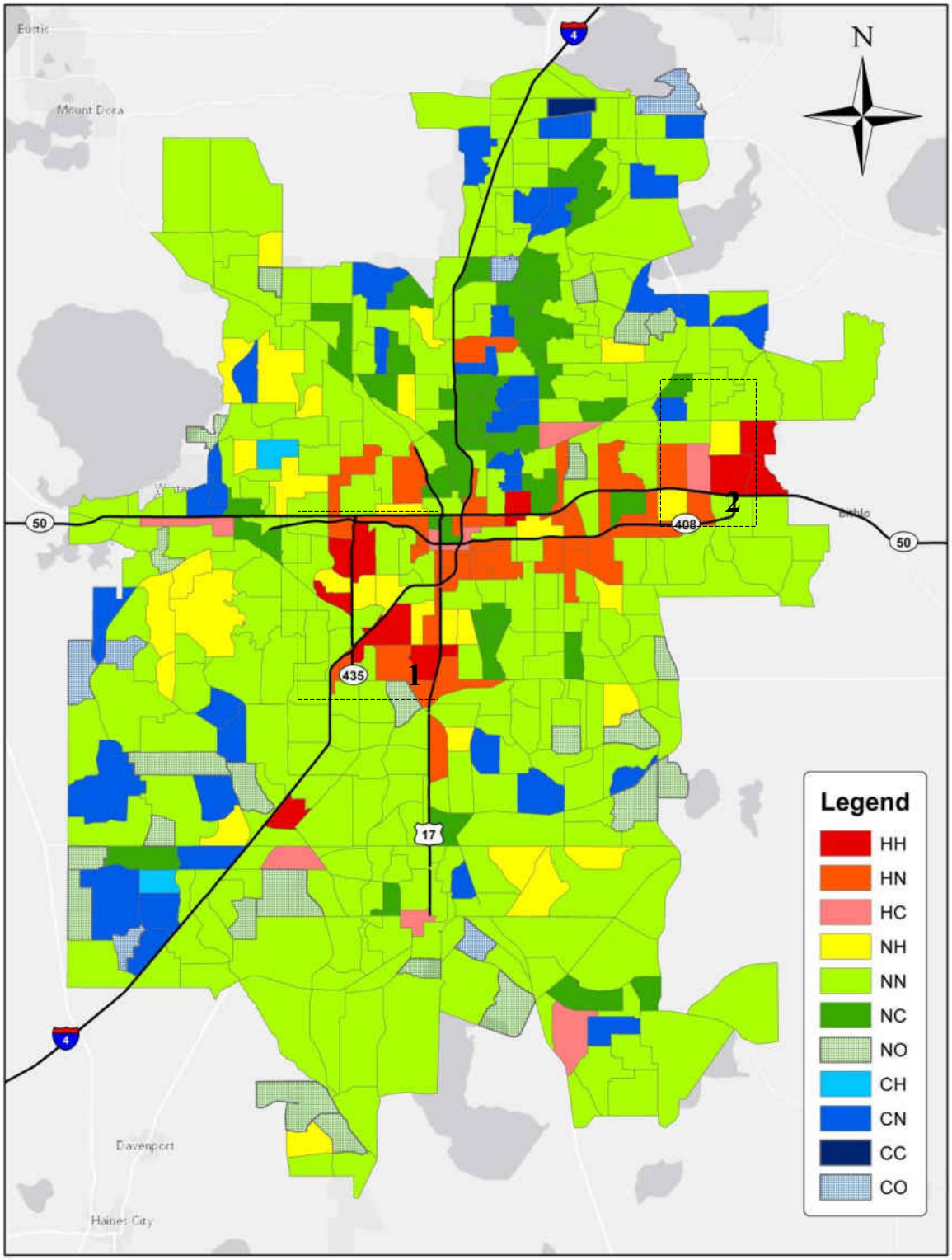


Figure 11-5: Distribution of zones by hot zone classification in the urban area (fatal-and-injury crashes)

As for the rural area, Figure 11-6 shows the spatial distribution of TSAZs by hot zone classification. Similarly in the total crash case, majority of HH/HN zones of FI crashes are located near main arterial roadways such as State Road 528, State Road 520 and State Road 91. Only few HH/HN zones are close to the urban area.

A HH zone in the northwest, which was also classified as HH zone for the total crashes, has a mixed land use of residential and commercial area. It was shown that only zone in the rural area is classified as CC zone for fatal-and-injury crashes, and it is residential areas. It was found that most of HH/HN zones for fatal-and-injury crashes are also categorized into HH/HN zones. It denotes that zones that are vulnerable to total crashes are also dangerous for fatal-and-injury crashes. It is possible that it is because crashes occurring in the rural area are severer than those in the urban area.

