2020

Econometric Frameworks for Multivariate Models: Application to Crash Frequency Analysis

Tanmoy Bhowmik
University of Central Florida

ECONOMETRIC FRAMEWORK FOR MULTIVARIATE MODEL: APPLICATION TO CRASH FREQUENCY ANALYSIS

by

TANMOY BHOWMIK
B.Sc. Bangladesh University of Engineering and Technology, 2014
M.Sc. University of Central Florida, 2018

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctoral 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
2020

Major Professor: Naveen Eluru

ABSTRACT

Econometric crash frequency models are a major analytical tool employed for examining the critical factors influencing crash occurrence. However, there are several methodological challenges associated with existing models suggesting a continual need to develop advanced econometric framework to address these gaps. The current dissertation contributes towards addressing the methodological challenges in crash frequency analysis for analyzing multiple crash frequency variables for the same study unit by proposing advanced econometric approaches. The first part of the dissertation contributes to safety literature by conducting a comparison exercise between the two major streams of multivariate approaches - (1) simulation-based approach and (2) analytical closed form approach - for analyzing the crash counts considering different crash types. In the second part of the dissertation, we propose an alternative and mathematically simpler approach for analyzing multiple crash frequency variables for the same study unit by recasting a multivariate distributional problem as a repeated measures univariate problem. The recasting allows us to estimate parsimonious model systems thus improving parameter estimation efficiency. The third part of the dissertation contributes to burgeoning econometric and safety literature by developing a joint modeling approach that can accommodate for several dependent variables within a parsimonious structure. By recasting the analysis levels for dependent variables, the proposed approach allows for flexible consideration of crashes by type and severity within a single framework. The final part of the dissertation contributes to literature on crash frequency analysis by accommodating population heterogeneity in the impact of exogenous variables. The empirical analysis in this dissertation is based on traffic analysis zone (TAZ) level crash count data for both motorized and non-motorized crashes from Central Florida for the year 2016.

Keywords: Multivariate approach, crash types, crash severities, panel univariate model, Fractional split model, latent class panel mixed model, unobserved heterogeneity, population heterogeneity

ACKNOWLEDGMENTS

The completion of undertaking could not have been possible without the participation and assistance of so many people whose names may not all be enumerated. Their contributions are sincerely appreciated and gratefully acknowledged. However, I would like to express my deep appreciation and indebtedness particularly to the following:

I would like to express my deep sense of gratitude to my honorable Professor, Dr. Naveen Eluru. His guidance, resourceful insights and wisdom guided and directed me to the way to complete my work in time. In spite of his heavy engagements, he managed to associate with my work in harmony with the intermittent need of solving different problems.

I would also like to owe my deepest gratitude to Dr. Shamsunnahar Yasmin for her valuable advice, continuous guidance, encouragement and support. Words cannot express the gratitude I have for the profound impact you have had on my development both professionally and personally. This dissertation would not have been possible without your help.

My sincere gratitude to my family for their constant encouragement; especially my parents and my wife (Kanta). Without their support and inspiration, I would not be able to come this far and accomplish my achievements. Further, it will be ungrateful of me if I don't talk about my six best friends in my life including Arafat, Sourav, Arpan, Shuva, Tarek and Shammya. I always believe one thing that making many friends in one year is relatively easy but maintain one friend for many years is very tough. I feel very lucky that I got six such friends in my life.

Finally, I would also like to gratefully acknowledge Signal Four Analytics (S4A) and Florida Department of Transportation (FDOT) for providing access to Florida crash and geospatial

data. Above all, thanks to the Almighty God who made all these possible and gave me the strength to grow in my life even with many difficulties and overcome it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Background

The negative consequences of road traffic crashes have a significant impact on the emotional and financial well-being of the society. In the United States, annually motor vehicle crashes are responsible for more than 33,000 deaths and cost approximately $230 billion to the economy (GHSA, 2009; National Highway Traffic Safety Administration (NHTSA), 2013). According to the Global Status Report on Road Safety (World Health Organization, 2015) traffic crashes are likely to become the seventh leading cause of death in 2030 if adequate countermeasures are not adopted. In addition to the alarmingly high number of fatalities, there are multiple worrying trends within these numbers. The increase in the number of fatalities year over year for 2015 and 2016 represent the two largest year over year increases over last three decades. Further, in 2016, the percentage of non-motorized road user fatalities as a proportion of total fatalities have increased. Given the impact of road traffic crashes on the society, it is not surprising that safety researchers are continually investigating approaches for crash occurrence reduction and crash consequence mitigation. In this research, we limit ourselves to approaches dealing with crash occurrence reduction. Econometric crash prediction models are typically employed for examining crash counts either at the micro (intersection or segment) or the macro-level (county or traffic analysis zone). The micro-level analysis aims to suggest specific geometric design and/or engineering solutions to reduce the number of crashes for the examined road entities while the macro-level studies are useful from a transportation planning perspective providing regional hotspot identification and remedial solutions. The various crash frequency dimensions explored in existing literature include total crashes, crashes by severity, crashes by collision type and crashes by vehicle

type for a spatial unit over a given time period (Abdel-Aty et al., 2005; Lee et al., 2015; Wang et al., 2017).

In recent decades, substantial progress in analysing crash frequency models has been made. Earlier research efforts typically adopted a univariate framework to study a single crash frequency variable (such as total crashes) or multiple crash frequency variables (such as crash frequency by injury severity). Univariate approaches are not appropriate for modeling multiple dependent variables for the same observational unit as these approaches do not account for common unobserved heterogeneity affecting the various dependent variables (see (Mannering et al., 2016) for a detailed review). Recognizing this drawback, several research efforts in recent years have been conducted to accommodate for the potential dependency across multiple dependent variables for each observational unit (Anastasopoulos, 2016; Mannering et al., 2016; Nashad et al., 2016). In these multivariate approaches, propensity equations for multiple dependent variables are developed to accommodate for the impact of observed factors. These propensity equations traditionally take the form of a negative binomial or log-normal formulation. However, there are still several methodological challenges associated with such existing models suggesting continual needs to develop advanced econometric framework to address these gaps.

1.2 Motivation for The Study

The multivariate approaches can broadly be classified along two major streams: (1) simulation-based approaches and (2) analytically closed-form based approaches. The main difference between these two streams lies in how the dependency across dimensions is captured. In simulation-based approaches, the different propensities are correlated by generating a common error term across dimensions. For each realization of the common error term, the likelihood function (or posterior probability in Bayesian regime) is computed. However, given the inherently unobserved nature of

the error term, an appropriate distributional assumption is necessary to generate a population function. For this reason, multiple error term draws are generated, and the likelihood function values are averaged across these repetitions. The accuracy of the approach is affected by number of dimensions as well as number of draws considered for the function evaluation. Further, the stability of the variance-covariance matrix is often sensitive to model specification and number of simulation draws (see (Bhat, 2011) for a discussion). In closed-form based approaches, the propensity equations for frequency dimensions are tied together by analytical multivariate distributional assumptions. For example, the different propensity error terms are assumed to follow a multivariate distribution or a more general copula distribution. Thus, whenever permissible, such model formulation yields an analytical formula for the probability computation (Bhat and Eluru, 2009; Nashad et al., 2016). These models can be estimated using traditional maximum likelihood approaches. In some cases, where such formulas are of very high dimensions they might not be analytically tractable. In this case, an alternative approach that approximates the analytical probability is adopted. A commonly used such approximation approach involves composite maximum likelihood frameworks (Bhat, 2014, 2011; Narayanamoorthy et al., 2013). The first task of the research effort is focused on comparing the performance of these two streams of multivariate approaches (simulation and analytical). In our analysis, the comparison is undertaken with the univariate models following negative binomial model structure.