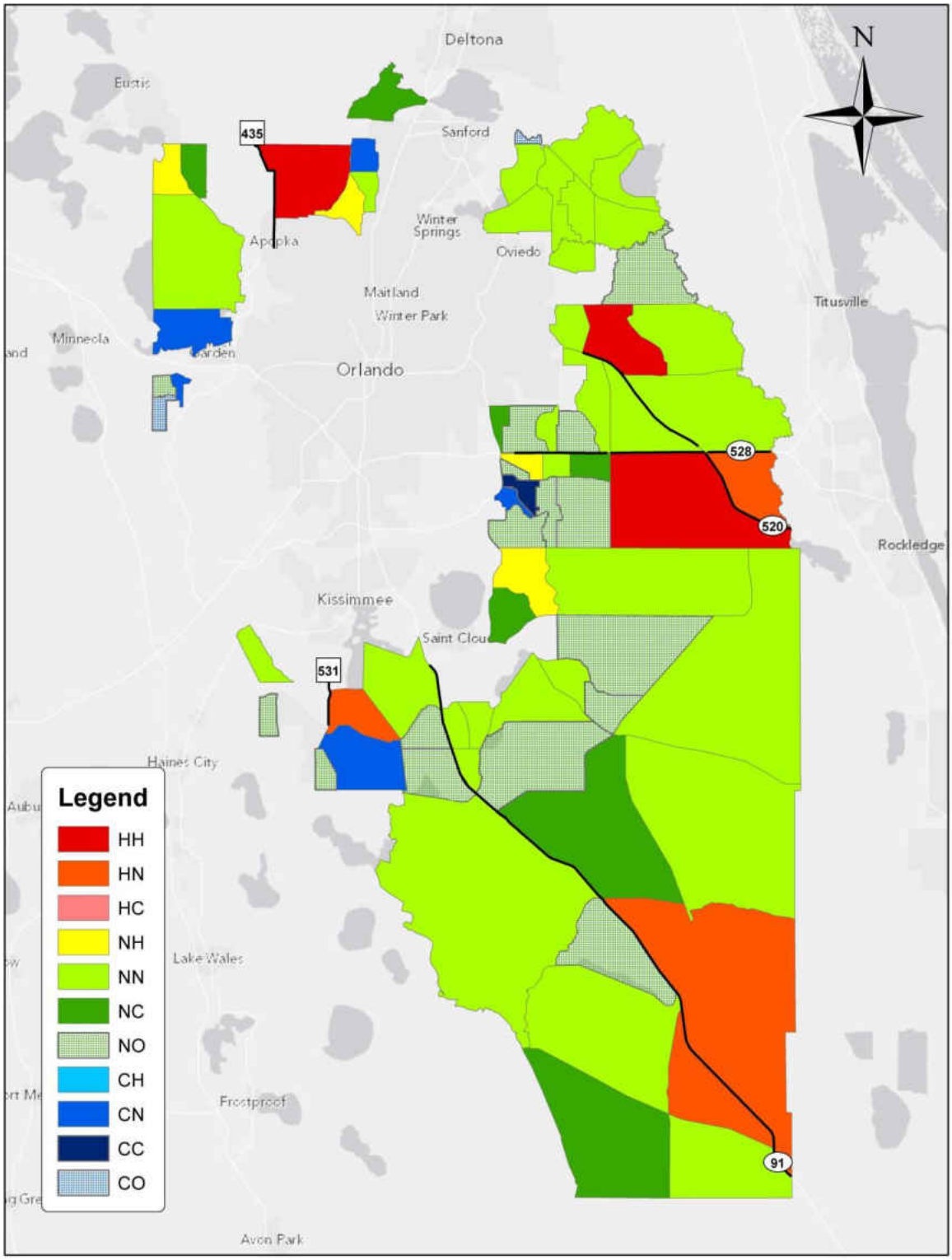


Figure 11-6: Distribution of zones by hot zone classification in the rural area (fatal-and-injury crashes)

Table 11-5 compares zonal features between the average value of all area, HH, and CC zones for fatal-and-injury crashes. In the urban area, 'Population density' of both HH and CC zones are larger than the average. However, 'Number of hotel, motel, and timeshare rooms per square mile' in HH zones is nearly triple of the average. In contrast, 'Number of hotel, motel, and timeshare rooms per square mile' in CC zones is only half of the average. 'Proportion of roadways with 55 mph or higher speed limits' in HH zones is 8.0%, and it is higher than that of the average (5.2%).

Table 11-5: Comparison of zonal features between the average, HH, and CC zones (fatal-and-injury crash)

Zonal factors	**Urban**			**Rural**		
	Average	**HH**	**CC**	**Average**	**HH**	**CC**
Population density	410.0	733.4	1612.2	124.4	287.2	18.6
Proportion of Hispanics	0.274	0.221	0.213	0.279	0.466	0.233
Number of hotel, motel, and timeshare rooms per square mile	139.7	338.5	61.83	51.05	1.227	0.000
Proportion of roadways with 55 mph or higher speed limits	0.052	0.080	0.000	0.075	0.000	0.000

As for the rural area, 'Population density' in the HH zones is much larger than the average but that in CC zones is quite smaller compared to the average. It implies that population-dense areas (i.e., residential areas) are more dangerous for fatal-and-injury crashes in the rural area. Also, it was observed that there is a big gap of 'Proportion of Hispanics' between the HH zones and the average. Hispanics in the HH zones for fatal-and-injury are 46.6% of total population whereas the average Hispanics are 27.9% in all rural area.

11.4 Summary and Conclusion

A novel screening methodology for integrating two levels was developed and used in this chapter for hot zone identification. TSAZs were classified into twelve categories with considerations made for both macroscopic and microscopic screening results. It is recommended that different strategies for each hot zone classification be developed because each category has distinctive traffic safety risks at each of the different levels. For HH zones, both macro-level treatments (i.e., education, campaigns, enforcement, etc.) and micro-level treatments (i.e., engineering solutions) are required to improve the traffic safety of the entire area. For example, assuming that one zone has a high safety risk related to bicycle crashes at both the macro- and microscopic levels, only applying engineering treatments at the network level (i.e., adding bike lanes) might not be effective or efficient because the zone also has zonal level factors that contribute to bicycle crashes. Therefore it would be ideal to begin bicycle safety campaigns and education programs at bike facilities.

On the other hand, HN and HC zones might need a greater level of focus on macro-level treatments because no specific safety problems emerge at the microscopic level. For CH zones, applying micro-level treatments for specific hotspots could alleviate traffic risks more efficiently than other types of measures. As seen in the results of this research, no HO zones were identified by the case study. However, they might be observed in other study areas. If HO zones exist, it would mean that such zones do not have major roadways or intersections, but rather only local residential roads with high traffic crash risk. Thus, we would need to screen residential areas and provide macro-level solutions to prevent local traffic crashes (such as installing a traffic-calming

zone). Admittedly, NC, NO, CC, and CO zones are not priority zones for safety treatments because they are safe for now. Nevertheless, it is necessary to keep monitoring these areas because traffic crash patterns are unstable and traffic crash risks can be transferred to these zones from other adjusted zones, especially for NC and NO zones.

CHAPTER 12 CONCLUSION

12.1 Summary

In this dissertation, many methodologies to improve macroscopic traffic safety analysis were suggested. In Chapter 3, it was attempted to find out relationships between residence characteristics and the number of at-fault drivers. It was revealed that traffic crashes are a socio-economic problem related to the deprived socio-economic status and specific demographic conditions. The final model revealed that there are several demographic, socioeconomic and commute patterns of residence zones that contribute to crash occurrence. The contributing factors found from Chapter 3 were referenced for the final models. For example, the age factor that were found significant in Chapter 3 was used in the final models (i.e., 'Proportion of young people' and 'Proportion of older people'), and also the economic status factors (i.e., 'Median family income') was applied in the final models (i.e., 'Proportion of households without vehicle').

Multivariate modeling for multiple crash types by transportation modes was proposed to improve model predictability in Chapter 4. It was shown that the multivariate spatial error model performs best. In the aspect of traffic safety, it was shown that there are many factors commonly significant across crashes by transportation modes as stated previously. However, it is also possible that there are additional shared factors that are non-observed or were omitted in the modeling process. Unsurprisingly, the error correlation of bicycle and pedestrian crashes is very high, even without the spatial correlation. It may imply the existence of unobserved shared

factors across these two different modes. This relationship may be caused by common inherent characteristics between bicycle and pedestrian crashes. It is interesting that the error correlation between motor vehicle and pedestrian crashes is also found somewhat high; which indicate that there are possible common omitted factors between these two types of crashes. However, the multivariate modeling was not applied in the final models of the dissertation because the final models were only developed for total crashes and fatal-and-injury crashes. Nevertheless, it is recommended to consider the unobserved factors shared by different crashes types, while choosing additional variables in future studies.