Despite the prevalence of multivariate approaches in safety literature, there are several challenges and gaps associated with such frameworks. In multivariate approaches (both simulation and closed-form streams), a separate crash propensity equation is adopted for each crash type. Thus, if there are D dependent variables and K independent variables, the order of observed parameters estimated in the model structure is $D*K$. With increasing dimension of D, the number

of parameters to be estimated increases rapidly. Thus, in models with D >3, the number of parameters to be estimated are prohibitively high. For example, consider a case of crash frequency for four crash types at an intersection (rear-end, side-swipe, angle and non-motorized). In the univariate models, for each of the crash types, Annual Average Daily Traffic (AADT) is likely to have a statistically significant impact. So, the typical multivariate model estimates 4 parameters for AADT. However, it is possible that the impact of AADT on side-swipe and angle crashes is not statistically different. Testing this is not straightforward in the multivariate model structure. The analyst will need to modify the model estimation code to restrict the parameters across the side-swipe and angle univariate models to be the same. Subsequently, the restricted model version data fit must be compared with the data fit of the unrestricted version using log-likelihood ratio (LR) test. Based on the result, the analyst can conclude if AADT does offer different impacts for side-swipe and angle crash profiles. Given the additional burden of these steps, the models employed in safety literature typically ignore if the variable impacts are really different across crash type propensities. The result is an ill-specified model structure with too many parameters. To be sure, the model estimates thus obtained are not incorrect. However, the estimation process could become inefficient particularly when sample sizes for crash frequency are small (<1000). The sample sizes for micro-level analysis can typically vary from 200-500 and the number of total parameters estimated has an impact of model estimation efficiency. Further, in simulation-based multivariate approaches, the influence of unobserved factors is typically accommodated as random effects and correlations across dimensions. The random effects accommodate for the influence of unobserved factors affecting crash propensity within the dimension. The correlations account for the influence of unobserved factors affecting multiple dependent variables. These effects require simulation for parameter estimation. The complexity of the model estimation is dependent on the

number of unobserved parameters estimated. With higher dimensions, the model estimation infrastructure can get computationally demanding (while not unmanageable with latest computing power). In our research, we propose to address these challenges by recasting the multivariate crash frequency modeling problem as a pooled univariate crash frequency (with unobserved heterogeneity accommodated) analysis problem.

Further, in multivariate count regression approaches described above, the impact of exogenous variables is quantified through the propensity component of count models. While several research efforts have developed multivariate crash frequency models for a small number of dimensions (such as 5); there is limited adoption of multivariate approaches for count variables in the presence of larger number of dependent variables (say greater than 15). As a result, there is limited adoption of research modeling crash severity frequency considering different crash types For example, consider the development of crash frequency models by crash type (say N types) and severity level (say K levels). In the currently employed approaches, the number of crash propensity equations to be estimated will be N*K. While the estimation of $N * K$ univariate model systems is repetitive, it is still feasible. However, accommodating for unobserved heterogeneity with a large number of dependent variables is substantially challenging. The probability evaluation with high dimensional integrals is potentially affected by several challenges including - requirements of generating high dimensionality of random numbers, empirical identification issues due to relatively flat objective functions in larger dimensions and longer computational run times. Furthermore, the stability of the variance-covariance matrix is often sensitive to model specification and number of simulation draws. In this context, the current dissertation contributes to burgeoning econometric and safety literature by developing a joint modeling approach that can accommodate for several dependent variables within a parsimonious structure.

1.3 Objective of the Dissertation

The current dissertation contributes towards addressing the methodological challenges in crash frequency analysis for analyzing multiple crash frequency variables for the same study unit by proposing advanced econometric approaches. The empirical analysis is based on traffic analysis zone (TAZ) level crash count data for both motorized and non-motorized crashes from Central Florida for the year 2016. A comprehensive set of exogenous variables including roadway, built environment, land-use, traffic, socio-demographic and spatial spillover characteristics are considered for the analysis. The proposed contributions are organized along four objectives that discussed in detail in subsequent paragraphs.

The first objective of the dissertation is focused on addressing the following question: which framework performs better in capturing potential correlation across multiple dependent variables in the current study context? Hence, we conduct a comparison exercise between the two major streams of multivariate approaches for analyzing the crash counts considering different crash types in the fast task of the dissertation. In safety literature, there are two ways to incorporate the potential correlation between multiple crash frequency variables: (1) simulation-based approach and (2) analytical closed form approach. The main difference between these two streams lies in how the dependency across dimensions is captured. However, so far there has not been a comprehensive comparison exercise between these two regimes. To that extent, the research effort proposed a comparison between the simulation-based multivariate model and copula based closed-form approach to analyze zonal level crash counts for different crash types. Further, the research builds on earlier copula based models by incorporating random parameters thus proposing a combination approach to incorporating unobserved heterogeneity. Within the proposed combination copula model, the empirical analysis involves estimation of count models using four

different copula structures which cover a wide range of dependency structures, including radial symmetry and asymmetry, and asymptotic tail independence and dependence.

The second objective of the dissertation is focused on addressing the following question: do we really need multivariate approaches for modelling multiple dependent variables? In this context, we propose an alternative and mathematically simpler approach for analyzing multiple crash frequency variables for the same study unit. Prior to presenting our alternative approach, challenges with the current simulation-based multivariate approaches in estimating observed and unobserved variable effects are discussed. Traditionally, simulation-based approaches are employed in crash frequency analysis for multiple crash frequency variables. However, in simulation-based models, the model estimation infrastructure can get computationally demanding with higher dimensions (while not unmanageable with latest computing power). Towards addressing these challenges, the proposed research presents an alternative formulation to analyze multiple crash frequency variables by recasting a multivariate distributional problem as a repeated measures univariate problem. To elaborate, instead of considering the crash frequency by crash type as a multivariate distribution, we represent it as repeated measures of crash frequency while recognizing that each repetition represents a different crash type. Thus, in this process we cast a multivariate distribution as a univariate distribution with repeated measures. The recasting allows us to estimate parsimonious model systems thus improving parameter estimation efficiency. Specifically, we employed a simpler panel random parameter based univariate model framework to analyze zonal level crash counts for different crash types. The performance of the proposed framework is then compared with the performance of the random parameter multivariate negative binomial model (RPMNB) using a host of metrics for estimation and hold-out sample.

The third objective of the dissertation is focused on addressing the following concern: can we develop a single framework for modelling zonal level crash severity counts across different crash types? In this context, the current objective contributes to burgeoning econometric and safety literature by developing a joint modeling approach that can accommodate for several dependent variables within a parsimonious structure. By recasting the analysis levels for dependent variables, the proposed approach allows for flexible consideration of crashes by type and severity within a single framework. Despite the recognition of distinct injury severity profiles across different crash types, there is limited adoption of research modeling severity frequency by crash types. The main challenge is with the number of dependent variables as accommodating unobserved heterogeneity for such large number of dimensions is substantially burdensome. In this context, we employ a Panel mixed Negative Binomial- Generalized Ordered Probit Fractional Spilt (PMNB-GOPFS) model where the first component (NB) accommodates for crash frequency by crash type and the later component (GOPFS) studies the fraction of severity outcome for different crash types. The dimension of the dependent variables analyzed is 24 $[(6 * 4)$ from 6 crash types and 4 severity levels). Further, a number of correlation terms are tested in the current research effort including: 1) common unobserved factors simultaneously affecting crash counts of different crash types ; 2) common unobserved factors simultaneously affecting crash severity proportions of different crash types ; and 3) common unobserved factors that simultaneously impact crash counts and severity proportions by different crash types. The analysis is further augmented by undertaking a prediction exercise using the final model parameter estimates for estimation and hold-out samples.

The final objective (fourth) of the dissertation focused on addressing the following issue: does the effect of observed and unobserved variables vary across the population or not? In this context, the current objective contributes to literature on crash frequency analysis by

accommodating population heterogeneity in the impact of exogenous variables. In conventional count models, the impact of exogenous factors is restricted to be the same across the entire region. However, it is possible that the influence of exogenous factors might vary across different TAZs. Ignoring such heterogeneous impact of variables might result in incorrect coefficient estimates. To that extent, the research effort proposes a latent segmentation based count model to capture the potential variation in the impact of exogenous variables. Specifically, we will formulate and estimate a latent segmentation based Negative Binomial (NB) to study the zonal level crash counts across different crash types.

1.4 Outline of the Dissertation

The remainder of the research proposal is divided into seven chapters which shows how each chapter position the current research effort within the larger context of the safety literature. From chapter three to six, the problem in context, an exhaustive literature review, limitation of earlier research, econometric framework adopted in the study and estimation results are discussed in detail to illustrate how each objective contributes in safety literature.

Chapter two discusses a detailed summary of the study area, data source, dependent and exogenous variables considered for the analysis. The research considers the Central Florida region which includes 4,747 traffic analysis zones (TAZs). The study is focused on crashes involving both motor vehicles and non-motorists at a zonal level for the year 2016. The data are compiled from Florida Department of Transportation (FDOT), Crash Analysis Reporting Systems (CARS) and Signal Four Analytics (S4A) databases. A host of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics are considered for the current research effort. Information about the variables are gathered from FDOT Transportation

Statistics Division, US Census Bureau, American Community Survey and Florida Geographic Data Library databases.

Chapter three contributes to objective one by comparing the performance of the simulation-based framework with closed form copula-based frameworks. The empirical analysis is based on the traffic analysis zone (TAZ) level crash count data for both motorized and non-motorized crashes from Central Florida for the year 2016. The crash data for 4,747 TAZs were sorted into the following four categories: (1) motorized intersection crashes, (2) motorized road segment crashes, (3) motorized off-road crashes and (4) non-motorized crashes. Using the four crash categories defined, we compare the performance of the random parameter multivariate negative binomial model with copula based negative parameter model. Within the copula framework, we estimate models for four copula structures: (1) Frank, (2) Gumbel, (3) Clayton and (4) Joe to cover a wide range of dependency structures, including radial symmetry and asymmetry, and asymptotic tail independence and dependence. The model frameworks are compared based on statistical fit and a host of comparison metrics for estimation sample and hold-out sample. Finally, the applicability of the model for hot zone identification is illustrated by generating plots identifying hot and cold zones by crash type in the Central Florida region.

Chapter four contributes to objective two by suggesting an alternative and mathematically simpler approach for analyzing multiple crash frequency variables for the same study unit. Basically, the proposed model attempts to contribute to simulation-based multivariate approaches by altering how the multiple dependent variables are analyzed. The proposed recasts a multivariate distributional problem as a repeated measure univariate problem. Prior to presenting our alternative approach, challenges with the current simulation-based multivariate approaches in estimating observed and unobserved variable effects are discussed. Specifically, we employed a simpler panel

random parameter based univariate model framework to analyze zonal level crash counts for different crash types as well as incorporating the presence of unobserved heterogeneity across crash types. The analysis is conducted using the zonal level crash records from Central Florida for the year 2016 considering a comprehensive set of exogenous variables. Further, the study evaluates the performance of the proposed approach by undertaking a comparison exercise with the traditional random parameter multivariate negative binomial model.

Chapter five contributes to objective three by developing a joint modeling approach that can accommodate for several dependent variables within a parsimonious structure. Specifically, we employ a Panel Mixed Negative Binomial- Generalized Ordered Probit Fractional Spilt (PM-NB-GOPFS) model where the first component (NB) accommodates for crash frequency by crash type and the later component (GOPFS) studies the fraction of severity outcome for different crash types. A number of correlation terms are tested in the current research effort including: 1) common unobserved factors simultaneously affecting crash counts of different crash types ; 2) common unobserved factors simultaneously affecting crash severity proportions of different crash types ; and 3) common unobserved factors that simultaneously impact crash counts and severity proportions by different crash types. The empirical analysis was conducted using the zonal level crash count data for the year 2016 from Central Florida while considering a comprehensive set of exogenous variables. The analysis is further augmented by undertaking a prediction exercise using the final model parameter estimates for estimation and hold-out samples.

Chapter six contributes to objective four analysis by accommodating population heterogeneity in the impact of exogenous variables. In conventional count models, the impact of exogenous factors is restricted to be the same across the entire region. However, it is possible that the influence of exogenous factors might vary across different TAZs. Ignoring such heterogeneous

impact of variables might result in incorrect coefficient estimates. To that extent, the research effort proposes a latent segmentation based count model to capture the potential variation in the impact of exogenous variables. Specifically, we will formulate and estimate a latent segmentation based Negative Binomial (NB) to study the zonal level crash counts across different crash types.

Chapter seven finally presents the summary of model findings and concluding thoughts followed by the contribution of the dissertation in the larger context of safety literature. The chapter also identifies limitations of the current dissertation and future directions of research.

CHAPTER 2: DATA PREPARATION

The previous chapter presented a summary of the objectives of the current dissertation. In this chapter, we briefly provide the details of the study area along with the information regarding the data source, dependent variables and exogenous attributes.

2.1 Study Area

Our study areas include the Central Florida region. Specifically, the research considers the region defined for Central Florida Regional Planning Model version 6.0 (CFRPM 6.0). The study area includes 4,747 traffic analysis zones (TAZs). Boundary of the study area encompasses nine counties (Brevard, Flagler, Lake, Marion, Orange, Osceola, Seminole, Sumter and Volusia) within District 5, Polk county within District 1 and part of Indian River county in District 4 of Florida Department of Transportation (FDOT). The location study area along with the zonal boundaries are shown in Figure 2.1.

2.2 Data Source

The study is focused on crashes involving both motor vehicles and non-motorists at a zonal level for the year 2016. The data are compiled from Florida Department of Transportation (FDOT), Crash Analysis Reporting Systems (CARS) and Signal Four Analytics (S4A) databases. CAR and S4A are long and short forms of crash reports in the State of Florida, respectively. The Long Form crash report is used mostly to give focus on injurious accident or crash concerning felonious activities (such as hit-and-run or driving under influence) whereas Short Form depicts the reports

based on all other traffic crashes. Both forms of reports are integrated to get a complete view on road crashes and having the objective to use it for the better understanding of current analysis.

2.3 Dependent Variable

2.3.1 Exploration of Analytical, Simulation and Combined Model Structures

At first, the crash data were sorted into two classes based on the road user group: motorist and non-motorist; further, within the motorized group, the records are classified into three categories based on the location of the crash: intersection, road segment and off-road. All the crash records are aggregated at a TAZ level using the Geographic Information System (GIS). A total of 112,376 motorized and 3,413 non-motorized crashes were reported in the Central Florida for the year 2016. Figure 2.2 describes the overall summary of all crash types in Central Florida for the year 2016 in terms of percentage For the motorists, road segment was found to be most unsafe place (48.5%) followed by intersection (38.9%). Table 2.1 presents the summary statistics of crash type variables. Further, we have partitioned the zonal level records into two datasets: 1) 3,800 TAZs for model estimation and 2) 947 TAZs for validation analysis.

2.3.2 Panel Mixed Approach to Modeling Crash Frequency by Crash Types

The study is focused on crashes involving both motor vehicles and non-motorists at a zonal level for the year 2016. At first, the crash data were sorted into two classes based on the road user group: motorist and non-motorist; within the motorized group, the records are further classified into five categories based on the manner of crash: rear-end, angular, sideswipe, all single vehicle and other multiple vehicle crashes. Based on the crash records, crash of different types are combined together as one category: left-turn, right-turn and angular crashes within angular class; off-road, rollover and other single vehicle in the all single vehicle category; and head-on and other multiple vehicle

crashes are in the other multiple vehicle crash types. All the crash records are aggregated at a TAZ level using the Geographic Information System (GIS). A total of 114,458 motorized and 3,413 non-motorized crashes were reported in the Central Florida for the year 2016. Figure 2.3 describes the overall summary of all crash types in Central Florida for the year 2016 in terms of percentage. Within the motorized crashes, rear-end is found to be the most prevalent one (44.09%) while sideswipe is less frequent with 10.82% among all other motorized crash types. Crash statistics at a zonal level for different types of crash are summarized in Table 2.2. From the total record, for the validation analysis, we set aside records from 932 TAZs and the remaining 3,815 TAZs are used for the estimation analysis.