In Chapter 5, MAUP study was carried out particularly to understand the zoning and scale effects in modeling in three types of geographic entities (i.e., CTs, BG, and TAZs). It was revealed that significant variables were not consistent across the geographic units. For example, TAZ based models have more roadway/traffic related explanatory variables whereas BG based models include more of the commute related variables.

In Chapter 6, a new zone system was developed exclusively for macro-level crash analysis. It was discussed that there are several possible limitations of TAZs for macroscopic crash modeling. First, TAZs might be too small for crash analysis. Second, since TAZs are often delineated by arterial roads, many crashes occur at or near the boundary of TAZs. The existence of many boundary crashes can result in inaccurate modeling results. One simple way to overcome these two possible limitations of TAZs is to combine small TAZs with similar traffic crash patterns (i.e., crash rates) into sufficiently large zones, which is called the regionalization

process. After the regionalization, ten new zones systems were created by different scales of zones. Among these zone systems, the zone system with the optimal scale was determined using Brown-Forsythe homogeneity of variance test. According to the result, zone systems have 500 to 700 zones were considered optimal for macroscopic crash analysis. In other words, zone systems with 500-700 zones can capture both local and global traffic crash characteristics. After all, the zone system with 500 zones was chosen as TSAZs since it minimizes boundary crashes.

In Chapter 7, various spatial autocorrelation conceptualization methods were explored. It was shown that there are spatial autocorrelations in the data set using Moran's *I* statistics. Overall three conceptualization methods were applied including 1) first order rook polygon contiguity, 2) inverse distance, and 3) inverse distance squared. It was shown that spatial autocorrelations are significant, regardless of conceptualization methods. Two SPFs for total and severe crashes were estimated and each model includes a spatial error term based on different conceptualization methods. It was revealed that there is no big difference in DIC among models using different conceptualization methods. Nevertheless, only the spatial error term conceptualized by first order rook polygon slightly improves the model performance both in total and severe crash models. Thus, final models will adopt a spatial error component based on first order rook polygon contiguity in the final model.

In Chapter 8, two methods were suggested to account for boundary crashes. First, a complex nested structure was constructed to estimated individual six sub-models for boundary and interior crashes by roadway types. It enables estimating boundary and interior crash models individually.

Second, a variable transformation was proposed to relate boundary crashes with adjacent multiple zones. It allows developing boundary crash models with zonal factors from neighboring zones. The variable transformation formula includes a weight to balance effects from the crash zone and its adjacent zones. The optimal weights were calculated for six sub-models. Optimal weights for FSB, OSB, and NSB total crash models were 0.7, 0.9, and 0.7, respectively. It implies that boundary crash types were greatly affected by the crash zone (70-90%) and rarely influenced by adjacent zones (10-30%). On the other hand, optimal weights for all interior crash models were 1.0. It is expected that the model based on the complex nested structure with the optimal variable transformation performs better than other models without nested structures.

In Chapter 9, NBPLSEM was adopted to estimate both total and fatal-and-injury crashes at the macroscopic level. NBPLSEM contains a spatial error term based on first rook contiguity that controls for spatial autocorrelations. NBPLSEM with six sub-models was estimated and the result shows that each sub-model has different significant variable sets. It also justifies that applying the nested structure is reasonable. Several goodness-of-fit measures such as MAD, RMSE, PMAD, and R_{FT}^2 were used for the comparison of models with different nested structures. The result revealed that the most complicated nested structure with the variable transformation outperforms all other models.

Hot zones/spots in Orange, Seminole, and Osceola Counties were identified both at the macroscopic and microscopic levels in Chapter 10. PSI, which is a measure of how many crashes can be effectively reduced, was chosen as a performance measure for the screening. Each TSAZ

was ranked for both total and fatal-and-injury crashes. The screening results from this chapter will be used for the integrated screening in Chapter 11.

As stated earlier, Chapter 11 proposes a new screening methodology for integrating two levels. TSAZs were classified into twelve categories with considerations made for both macroscopic and microscopic screening results. It is recommended that different strategies for each hot zone classification be developed because each category has distinctive traffic safety risks at each of the different levels. For HH zones, both macro-level treatments (i.e., education, campaigns, enforcement, etc.) and micro-level treatments (i.e., engineering solutions) are required to improve the traffic safety of the entire area. For example, assuming that one zone has a high safety risk related to bicycle crashes at both the macro- and microscopic levels, only applying engineering treatments at the network level (i.e., adding bike lanes) might not be effective or efficient because the zone also has zonal level factors that contribute to bicycle crashes. Therefore it would be ideal to begin bicycle safety campaigns and education programs at bike facilities. On the other hand, HN and HC zones might need a greater level of focus on macro-level treatments because no specific safety problems emerge at the microscopic level. For CH zones, applying micro-level treatments for specific hotspots could alleviate traffic risks more efficiently than other types of measures. As seen in the results of this research, no HO zones were identified by the case study. However, they might be observed in other study areas. If HO zones exist, it would mean that such zones do not have major roadways or intersections, but rather only local residential roads with high traffic crash risk. Thus, we would need to screen residential areas and provide macro-level solutions to prevent local traffic crashes (such as

installing a traffic-calming zone). Admittedly, NC, NO, CC, and CO zones are not priority zones for safety treatments because they are safe for now. Nevertheless, it is necessary to keep monitoring these areas because traffic crash patterns are unstable and traffic crash risks can be transferred to these zones from other adjusted zones, especially for NC and NO zones.

12.2 Research Implications

The findings from Chapter 3 may give some implications to safety researchers. Most of safety studies have only focused on physical roadway characteristics of segments, intersections, corridors or zones where traffic crashes have occurred. On the contrary, this chapter explored the residence zone's characteristics of dangerous drivers, which are believed to largely influence individual drivers' attitudes/habits. Consequently, it was shown that several key residence zone's factors contribute to crash occurrence. Further studies are required to analyze the residence zone's characteristics of drivers who were involved and caused various types of crashes. For practitioners or policy makers, the results can provide meaningful guidance for designing and tailoring specific education, engineering and awareness campaigns and stricter enforcement to reduce traffic crashes.

The implications from Chapter 4 are as follows: First, the multivariate modeling is recommended if multiple crash types (i.e., crashes by transportation modes, severity levels, or by time periods, et cetera) are simultaneously analyzed. The multivariate modeling can provide better predictability for multiple crash types by considering commonly unobservable factors among crash types. Second, practitioners are suggested to analyze crashes by multiple transportation

modes because they are very highly correlated as shown in Chapter 4. By doing so, not only better reliable and accurate screening results are available but also more efficient and comprehensive safety treatments can be planned and provided.