2.3.3 Econometric Approach for Modeling Crash Counts by Crash Type and Severity

At first, the crash data were sorted into two classes based on the road user group: motorist and non-motorist; within the motorized group, the records are further classified into five categories based on the manner of crash: rear-end, angular, sideswipe, head-on and single vehicle crashes. Then for each crash types, crashes are further classified by injury severity levels such as fatal (K), incapacitating (A), non-incapacitating (B), possible injury (C), and property damage only (O) crashes. Based on crash records, fatal and incapacitating injuries are combined as one category and defined as severe injury. Finally, the crash records are aggregated at a zonal level and the corresponding severity proportions by crash type are as follows: (1) proportion of no injury (property damage only) crashes, (2) proportion of minor injury crashes, (3) proportion of non-incapacitating injury crashes, and (4) proportion of severe injury crashes.

The crash counts and severity outcome proportions for each crash type are presented in Figure 2.4. From the Figure 2.4, we can observe that number of no injury crashes has the highest proportion followed by proportion of minor injury crashes. Further, in terms of crash types, the

figure 2.4 shows that non-motorists are more prone to severe crashes whereas the injury outcomes are higher for motorists involved in head-on crashes. On the other hand, in approximately 84% and 72% sideswipe and rear-end crashes, respectively, the outcomes were no injury. The most commonly used approach of modeling severity frequency or proportion without considering crash type would result in an inaccurate aggregation. From the figure (2.4), it is evident that severity proportions by crash type vary significantly across crash types.

2.3.4 Accommodating Population Heterogeneity Within A Panel Model Framework

We used the zonal level crash counts of same six crash types as described in section 2.3.3.

2.4 Exogenous Variable Considered

A host of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics are considered for the current research effort. Information about the variables are gathered from FDOT Transportation Statistics Division, US Census Bureau, American Community Survey and Florida Geographic Data Library databases. In addition to crash records, explanatory attributes are also aggregated at a zonal level using the GIS. Roadway attributes included are road lengths for different functional class, proportion of rural and urban road, proportion of road with different number of lanes (1, 2, and 3 or more), number of intersections and signals, mean and variance of speed limit, length of road with different speed limit (≤40mph, 41-54mph and ≥55mph), average width of inside and outside shoulder, average width of bike lane and sidewalk. Land use attributes mainly provide the land use category information including area of urban, residential, industrial, institutional, recreational, office and land use mix while information about the number of business centers, commercial centers, schools, hospitals, recreational centers, restaurants and shopping centers are considered in the built

environment characteristics. Land use mix is defined as: $\left[\frac{-\sum_k(p_k(lnp_k))}{lnN}\right]$, where k is the category of land-use, p_k is the proportion of the developed land area devoted to a specific land-use *k*, *N* is the number of land-use categories in a STAZ. In our study, six land use types were considered including residential, park facilities, industrial, institutional, agricultural and office areas. Institutional land use refers to land uses that cater to community's social and educational needs (schools, town hall, police station) while park facilities refer to land used for recreational or entertainment purposes. The value of this index ranges from zero to one - zero (no mix) corresponds to a homogenous area characterized by single land use type and one to a perfectly heterogeneous mix).Further, for traffic characteristics, average annual daily traffic (AADT), average annual daily truck traffic (truck AADT), vehicle miles traveled (VMT), truck vehicle miles traveled (truck VMT) and proportion of heavy traffic are considered. In sociodemographic attributes, population and household density, proportion of means of transportation used by commuter for their work trips (car, transit, bike and walk) and proportion of household by vehicle ownership level (0, 1, 2, 3 and 4 or more) are included.

In case of aggregate level models, for any spatial unit, there is a possibility that crashes are more affected by the neighbouring units rather than the actual unit, specially for those crashes which occurred in the boundary region. Several research efforts have acknowledged the importance of spatial spillover effects (see (Aguero-Valverde and Jovanis, 2006; Cai et al., 2016; Quddus, 2008)). In safety literature, there are two ways to incorporate the effect of spatial effect: 1) Spatial error correlation and 2) Spatial spillover effect (see (Cai et al., 2016) for details). The current research effort follows the second method in which the dependency is captured through the observed attributes (Cai et al., 2016; Narayanamoorthy et al., 2013). For every zone, neighbouring zones are identified and based on the neighbouring zone, exogenous variables are

estimated (similar to the actual TAZ). For example, proportion of urban road in the actual TAZ is computed by taking the ratio of the length of urban road to the total road in that specific TAZ. In terms of spillover effect, for every TAZs, we have the neighbouring TAZs and based on that we take the sum of the urban and total road and estimate the proportion of urban road by taking the ratio of it. The reader would note that, targeted TAZs are not considered in the neighbouring TAZs. Across the dataset, the number of surrounding zones range from 1 to 21 with an average value of 6.43.

Table 2.3 summarizes sample characteristics of the explanatory variables with the appropriate definition considered for final model estimation along with the minimum, maximum and mean values at a zonal level. While we estimated spatial spill-over variables for all variables, we only present the variables that offered significant effects in the model. In estimating the model, several functional forms, combination of variables and interaction terms are considered and those that provides the best fit are retained in the final specification. The final specification of the model was based on removing the statistically insignificant variables in a systematic process based on 90% confidence level.

2.5 Summary

In this chapter, data source employed along with the data preparation procedure are discussed in detail. Moreover, descriptive statistics for both dependent and exogenous variables are provided. The next four chapters describe the four objectives of the current research effort and shows how it contributes to the safety literature on crash frequency analysis.