Chapter 5 has important implications for both researchers and practitioners. For several pragmatic reasons (i.e., incorporating with transportation planning, etc.), TAZs are preferred by practitioners, although TAZs have not been examined thoroughly for traffic safety analysis. Therefore, researchers should focus on MAUP and investigate if TAZs are desired zone system for the traffic safety modeling. If TAZs have limitations or problems for crash modeling, it is recommended, as highlighted in Chapter 6, to develop a new zonal system for the macroscopic safety analysis.

Chapter 6 carries many implications for researchers and practitioners. From an application perspective, it is expected that TSAZs have advantages as follows: First, TSAZs are still based on TAZs. So it allows for transportation planners to incorporate traffic safety into long range transportation plans (LRTPs). Second, TSAZs were developed using relatively simple methods (i.e., regionalization and Brown-Forsythe tests). Thus, practitioners can develop TSAZs by themselves without much time and effort. Third, TSAZs do not contain extremely small zones whereas TAZs have very small zones especially in the downtown area. It makes zonal-level screening much more efficient at the macroscopic level. In the research point of view, TSAZs can produce models with better fit compared to TAZs. In other words, more accurate models can be developed based on TSAZs. Besides, TSAZs can reflect both local and global characteristics

of crash patterns as shown in the Brown-Forsythe test results. If the study area does not have developed TAZs, it is also possible to construct TSAZs by using other geographic units such as CTs or BGs. In this case, the TSAZs still have homogenous traffic crash patterns in each zone even though the TSAZs are not developed based on TAZs, Thus, it is expected that the aggregated TSAZs are more appropriate for the macroscopic crash analysis compared to other zonal systems without consideration of traffic crash patterns.

The findings from Chapter 7 are useful for traffic safety researchers and spatial analysts. Few researchers have focused on the spatial autocorrelation conceptualization issues for the traffic safety analysis. Although many zonal-level studies have adopted first order rook polygon contiguity based spatial effect term in their models, it has not been proven with empirical evidence. In Chapter 7, it was clearly revealed that the first order rook polygon contiguity based spatial error term is better than others. It implies that spatial errors are correlated only among neighboring zones which are spatially connected.

Chapter 8 provides an important implication for traffic safety researchers. It is necessary to separate boundary and interior crashes since they have different assumptions for the statistical modeling. Interior crashes are hypothesized to be influenced solely by the zone within which they are spatially located. Eventually, it was found that these assumptions are valid. Therefore, it is recommended for researchers to estimate separate models for boundary and interior crashes.

Chapter 9 also has several implications. Although the nested structure suggested in Chapter 8 is complicated, it is worth to apply since it improves the predictability significantly. Moreover, each sub-model has its own significant variable set. It also supports the validity of the nested structure. Researchers have difficulties to handle crashes occurring on or near boundaries of the zone, and in the literature several simplistic and inaccurate assumptions were made. The methodologies suggested in Chapter 8 and 9 (i.e., nested structure, variable transformation, optimal weights for boundary crashes, etc.) can efficiently solve the boundary crash issues.

The findings from Chapter 10 are very helpful for practitioners. The screening results revealed that most of the hot zones in the rural areas were close to the fringe of the urban areas, or they contained major arterials or full access control roads. With regards to urban areas, the hot zones mostly were located in area from downtown to eastern Orlando along the major arterial state road (SR50). As for fatal-and-injury crashes, the hot zones had patterns very close to those of the total crash hot zones but the hot zones for fatal-and-injury crashes were closer to high speed roads in both urban and rural areas The microscopic screenings were also conducted for roadway segments and intersections which have AADT information for both total and fatal-and-injury crashes.

Chapter 11 suggests key implications both for practitioners and researchers. It is expected that the innovative integrated screening approach proposed in Chapter 11 can provide a comprehensive perspective by balancing macroscopic and microscopic screening results. It is recommended for practitioners to apply different safety strategies for each hot zone classification

because each category has distinctive traffic safety risk at each of the different levels. By doing so, they can plan and provide effective traffic safety treatments to reduce traffic crashes. It is also the researchers' responsibility to suggest and evaluate effective traffic safety programs for zones at both macroscopic and microscopic levels (HH), zones only which is risky at the microscopic (CH, NH, etc.), and zones that have safety problems solely at the macroscopic level (HN, HC, etc.).

12.3 Conclusion

Many studies have been done to analyze traffic crashes at the macroscopic level. Nevertheless, there are several issues in the macroscopic analysis and some of the issues were addressed in dissertation. First, there have been no geographic units exclusively developed for the traffic safety analysis. Admittedly, TAZs, which are related to the transportation/traffic analysis, have been widely used for the macroscopic analysis. However, TAZs have critical drawbacks such as boundary crashes and too small sizes in the urban area, as stated previously. In order to address this problem, TSAZs were developed by aggregating current TAZs into sufficiently large and homogenous zones. Brown-Forsythe test was conducted to determine the optimal scale that can keep both local and global traffic crash patterns. TSAZs have several advantages compared to current TAZs for the macroscopic analysis. It was revealed that 500-700 zones are optimal scale for the macroscopic traffic crash analysis. Eventually, 500 zones were chosen as TSAZs since it can minimize the number of boundary crashes between 500 and 700 zones. Approximately 10% of boundary crashes have been eliminated after the regionalization but more than 60% of crashes still occur on or near the boundary of TSAZs.

Subsequently, a nested structure was proposed to estimate safety performance models separately for boundary and interior crashes. This nested structure allows different contributing factors for different crash types, so this model can provide more accurate and predictable results than a single model. The six types of crashes in each model are varied based on their locations (boundary or interior) and roadway types (freeways-and-expressways, other state roads, or non-state roads). They are FSB, FSI, OSB, OSI, NSB, and NSI. For boundary crash models, explanatory variables were transformed to reflect both crash zone and its adjacent zones. Also, optimal weights, which can balance influences between the crash zone and its neighboring zones efficiently, were calculated and used for the variable transformation.

Furthermore, a Bayesian Poisson Lognormal Spatial Error Model (BPLSEM) was adopted for the SPF analysis in this nested structure. The BPLSEM contains a disturbance term for handling the over-dispersion problem, and its spatial error term can control for the spatial autocorrelation of the crash data. In addition, the PSI (Potential for Safety Improvements), the difference between the expected crash count and the predicted crash count, was used as a performance measure to identify hot zones.

After identifying hot spots/zones at both macroscopic and microscopic levels, the screening results from two levels were combined. However, this integration task was challenging because it was necessary to 1) combine various SPFs from different scales, areas, and roadway types; 2) determine an appropriate weight for each group; and 3) choose a performance measure for the final results. In order to solve the above mentioned problems, this study then developed a new

criterion to identify whether a zone has safety issues at the macroscopic and microscopic levels. All TSAZs were classified into twelve categories that include two levels (macroscopic and microscopic levels) and four safety categories (Hot, Normal, Cold, and No data). These categories are: HH, HN, HC, HO, NH, NN, NC, NO, CH, CN, CC, and CO. The first character of the classification represents the macroscopic safety risk, and the second character illustrates the microscopic safety risk. At the macroscopic level, TSAZs were ranked by their zonal PSIs; at the microscopic level, the calculation of average PSI was more complicated because each TSAZ had several intersections and segments. The PSIs of the intersections in each TSAZ were averaged by the number of intersections, and the zones were ranked by their averaged intersection PSI. Simultaneously, the PSIs of segments in each zone were averaged by the total length of the segments in the zone, and zones were ranked by their averaged segment PSI. After that, both the intersection and segment PSI ranks were averaged; the TSAZs were ranked by the final averaged intersection and segment PSIs. As was the case at the macroscopic level, TSAZs with top 10% micro-level PSIs were categorized as "Hot" zones at the microscopic level.

Finally, the percentile ranks of the PSIs were used in the integration (instead of the original PSIs) because the units of PSI intersections and PSI segments were different. Hot TSAZs for both total crashes and fatal-and-injury crashes were analyzed, in order to be consistent with the HSM. Moreover, by doing so the results also allowed an examination of whether there are any differences with regards to hot zone locations among various crash severity levels. The total crash hot zone screening results display the overall crash distributions within the study area, whereas the fatal-and-injury crash hot zone screening results represent severer crash distributions.

The integrated screening methodology suggested in this dissertation provides a comprehensive perspective on appropriate safety treatments by balancing the accuracy and efficiency of screening. Also, it is recommended that different strategies for each hot zone classification be developed because each category has distinctive traffic safety risks at each of the different levels.

Specific safety treatments to reduce traffic crash risks can be designed in consideration of modeling results suggested in Chapter 9. It was shown that proportion of young people has a positive effect on crash counts on freeway-and-expressway. Thus, it is recommended to provide safety education or campaign for young people at high schools and colleges, in order to reduce both total crashes and fatal-and-injury crashes on freeway-and-expressways. Furthermore, it was revealed that proportion of Hispanics has a positive relationship with total and fatal-and-injury crash occurrences on other state roads. Hence, the crashes happening on other state roads can be reduced by providing campaigns and flyers in Spanish language to zones with larger Hispanic population. Similarly, households without vehicle were also positively associated with the number of crashes on other state roads. In this dissertation, public transportation related crashes or non-motorized crashes were not separately considered for the modeling. Nevertheless, one of the contributing factors for other state roads, proportion of households without car is found that it has positive effects on crash counts. People from households without car should use alternative modes (i.e., public transportation, bicycles, walking, etc.). It seems the increased crashes by households without car were related to public transportation, bicyclists and/or pedestrians. As a result, it is suggested that providing sidewalks with enough width, sidewalk barriers, and bicycle exclusive lanes for alleviating crashes on other state roads in zones with higher proportion of

households without households. Furthermore, it was revealed that zones with many accommodation rooms have more crashes. It may imply that tourists who are not familiar with local roadways and rules are exposed more to traffic crashes. Thus, it is necessary to provide traffic safety information to tourists at hotel, motel, attractions, or airports. Otherwise, a higher design standard for roadways may reduce traffic crashes occurred by tourists. Lastly, poor pavement conditions are likely to increase both total and fatal-and-injury crash occurrence on non-state roads. It is expected that resurfacing segments with poor pavement condition to reduce traffic crash risks on local or residential roads.

Although many issues were addressed in this dissertation, it should be noted that there are several limitations to the integration method suggested in this dissertation. First, only data from three counties were used to estimate the SPFs. It is suggested that future research evaluate the transferability of the SPFs developed in this study for other areas in Florida. Second, the variance of PSIs of microscopic components was not considered when calculating the average PSI in each zone. It may neglect the variance of PSIs between intersections/segments in a zone. Suppose that two zones were classified as hot zones at the microscopic level, and one zone could have consistently high PSI segments/intersections with low variances, whereas the other zone could have segments/intersections with high variances in their PSIs. In the former case, the zone would be uniformly risky at the microscopic level, so area-wide engineering treatments should be considered. In contrast, the zone in the latter case would have some extremely high risk segments/intersections but other segments/intersections would not be such dangerous. In this case, it is recommended that countermeasures be applied only for the specific sites. Also, if the

practitioners are more interested in crash costs, PSIs can be replaced with other performance measures such as the EPDO (Equivalent Property Damage Only) crash frequency method. Lastly, only total crashes and fatal-and-injury crashes were analyzed in this dissertation. The proposed integrated screening method would be also useful because hot zones for other various types of crashes could be identified considering both macroscopic and microscopic levels, so practitioners could comprehensively understand the hot zone locations of specific crash types and provide appropriate traffic safety treatments.

REFERENCES

Abbas, K. A., 2004. Traffic safety assessment and development of predictive models for accidents on rural roads in Egypt. Accident Analysis and Prevention 36 (2), pp. 149–163.

Abbess, C., Jarret, D., and Wright, C. C., 1981. Accidents at blackspots: estimating the effectiveness of remedial treatment, with special reference to the "regression-to-the-mean" effect. Traffic Engineering and Control 22(10), pp. 535-542.

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

Abdel-Aty, M. A., and Radwan, A. E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention 32, pp .663-642.

Abdel-Aty, M. A., Siddiqui, C., and Huang, H., 2011. Integrating trip and roadway characteristics in managing safety at traffic analysis zones. Transportation Research Record 2213, pp. 20-28.

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

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

Aguero-Valverde, J., and Jovanis, P. P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transportation Research Record: Journal of the Transportation Research Board, Volume 2136, pp. 82-91.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), pp. 716-723.

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

Baass, K. G., 1980. Design of zonal systems for aggregate transportation planning models. Transportation Research Record, Volume 807, pp. 1-6.

Baker, S. P., Braver, E. R., Chen, L. H., Pantula, J. F., and Massie, D., 1998. Motor vehicle occupant deaths among Hispanic and black children and teenagers. Archives of Pediatrics and Adolescent Medicine, 152(12), pp. 1209-1212.

Bates, L. J., Soole, D., and Watson, B. 2012. The effectiveness of traffic policing in reducing traffic crashes. Policing and Security in Practice: Challenges and Achievements, pp. 90-109.

Beck, L. F., Dellinger, A. M., and O'Neil, M. E., 2007. Motor vehicle crash injury rates by mode of travel, United States: using exposure-based methods to quantify differences. American Journal of Epidemiology, 166(2), pp. 212-218.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society B 36(2), pp. 192-236.

Bijleveld, F. D., 2005. The Covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis and Prevention 37(4), pp. 591–600.

Blatt, J., and Furman, S. M., 1998. Residence location of drivers involved in fatal crashes. Accident Analysis and Prevention 30(6), pp. 705-711.