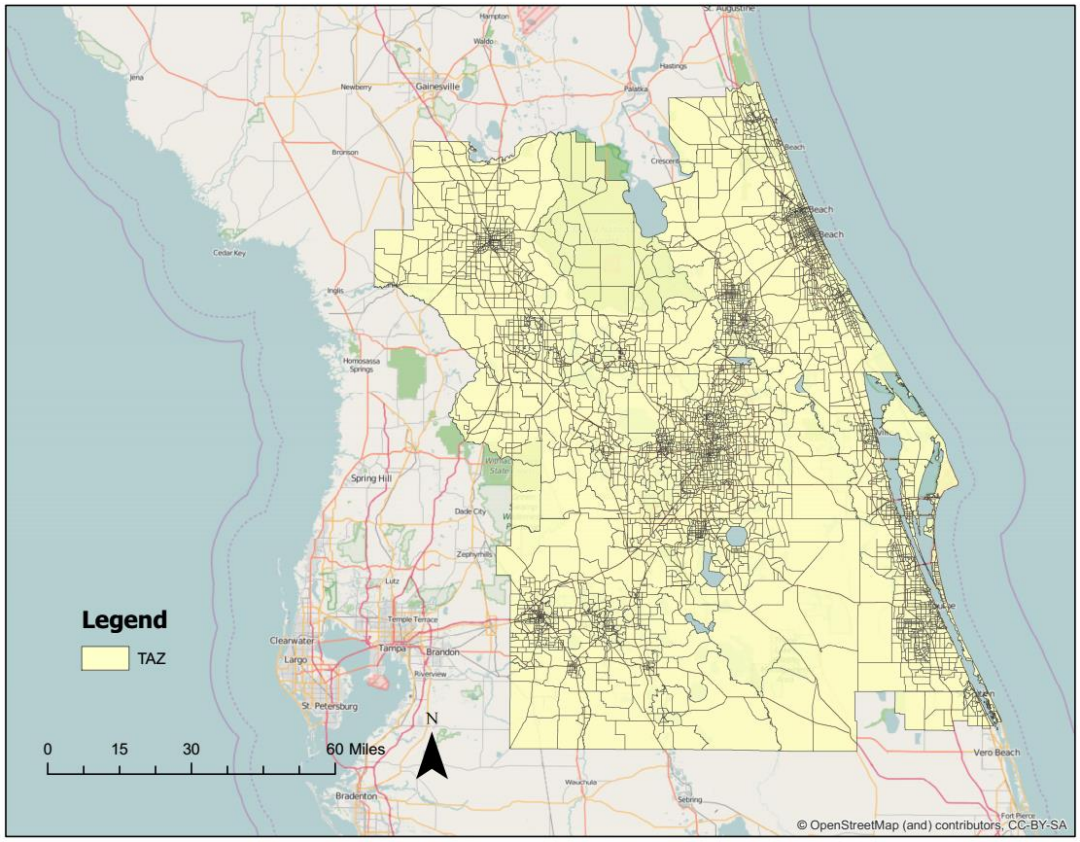

Figure 2.1 Location of Study Region

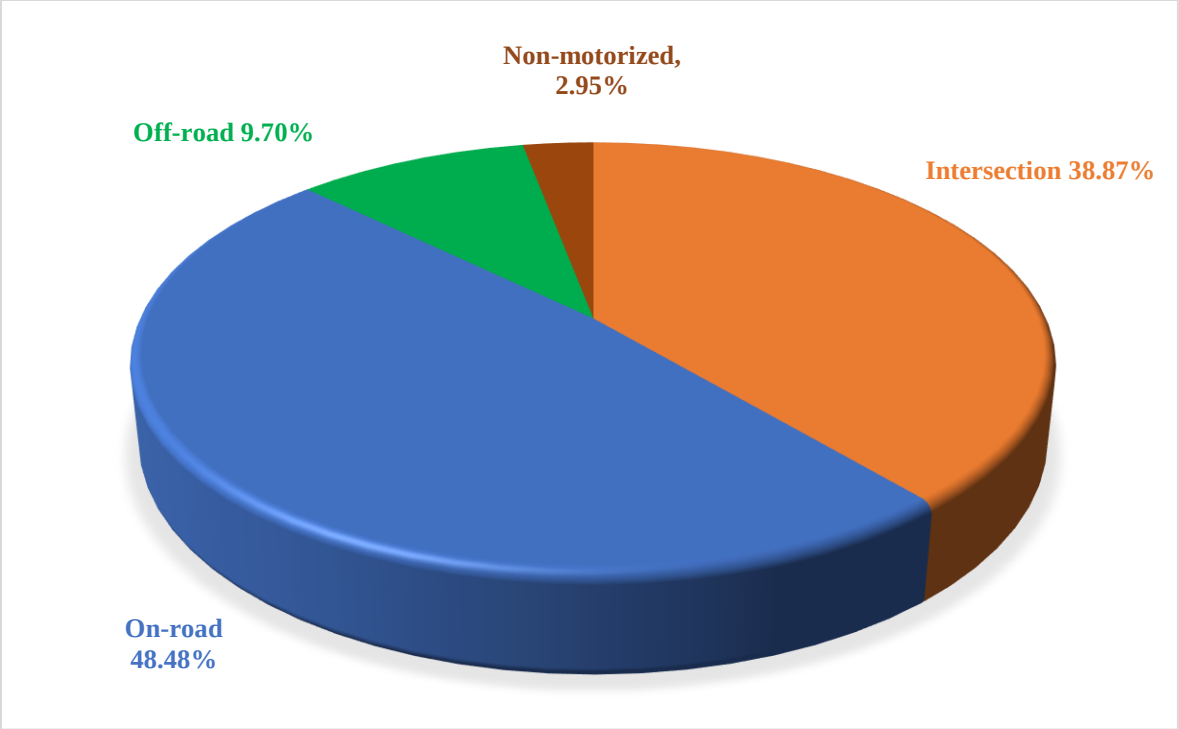

Figure 2.2 2016 Crashes by types (%) in Central Florida (Crash Location Wise)

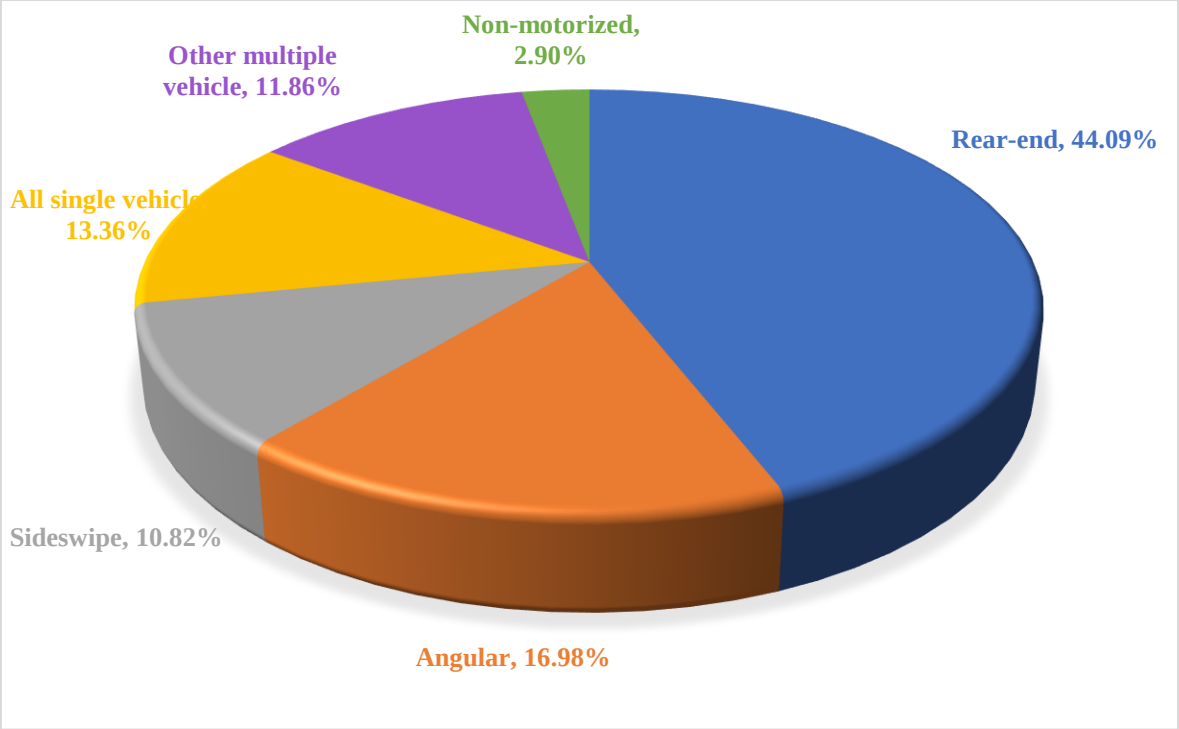

Figure 2.3 2016 Crashes (%) in Central Florida (Crash Type Wise)

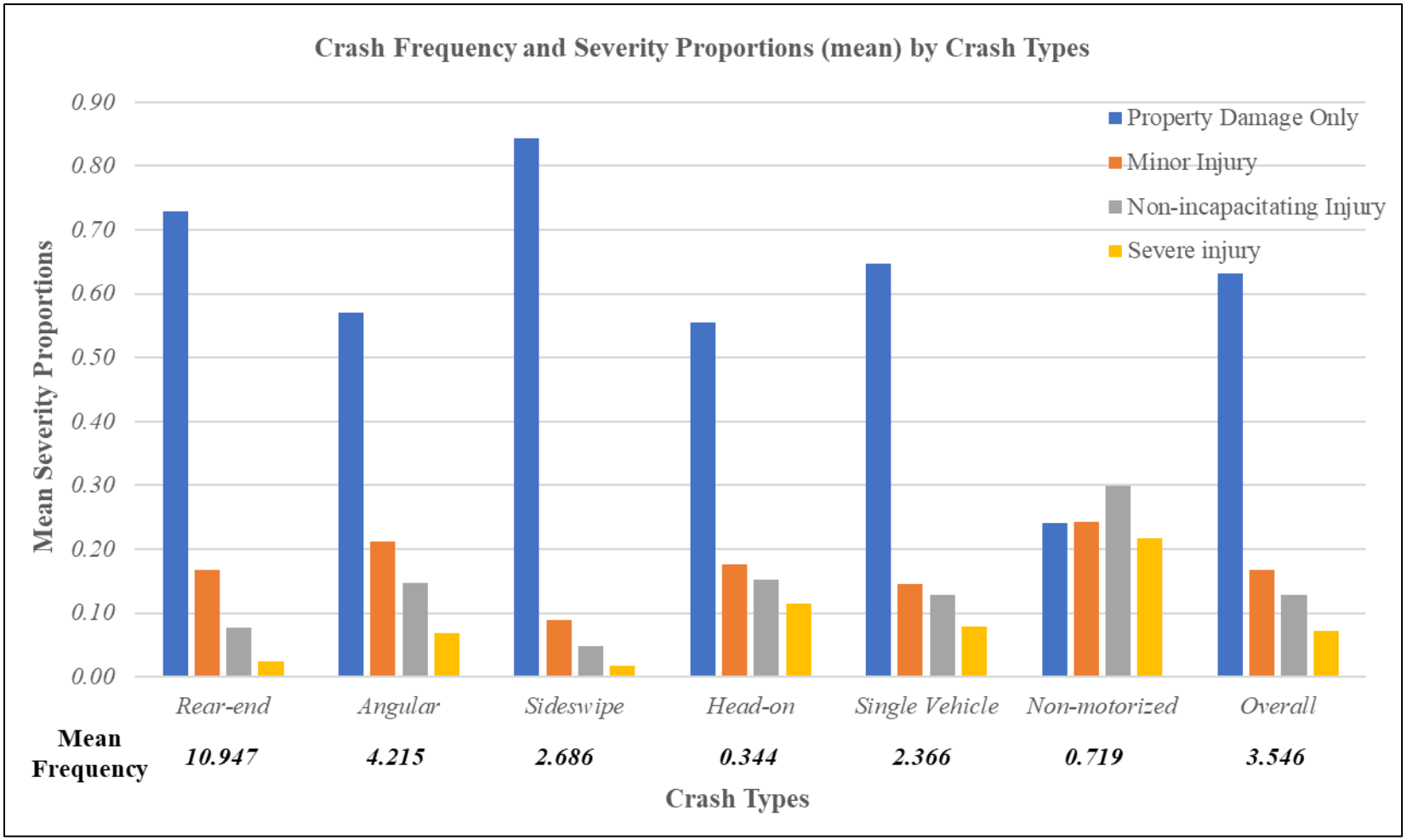

Figure 2.4 Crash Frequency and Severity Proportions (mean) by Crash Types

Table 2.1 Descriptive Statistics of Dependent Variables (Copula Approach)

Variable Names	Definition	Zones (N=4,747)			
		Minimum	Maximum	Mean	Standard Deviation
Motorized Intersection Crash	Total number of crashes occurred at or within the influence area of intersection in a TAZ	0.000	171.000	9.480	13.490
Motorized road segment Crash	Total number of crashes occurred on roadway segments and outside the influence area of intersection in a TAZ	0.000	283.000	11.826	20.700
Motorized Off-road Crash	Total number of crashes occurred outside the influence area of roadway in a TAZ	0.000	51.000	2.367	3.573
Non-motorized Crash	Total number of non-motorized (pedestrian and bicyclist) crash in a TAZ	0.000	12.000	0.719	1.318

Table 2.2 Descriptive Statistics of Dependent Variables (Panel Approach)

Variable Names	Definition	Zones (N=4,747)			
		Minimum	Maximum	Mean	Standard Deviation
Rear-end Crash (motorized)	Total number of rear-end crash (motorized) occurred in a TAZ	0.000	243.000	10.948	18.517
Angular Crash (motorized)	Total number of left turn, right turn and angular crash (motorized) occurred in a TAZ	0.000	104.000	4.216	6.817
Sideswipe Crash (motorized)	Total number of sideswipe crash (motorized) occurred in a TAZ	0.000	66.000	2.686	5.228
All Single Vehicle Crash (motorized)	Total number of off-road, rollover and other-single vehicle crash (motorized) occurred in a TAZ	0.000	62.000	3.317	4.480
Other-multiple Vehicle Crash (motorized)	Total number of head-on and other-multiple vehicle crash (motorized) occurred in a TAZ	0.000	112.000	2.945	4.549
Non-motorized Crash	Total number of non-motorized (pedestrian and bicycle) crash in a TAZ	0.000	12.000	0.719	1.318

Table 2.3 Summary Statistics of Exogenous Variables (Zonal Level)

Variables	Definition	Zonal (N=4,747)			
		Minimum	Maximum	Mean	Std. Deviation
Roadway Characteristic					
Proportion of rural road	(Rural road length/total road length)	0.000	1.000	0.121	0.309
Proportion of urban road	(Urban road length/total road length)	0.000	1.000	0.806	0.381
Proportion of arterial road	(Arterial road length/total road length)	0.000	1.000	00377	0.393
Number of Intersection	Ln (no of intersection)	0.000	4.682	1.921	1.053
Signal intensity	Total number of traffic signal per	0.000	1.000	0.038	0.096
Average speed limit	Ln (mean speed limit in mph)	0.000	4.248	3.228	1.279
Variance of speed limit	Ln (variance of speed limit in mph)	0.000	6.686	2.325	2.041
Average bike lane length	Ln (average length of bike lane in feet)	0.000	1.662	0.044	0.147
Average inside shoulder width	Ln (average inside shoulder width in feet)	0.000	2.650	0.288	0.445
Average outside shoulder width	Ln (average outside shoulder width in feet)	0.000	2.977	0.964	0.579
Average sidewalk width	Ln (average sidewalk width in feet)	0.000	2.977	0.964	0.579
Divided road length	Ln of (divided road length in meter)	0.000	1.547	0.037	0.096
Road ≥55mph	Proportion of road length greater than	0.000	1.000	0.088	0.174
Land-use Attributes					
Urban area	Ln (urban area+1) in acre	0.000	9.440	4.921	1.970
Recreational area	Ln (recreational area+1) in acre	0.000	9.814	0.470	1.408
Office area	Ln (office area+1) in acre	0.000	6.440	0.877	1.383
Residential area	Ln (residential area+1) in acre	0.000	8.131	3.811	2.075
Industrial area	Ln (industrial area+1) in acre	0.000	7.067	1.118	1.306
Institutional area	Ln (institutional area+1) in acre	0.000	6.617	1.946	1.589
Land use mix	Land use mix = $\left[\frac{-\sum_k (p_k (lnp_k))}{lnN}\right]$, where k is the category of land-use, p is the proportion of the developed land area for specific land-use, N is the number of land-use categories	0.000	0.946	0.369	0.221

Built Environment Characteristics					
No of business center	Z score[1]: No of business center	-0.138	19.664	0.000	1.000
No of commercial center	Z score: No of commercial center	-0.270	9.521	0.000	1.000
No of educational center	Z score: No of educational center	-0.487	11.610	0.000	1.000
No of recreational center	Z score: No of park and recreational center	-0.475	16.678	0.000	1.000
No of restaurant	Z score: No of restaurant	-0.464	11.021	0.000	1.000
No of shop	Z score: No of shopping center	-0.442	19.728	0.000	1.000
Traffic Characteristics					
VMT	Vehicle miles travelled	0.000	15.026	7.914	3.368
Truck VMT	Tuck vehicle miles traveled	0.000	13.049	3.474	2.864
Proportion of heavy	Total truck AADT/ Total AADT	0.000	0.369	0.068	0.046
Sociodemographic Characteristics					
Population density	Total population/Total area of TAZ in acre	0.000	21.293	2.364	2.233
household density	Total number of household/Total area of	0.000	8.556	0.902	0.878
Average TAZ income	Ln (Average TAZ income+1)	0.000	12.534	11.065	0.386
Proportion of commuter	Total number of commuter/total population	0.000	0.778	0.408	0.085
Non-motorist commuter	Ln (NMT means to work for a TAZ)	0.000	5.261	1.278	1.098
Proportion of senior people	Total number of people over 65 years/total population in TAZ	0.000	0.821	0.206	0.114
Proportion of African-American people	Total number of African-American people /total population in TAZ	0.000	0.969	0.142	0.159
Proportion of household with no vehicle	Number of household with no vehicle/total household	0.000	0.471	0.069	0.065
Spatial Spillover Effect					
Office area	Ln (∑office area+1) in acre in surrounding	0.000	7.670	2.849	1.869
Signal intensity	∑signal/∑intersection in neighbour's zone	0.000	1.000	0.042	0.050
Proportion of major road	(∑Major road length/∑total road length) in surrounding zones	0.000	1.000	0.619	0.249

[1] Z-score represents the standardized form of the actual variable.