Cafiso, S., Di Silvestro, G., Persaud, B., Begum, M.A., 2010. Revisiting the variability of the dispersion parameter of safety performance functions using data for two-lane rural roads.Transportation Research Record 2148, pp. 38-46.

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

Chang, K. T., Khatib, Z., and Ou, Y. M., 2002. Effects of zoning structure and network detail on traffic demand modeling. Environment and Planning B: Planning and Design 29, pp. 37-52.

Chung, K., and Ragland, D., 2007. A method for generating a continuous risk profile for highway collisions, UC Berkeley Traffic Safety Center.

Clark, D. E., 2003. Effect of population density on mortality after motor vehicle collision. Accident Analysis and Prevention 35, pp. 965-971.

Congdon, P., 2001. Bayesian statistical modelling. John Wiley & Sons, Ltd., United Kingdom.

Congdon, P., 2003. Applied Bayesian modelling. John Wiley & Sons, Ltd., United Kingdom.

Cottrill, C. D., and Thakuriah, P., 2010. evaluating pedestrian crashes in areas with high low-income or minority populations. Accident Analysis and Prevention 42 (6), pp. 1718-1728.

Cutrona, C. E., Wallace, G., and Wesner, K. A., 2006. Neighborhood characteristics and depression: an examination of stress processes. Current Directions in Psychological Science, 15(4), pp. 188-192.

Daniels, S., Brijs, T., Nuyts, E., Wets, G., 2010. Explaining variation in safety performance of roundabouts. Accident Analysis and Prevention 42(2), pp. 393-402.

Dawkins, C. J., Shen, Q., and Sanchez, T. W., Race, space, and unemployment duration. Journal of Urban Economics 58, pp. 91-113.

Ding, C., 1998. The GIS-based human-interactive TAZ design algorithm: examining the impacts of data aggregation on transportation-planning analysis. Environment and Planning B: Planning and Design 25, pp. 601-616.

El-Basyouny, K., and Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models. Transportation Research Record 1950, pp. 9–16.

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

El-Basyouny, K., and Sayed, T., 2010. Application of generalized link functions in developing accident prediction models. Safety Science 48, pp. 410-416.

Elvik, R., and Mysen, A. B., 1999. Incomplete accident reporting—meta-analysis of studies made in 13 countries. Transportation Research Record, No.1665, pp.133-140.

FHWA, 2005. SAFETEA-LU: Safe, accountable, flexible, efficient transportation equity act: a legacy for users. Federal Highway Administration (FHWA), www.fhwa.dot.gov/safetealu, Accessed Jan 1, 2013.

Fotheringham, A. S., and Rogerson, P. A., 1993. GIS and spatial analytical problems. International Journal of Geographical Information Systems 7(1), pp. 3- 19.

Freeman, M. F., and Tukey, J. W., 1950. Transformations related to the angular and the square root. The Annals of Mathematical Statistics, pp. 607-611.

Fridstrom, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., and Thomsen, L. K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accident Analysis and Prevention, 27(1), 1-20.

Fridstrom, L., and Ingebrigtsen, S., 1991. An aggregate accident model based on pooled, regional time-series data, Accident Analysis and Prevention 23(5), pp. 363-378.

Gehlke, C. E., and Biehl, K., 1934. Certain effects of grouping upon the size of the correlation coefficient in census tract material, Journal of the American Statistical Association 29 (185), Supplement: Proceeding of the American Statistical Journal, pp. 169-170.

Girasek, D. C., and Taylor, B., 2010. An exploratory study of the relationship between socioeconomic status and motor vehicle safety features. Traffic Injury Prevention 11(2), pp. 151-155.

Gruenewald, P. J., Freisthler, B., Remer, L., LaScala, E. A., and Treno, A., 2006. Ecological models of alcohol outlets and violent assaults: crime potentials and geospatial analysis. Addiction 101(5), pp. 666-677.

Guevara, F. L., Washington, S. P., and Oh, J., 2004. Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record 1897, pp. 191-199.

Guo, D., 2008. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). International Journal of Geographical Information Science, 22(7), pp. 801-823.

Guo, D., and Wang, H., 2011. Automatic region building for spatial analysis, Transactions in GIS 15, pp. 29-45.

Gurstein, P., 1996. Planning for telework and home-based employment: Reconsidering the home/work separation. Journal of Planning Education and Research, 15(3), pp. 212-224.

Gyimah-Brempong, K., 2006. Neighborhood income, alcohol availability, and crime rates. The Review of Black Political Economy 33(3), pp. 21-44.

Hadayeghi, A., Shalaby, A. S., and Persaud, B. N., 2003. Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record 1840, pp. 87-95.

Hadayeghi, A., Shalaby, A. S., Persaud, B. N., and Cheung, C., 2006. Temporal transferability and updating of zonal level accident prediction models. Accident Analysis and Prevention 38, pp. 579-589.

Hadayeghi, A., Shalaby, A. S., and Persaud, B. N., 2010a. Development of planning-level transportation safety models using full Bayesian semiparametric additive techniques. Journal of Transportation Safety and Security 2(1), pp. 45-68.

Hadayeghi, A., Shalaby, A. S., and Persaud, B. N., 2010b. Development of planning-level transportation safety tools using geographically weighted poisson regression. Accident Analysis and Prevention 42, pp. 676-688.

Harper, J. S., Marine, W. M., Garrett, C. J., Lezotte, D., and Lowenstein, S. R., 2000. Motor vehicle crash fatalities: a comparison of hispanic and non-hispanic motorists in colorado. Annals of Emergency Medicine 36(6), pp. 589-596.

Hauer, E. and Hakkert, A. S., 1989. Extent and some implications of incomplete accident reporting. Transportation Research Record, No.1185, Transportation Research Board, National Research Council, Washington, D.C., pp.1-10.

Hauer, E., 1996. detection of safety deterioration in a series of accident counts, Transportation Research Record 1542, pp. 38-43.

Hauer, E., Ng, J. C. N., and Lovell, J., 1988. Estimation of safety at signalized intersections. Transportation Research Record 1185, pp. 48–61.

Hausman, J A., Hall, B. H., and Griliches, Z., 1984. Econometric models for count data with an application to the patents-R&D relationship. Econometrica 52 (4), pp. 909–938.

Heydecker, B. J., and Wu, J., 1991. Using the information in road accident records proc. 19th PTRC Summer Annual Meeting, London.

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

Hirst, W. M., Mountain, L.J., Maher, M.J., 2004. Sources of error in road safety scheme evaluation: a method to deal with outdated accident prediction models. Accident Analysis and Prevention 36 (5), pp. 717–727.

Hoyert, D. L., and Xu, Ji., 2012. National vital statistics report 61(6), US Department of Health and Human Services, Centers for Disease Control and Prevention.

Huang, H., Abdel-Aty, M. A., and Darwiche, A. L., 2010. County-level crash risks analysis in Florida: Bayesian spatial modeling. Transportation Research Record 2148, pp. 27-37.

Huang, H., Xu, P., and Abdel-Aty, M. A., 2013. Transportation safety planning: spatial analysis approach. Presented in Transportation Research Board 92nd Annual Meeting.

Johansson, P., 1996. Speed limitation and motorway casualties: a time series count data regression approach. Accident Analysis and Prevention 28(1), pp. 73–87.

Jones, B., Janssen, L., and Mannering, F., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. Accident Analysis and Prevention 23 (2), pp. 239–255.

Joshua, S. C., and Garber, N. J., 1990. Estimating truck accident rate and involvements using linear and poisson regression models. Transportation Planning and Technology 15(1), pp. 41–58.

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

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

Keeney, R. L., 1980. Evaluating alternatives involving potential fatalities, Operation Research 28, pp. 188-205.

Kenessey, Z., 1987. The primary, secondary, tertiary and quaternary sectors of the economy. Review of Income and Wealth 33(4), pp. 359-385.

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

Kim, K., Brunner, I. M., and Yamashita, E. Y., 2006. Influence of land use, population, employment, and economic activity on accidents, Transportation Research Record 1953, pp.56-64.

Kononov, J., and Allery, B., 2003. Level of service of safety: conceptual blueprint and analytical framework. Transportation Research Record 1840, pp 57-66.

Kumala, R., 1995. Safety at rural three- and four-arm junctions: development and applications of accident prediction models, vol. 233. VTT Publications, Technical Research Centre of Finland, Espoo, Finland.

Lai, P.-C., So, F.-M., and Chan, K.-W., 2008, Spatial epidemiological approaches in disease mapping and analysis, CRC Press.

LaScala, E. A., Gerber, D., and Gruenewald, P. J., 2000. Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. Accident Analysis and Prevention 32, pp. 651-658.

Lawson, A. B., Browne, W. J., and Rodiero, C. L. V, 2003. Disease mapping with WinBUGS and MLwiN. John Wiley & Sons Ltd., United Kingdom.

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

Lee, J., Abdel-Aty, M., & Choi, K. (2014). Analysis of residence characteristics of at-fault drivers in traffic crashes. Safety Science 68, pp. 6-13.

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

LeSage, J. P., and Pace, R. K., 2004. Models for spatially dependent missing data, The Journal of Real Estate Finance and Economics 29(2), pp. 233-254.

Levine, N., Kim, K. E., and Nitz, L. H., 1995. Spatial analysis of Honolulu motor vehicle crashes: II. zonal generators. Accident Analysis and Prevention 27 (5), pp. 675-685.

Lord, D., 2006. Modeling motor vehicle crashes using poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38(4), pp. 751–766.

Lord, D., and Bonneson, J. A., 2007. Development of accident modification factors for rural frontage road segments in Texas. Transportation Research Record 2023, pp. 20–27.

Lord, D., Geedipally, S. R., and Guikema, S., 2010. Extension of the application of Conway–Maxwell–Poisson models: analyzing traffic crash data exhibiting underdispersion. Risk Analysis 30(8), pp.1268-1276.

Lord, D., Manar, A., and Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-V/C ratio for rural and urban freeway segments. Accident Analysis and Prevention 37 (1), pp. 185–199.

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

Lord, D., and 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), pp. 751–770.

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

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

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

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

McFadden, D., 1974. The measurement of urban travel demand. Journal of Public Economics 3, pp. 303-328.

Males, M. A., 2009. Poverty as a determinant of young drivers' fatal crash risks. Journal of Safety Research 40, pp. 443-448.

Malyshkina, N., and Mannering, F., 2010b. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accident Analysis and Prevention 42(1), pp. 131–139.

Martinez, R., and Veloz, R. A. 1996. A challenge in injury prevention-the Hispanic population. Academic Emergency Medicine, 3(3), 194-197.

Maycock, G., and Hall, R. D., 1984. Accidents at 4-arm roundabouts. TRRL Laboratory Report 1120, Transportation and Road Research Laboratory, Crowthorne, Unied Kingdom.

Meyer, M. D., and Miller, E. J., 2001. Urban transportation planning: a decision-oriented approach. 2nd ed. New York: McGraw-Hill.

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

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

Miaou, S.-P., 1996. Measuring the goodness-of-fit of accident prediction models. FHWA-RD-96-040. Federal Highway Administration, Washington, D.C.

Miaou, S.-P., and Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. Transportation Research Record 1840, pp. 31–40.

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

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

Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor vehicle accident frequencies. Transportation 25 (4), pp. 395–413.

Moran, P. A. P., 1950. Notes on continuous stochastic phenomena. Biometrika 37(1), pp. 17-23.

Naderan, A., and Shahi, J., 2010. Aggregate crash prediction models: introducing crash generation concept. Accident Analysis and Prevention 42, pp. 339-346.

New South Wales Roads and Traffic Authority, 1996. Road whys speeding module presenter's booklet Regret is such a short distance.

Ng, K., Hung, W., and Wong, W., 2002. An algorithm for assessing the risk of traffic accident. Journal of Safety Research 33(3), pp. 387–410.

NHTSA, 2013. 2012 Motor vehicle crashes: overview, U.S. Department of Transportation, National Highway Traffic Safety Administration.

Nicholson, A. J., 1985. The variability of accident counts. Accident Analysis and Prevention 17(1), pp. 47-56.

Noland, R. B, 2003. Traffic Fatalities and Injuries: The effect of changes in infrastructure and other trends. Accident Analysis and Prevention 35, pp. 599-611.

Noland, R. B., and 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 and Prevention 36, pp. 525-532.

Noland, R. B., and Quddus, M. A., 2004. Analysis of pedestrian and bicycle casualties with regional panel data. Transportation Research Record, No. 1897, pp. 28-33.

Openshaw, S., 1984. Ecological fallacies and the analysis of areal census data. Environment and Planning A 16(1), pp. 17-31.

O'Sullivan, D., and Unwin, D., 2002. Geographic information analysis. John Wiley and Sons.

Owsley, C., 2002. Driver capabilities. In: Proceedings of the Transportation in an Aging Society: A Decade of Experience. Transportation Research Board, National Research Council, The National Academies, Washington, D.C.

Park, E. S., and Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, pp. 1–6.

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

Persaud, B. P., and Nguyen, T., 1998. Disaggregate safety performance models for signalized intersections on ontario provincial roads. Transportation Research Record 1635, pp. 113–120.