Proportion of HH with no vehicle	∑household with 0 vehicle/∑household of neighbouring zones	0.000	0.347	0.067	0.054
Non-motorist commuter	(∑commuter by walk and cycle/∑population) of neighbouring zones	0.000	6.703	3.174	1.257
Average sidewalk width	Ln (average sidewalk width in feet) in surrounding zones	0.000	2.127	1.089	0.334

CHAPTER 3: EXPLORATION OF ANALYTICAL, SIMULATION AND COMBINED (ANALYTICAL+SIMULATION) MODEL STRUCTURES

The negative consequences of road traffic crashes have a significant impact on the emotional and financial well-being of the society. In the United States, annually motor vehicle crashes are responsible for more than 33,000 deaths and cost approximately $230 billion to the economy (GHSA, 2009; NHTSA, 2015). According to the *Global Status Report on Road Safety* (WHO, 2018) traffic crashes are likely to become the seventh leading cause of death in 2030 if adequate countermeasures are not adopted. Given the impact of road traffic crashes on the society, it is not surprising that safety researchers are continually investigating approaches for crash occurrence reduction and crash consequence mitigation. In this research, we limit ourselves to approaches dealing with crash occurrence reduction. Econometric crash prediction models are typically employed for examining crash counts either at the micro (intersection or segment) or the macro-level (county or traffic analysis zone). The micro-level analysis aims to suggest specific geometric design and/or engineering solutions to reduce the number of crashes for the examined road entities while the macro-level studies are useful from a transportation planning perspective providing regional hotspot identification and remedial solutions. The various crash frequency dimensions explored in existing literature include total crashes, crashes by severity, crashes by collision type and crashes by vehicle type for a spatial unit over a given time period (Abdel-Aty et al., 2005; Lee et al., 2015; Wang et al., 2017).

In recent decades, substantial progress in analysing crash frequency models has been made. Earlier research efforts typically adopted a univariate framework to study a single crash frequency variable (such as total crashes) or multiple crash frequency variables (such as crash frequency by

injury severity). Univariate approaches are not appropriate for modeling multiple dependent variables for the same observational unit as these approaches do not account for common unobserved heterogeneity affecting the various dependent variables (see (Mannering et al., 2016) for a detailed review). Recognizing this drawback, several research efforts in recent years have been conducted to accommodate for the potential dependency across multiple dependent variables for each observational unit (Anastasopoulos, 2016; Mannering et al., 2016; Nashad et al., 2016). In these multivariate approaches, propensity equations for multiple dependent variables are developed to accommodate for the impact of observed factors. These propensity equations traditionally take the form of a negative binomial or log-normal formulation. These multivariate approaches can broadly be classified along two major streams: (1) simulation-based approaches and (2) analytically closed-form based approaches.

The main difference between these two streams lies in how the dependency across dimensions is captured. In simulation-based approaches, the different propensities are correlated by generating a common error term across dimensions. For each realization of the common error term, the likelihood function (or posterior probability in Bayesian regime) is computed. However, given the inherently unobserved nature of the error term, an appropriate distributional assumption is necessary to generate a population function. For this reason, multiple error term draws are generated, and the likelihood function values are averaged across these repetitions. The accuracy of the approach is affected by number of dimensions as well as number of draws considered for the function evaluation. Further, the stability of the variance-covariance matrix is often sensitive to model specification and number of simulation draws (see (Bhat, 2011) for a discussion). In closed-form based approaches, the propensity equations for frequency dimensions are tied together by analytical multivariate distributional assumptions. For example, the different propensity error

terms are assumed to follow a multivariate distribution or a more general copula distribution. Thus, whenever permissible, such model formulation yields an analytical formula for the probability computation (Bhat and Eluru, 2009; Nashad et al., 2016; Soulez et al., 2007). These models can be estimated using traditional maximum likelihood approaches. In some cases, where such formulas are of very high dimensions they might not be analytically tractable. In this case, an alternative approach that approximates the analytical probability is adopted. A commonly used such approximation approach involves composite maximum likelihood frameworks (Bhat, 2014, 2011; Narayanamoorthy et al., 2013).

3.1 Earlier Research

A summary of research efforts from the two streams described above are presented in Table 3.1 with information on the study unit, methodological framework, estimation technique, dependent variables and the number of dimensions employed. From the table, several observations can be made. First, simulation approaches employ maximum simulated likelihood approach (MSL) in the classical framework and Markov Chain Monte Carlo (MCMC) approach in the Bayesian realm for model estimation. Second, within the simulated framework, various model structures developed include multivariate Poisson regression model, multivariate Poisson lognormal model, multinomial-generalized Poisson model, multivariate Poisson gamma mixture count model, multivariate Poisson lognormal spatial and/or temporal model, grouped random parameter multivariate spatial model, Integrated Nested Laplace Approximation Multivariate Poisson Lognormal model, Bayesian latent class flexible mixture multivariate model, flexible Bayesian semiparametric approach and multivariate random-parameters zero-inflated negative binomial model. Third, an alternative framework that builds on the fractional split model has also been identified as a credible alternative to the traditional multivariate approaches. Instead of using

propensity per dimension, exogenous variable affects all dependent variables through a unified mechanism thus offering a more parsimonious specification. Fourth, only a small number of studies – 3 studies to be precise - have employed the closed-form approach for developing multivariate models in crash frequency analysis. Fifth, it is important to recognize that the analytical approach based systems are geared toward accommodating for the influence of unobserved factors across multiple dependent variables. However, in these approaches, the influence of unobserved factors on the individual dependent variables in the form of random parameters are rarely considered. Finally, the various independent variables examined include roadway, traffic, land-use, sociodemographic and socioeconomic characteristics.

3.2 Current Study

From the literature review, it is evident that simulation-based approaches are more commonly employed in crash frequency analysis. The preponderance of simulation-based approaches can be attributed to advancements in simulation approaches and enhanced access to computing power. These simulation-based approaches accommodate for (1) common unobserved factors affecting each dependent variable by allowing for random parameters and (2) common unobserved factors affecting multiple dependent variables by allowing for correlations across dependent variables. More recently, closed-form copula-based approaches are suggested as a viable alternative to modeling crash frequency. The likelihood function, while analytically closed-form, is complicated in the copula regime. Given the analytical formulation these frameworks rely on maximum likelihood (as opposed to maximum simulated likelihood) and are less prone to error. However, in these approaches, unobserved heterogeneity in the form of random parameters is rarely considered as it will introduce simulation within a complex analytical formulation. To elaborate, current copula model systems assume that all the exogenous variables have the same influence on crash

count propensity across the entire population. However, in some cases, this assumption might be erroneous. For example, let us consider the effect of average sidewalk width on non-motorized crash counts. Increased sidewalk width is associated with higher pedestrian activity (exposure) and as a result possibly more crashes. However, at the same time, the presence of sidewalk provides additional safety to the non-motorists from colliding with a motorized vehicle. Also, the higher number of pedestrian and bicyclist on the road might make the drivers more familiar with pedestrian activity and thus more cautious in their driving behavior that potentially could result in a reduced number of non-motorized crashes. Therefore, the effect of sidewalk width could be different across the TAZs and it is useful to allow for the effect of sidewalk width on non-motorized crash counts to vary across TAZs by considering a distributional assumption across the TAZs. The proposed effort develops a random parameter copula model structure that builds an approach for employing an analytical multivariate model embedded within a simulation framework for crash frequency analysis. . Subsequently, we compare the performance of the proposed model (random parameter copula models) with the most commonly employed simulation-based approach and analytical closed-form copula models. To the best of authors' knowledge, this study is the first of its kind to incorporate attribute variability (random effect) effect within the copula framework. For the comparison exercise, a negative binomial kernel is employed across all model structures. The reader would note that the comparison exercise could be extended to other model structures in a straightforward fashion.