Philbrook, J. K., and Franke-Wilson, N. A., 2009. The effectiveness of a peer lead smart driving campaign on high school students' driving habits. The Journal of Trauma and Acute Care Surgery, 67(1), S67-S69.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., and Wets, G., 2012. Developing zonal crash prediction models with a focus on application of different exposure measures. Presented in Transportation Research Board 91st Annual Meeting.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., and Wets, G., 2013a. Evaluating the road safety effects of a fuel cost increase measure by means of zonal crash prediction modeling. Accident Analysis and Prevention 50, pp. 186-195.

Pirdavani, A., Brijs, T., Bellemans, T., and Wets, G., 2013b. Spatial analysis of fatal and injury crashes in Flanders, Belgium: application of geographically weighted regression technique. Presented in Transportation Research Board 92nd Annual Meeting.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., and Wets, G., 2013c. Assessing the impacts of a teleworking policy on crash occurrence: The case of Flanders, Belgium. Retrieved Dec 16, 2013, from https://doclib.uhasselt.be/dspace/handle/1942/14564

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

Pucher, J., and Dijkstra, L., 2003. Promoting safe walking and cycling to improve public health: lessons from the Netherlands and Germany. American Journal of Public Health, 93(9), pp. 1509-1516.

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

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

Ragland, D. R., and Chan, C.-Y., 2007. High collision concentration location: Table C evaluation and recommendations, UC Berkeley Traffic Safety Center.

Romano E., Tippetts, S., Blackman, K., and Voas, R., 2006. Language, income, education and alcohol-related fatal motor vehicle crashes. Journal of Ethnicity in Substance Abuse 5(2), pp. 119-137.

Root, E. D., Meyer, R. E., and Emch, M., 2011. Socioeconomic context and gastroschisis: Exploring associations at various geographic scales. Social Science and Medicine, 72(4), pp. 625-633.

Root, E. D., 2012. Moving neighborhoods and health research forward: using geographic methods to examine the role of spatial scale in neighborhood effects on health. Annals of the Association of American Geographers, 102(5), pp. 986-995.

Ross, C. E., 2000. Neighborhood disadvantage and adult depression. Journal of Health and Social Behavior, 41(2), pp. 177-187.

Santos, A., McGuckin, N., Nakamoto, H. Y., Gray, D., and Liss, S., 2011. Summary of Travel Trends: 2009 National Household Travel Survey, U.S. Department of Transportation. FHWA-PL-11-022.

Sayed, T., Rodriguez, F., 1999. Accident prediction models for urban unsignalized intersections in British Columbia, Transportation Research Record 1665, pp. 93-99.

Schwartz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), pp. 461-464.

Shankar, V., Mannering, and F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural accident frequencies. Accident Analysis and Prevention 27(3), pp. 371–389.

Shankar, V. N., Albin, R. B., Milton, J. C., and Mannering, F. L., 1998. Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using random effects negative binomial model. Transportation Research Record 1635, pp. 44–48.

Siddiqui, C., 2009. Macroscopic Traffic safety analysis based on trip generation characteristics (Master's thesis). Retrieved Dec 17, from http://purl.fcla.edu/fcla/etd/CFE0002871.

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

Siddiqui, C., and Abdel-Aty, M. A., 2012, On the nature of modeling boundary pedestrian crashes at zones. Presented in Transportation Research Board 91st Annual Meeting.

Smith, G. D., Neation, J. D., Wentworth, D., Stamler, R., and Stamler, J., 1996. Socioeconomic differentials in mortality risk among men screened for the multiple risk factor intervention trial: I. white men. American Journal of Public Health, 86 (4), pp. 486-496.

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

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

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D., 2003. WinBUGS user manual. Cambridge: MRC Biostatistics Unit. United Kingdom.

Stamatiadis, N., and Puccini, G., 2000. Socioeconomic descriptors of fatal crash rates in the southeast USA. Injury Control and Safety Promotion 7(3), pp. 165-173.

Sundquist, K., M. Winkleby, H. Ahlen, and S. E. Johansson. neighborhood socioeconomic environment and incidence of coronary heart disease: a follow-up study of 25,319 women and men in Sweden. America Journal of Epidemiology, 159 (7), 2004, pp. 655-662.

Thomas, I., 1996. Spatial data aggregation: exploratory analysis of road accidents. Accident Analysis and Prevention 28 (2), pp. 251-264.

Tighe, S., Li, N., Falls, L. C., and Haas, R., 2000. Incorporating road safety into pavement management. Transportation Research Record: Journal of the Transportation Research Board, 1699(1), pp. 1-10.

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics, 31(3), pp. 221-229.

Ukkusuri, S., Hasan, S., and Aziz, H. M. A., 2011. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transportation Research Record: Journal of the Transportation Research Board 2237, pp. 98-106.

U.S. Census Bureau, 1994. Geographic areas reference manual, U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau.

U.S. Census Bureau, 2011. 2010 Census traffic analysis zone program MAF/TIGER partnership software participant guidelines, U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau.

Ukkusuri, S., Hasan S., and Aziz, M. H. A., 2011. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency, Transportation Research Record 2237, pp. 98-106.

Viegas, J. M., Martinez, L. M., and Siva, E. A., 2009. Effects of the modifiable areal unit problem on the delineation of traffic analysis zones. Environment and Planning B: Planning and Design 36 (4), pp. 625-643.

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

Wang, X., Jin, Y., Abdel-Aty, M., Tremont, P.J., and Chen X., 2012. Macrolevel model development for safety assessment of road network structures. Transportation Research Record: Journal of the Transportation Research Board 2280, pp. 100-109.

Wang, Y., and Kockelman, K., 2013. A conditional autoregressive model for spatial analysis of pedestrian crash counts across neighborhoods. Presented in Transportation Research Board 92nd Annual Meeting.

Washington, S., 2006. Incorporating safety into long-range transportation planning. National Cooperative Highway Research Program Report 546. Transportation Research Board.

Washington, S., van Schalkwyk, I., You, D., Shin, K., and Samuelson, J. P., 2010. PLANSAFE: forecasting the safety impacts of socio-demographic changes and safety countermeasures. National Cooperative Highway Research Program 8-44(2). Transportation Research Board.

Whittam, K. P., Dwyer, W. O., Simpson, P. W., and Leeming, F. C., 2006. Effectiveness of a media campaign to reduce. Journal of Applied Social Psychology, 36(3), pp. 614-628.

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

Winkelman, R., 2003. Econometric analysis of count data, fourth ed. Springer, New York.

Ye, F., and Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models. Transportation Research Record 2241, pp. 51-58.

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

You, J., Nedović-Budić, Z., and Kim, T. J., 1997. A GIS-based traffic analysis zone design: technique. Transportation Planning and Technology, Vol. 21, pp. 45-68.

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

Zhang, M., and Kukadia, N., 2005. Metrics of urban form and the modifiable areal unit problem. Transportation Research Record: Journal of the Transportation Research Board 1902, pp. 71-79.