The empirical analysis is based on the traffic analysis zone (TAZ) level crash count data for both motorized and non-motorized crashes from Central Florida for the year 2016. The crash data for 4,747 TAZs were sorted into the following four categories: (1) motorized intersection crashes, (2) motorized road segment crashes, (3) motorized off-road crashes and (4) non-motorized

crashes. Using the four crash categories defined, we compare the performance of the random parameter multivariate negative binomial model with random parameter copula-based multivariate negative binomial model. Within the copula framework, we estimate models for four copula structures: (1) Frank, (2) Gumbel, (3) Clayton and (4) Joe. We examine the performance of these two frameworks in terms of model fit and prediction power for two datasets 1) estimation sample (records that are used for analysis - 3,800 TAZs) and 2) validation sample (set aside for validation analysis - 947 TAZs). In our models, we consider exogenous variables from roadway characteristics, land-use attributes, built environment characteristics, traffic characteristics, sociodemographic characteristics, and spatial spillover effects. The model comparison exercise is augmented with spatial representation of hot and cold zones by crash type for policy implications and prioritizations.

The rest of the chapter is organized as follows: The next section presents the methodological framework adopted in the analysis while the model findings are presented in section 3.4. The comparison exercise are offered in the section 3.5 followed by the spatial distribution in section 3.6. Finally, a summary of model findings and conclusions are presented in Section 3.7.

3.3 Econometric Framework

In this section, we briefly provide details of the model frameworks employed in our study. The model structure description order is as follows: (a) independent negative binomial model, (b) Simulation-Based Random Parameter Multivariate NB (RPMNB) Model, (c) Copula-Based Multivariate NB Model and (d) Copula-Based Random Parameter Multivariate NB Model. The mathematical frameworks build on simpler approaches whenever appropriate.

3.3.1 Independent Negative Binomial (NB) Model

Let us assume that i ($i = 1,2,3, \dots, N, N = 3{,}800$) be the index for TAZ. Let j be the index representing different crash type, where ($j = 1,2, \dots, J, J = 4$), the index j may take the values of motorized intersection (j =1), motorized road segment (j =2), motorized off-road (J =3) and non-motorized (j =4) crashes. Using these notations, the equation system for modeling crash count across different crash type j in the usual negative binomial (NB) formulation can be written as:

$$P(c_{ij}|\mu_{ij}, \alpha_j) = \frac{\Gamma\left(c_{ij} + \frac{1}{\alpha_j}\right)}{\Gamma(c_{ij} + 1)\Gamma\left(\frac{1}{\alpha_j}\right)}\left(\frac{1}{1 + \alpha_j\mu_{ij}}\right)^{\frac{1}{\alpha_j}}\left(1 - \frac{1}{1 + \alpha\mu_{ij}}\right)^{c_{ij}} \quad (1)$$

where, c_{ij} be the index for crash counts specific to crash type j occurring over a period of time in TAZ i. $P(c_{ij})$ is the probability that TAZ i has c_{ij} number of crashes for crash type j. $\Gamma(\cdot)$ is the gamma function, α_j is NB over dispersion parameter and μ_{ij} is the expected number of crashes occurring in TAZ i over a given time period for crash type j. Given this set up, the mathematical formulations of the econometric frameworks considered in the current study context is presented in this section.

With the NB probability expression as presented in equation 1, we can express μ_{ij} as a function of explanatory variables by using a log-link function as follows:

$$\mu_{ij} = E(c_{ij}|\boldsymbol{z}_{ij}) = exp((\boldsymbol{\delta}_j)\boldsymbol{z}_{ij} + \varepsilon_{ij}) \quad (2)$$

where, $\boldsymbol{z}_{ij}$ is a vector of explanatory variables associated with TAZ i and collision type j. $\boldsymbol{\delta}_j$ is a vector of coefficients to be estimated. ε_{ij} is a gamma distributed error term with mean 1 and variance α_j.

Thus, the likelihood function for the probability can be expressed as:

$$L_{i,j} = P(c_{ij}) \quad (3)$$

Finally, the log-likelihood function is:

$$LL_j = \sum_i Ln(L_i) \quad (4)$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 4.

3.3.2 Simulation-Based Random Parameter Multivariate NB (RPMNB) Model

The focus of RPMNB (referred as multivariate NB model in the following sections for simplicity) model is to examine number of crashes across different collision types jointly. As we consider four different crash types in the current analysis, in estimating RPMNB model, we examine four different NB models for four different collision types simultaneously. The expected crash counts TAZ i over a given time period for crash type j presented in equation 2 is updated in the RPMNB model as following:

$$\mu_{ij} = E(c_{ij}|\boldsymbol{z}_{ij}) = exp((\boldsymbol{\delta}_j + \boldsymbol{\zeta}_{ij})\boldsymbol{z}_{ij} + \varepsilon_{ij} + \eta_{ij}) \quad (5)$$

where, $\boldsymbol{\zeta}_{ij}$ is a vector of unobserved factors on crash count propensity associated with crash type j for TAZ i and its associated zonal characteristics, assumed to be a realization from standard normal distribution: $\boldsymbol{\zeta}_{ij} \sim N(0, \boldsymbol{\pi}_j{}^2)$. η_{ij} captures unobserved factors that simultaneously impact number of crashes across different crash types for TAZ i. Here it is important to note that the unobserved heterogeneity between total number of crashes across different crash types can vary across TAZs. Therefore, in the current study, the correlation parameter η_{ij} is parameterized as a function of observed attributes as follows:

$$\eta_{ij} = \boldsymbol{\gamma_j s}_{ij} \quad (6)$$

where, $\boldsymbol{s}_{ij}$ is a vector of exogenous variables, $\boldsymbol{\gamma}_j$ is a vector of unknown parameters to be estimated (including a constant). In the current analysis, the RPMNB model only allows for a positive correlation for total number of crashes across different crash types.

In examining the model structure of crash count across different crash types, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\zeta}$ and $\boldsymbol{\gamma}$ represented by $\boldsymbol{\Omega}$. In this framework, it is assumed that these elements are drawn from independent normal distributions: $\Omega \sim N(0, (\boldsymbol{\pi}_j^{\mathbf{2}}, \boldsymbol{\sigma}_j^2))$. Thus, conditional on $\boldsymbol{\Omega}$, the likelihood function for the joint probability can be expressed as:

$$L_i = \int_{\boldsymbol{\Omega}} \prod_{j=1}^{J} \left(P(c_{ij})\right) f(\boldsymbol{\Omega}) d\boldsymbol{\Omega} \tag{7}$$

Finally, the log-likelihood function is:

$$LL = \sum_i Ln(L_i) \tag{8}$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 8. The parameters to be estimated in the RPMNB model are: $\boldsymbol{\delta}_j$, α_j, $\boldsymbol{\pi}_j$, and $\boldsymbol{\sigma}_j$.

3.3.3 Copula-Based Multivariate NB Model

The focus of our study is to estimate a copula-based multivariate NB modeling framework (see (Bhat and Eluru, 2009; Bhowmik et al., 2018; Yasmin et al., 2014) for a detailed description). The econometric framework for the copula-based model is presented in this section. Let's assume v_{ij} is the expected number of crashes occurring in TAZ i over a given time period for crash type j. We can express v_{ij} as a function of explanatory variable ($\boldsymbol{x}_{ij}$) by using a log-link function as:

$v_{ij} = E(c_{ij}|\boldsymbol{x}_{ij}) = exp(\boldsymbol{\beta}_j \boldsymbol{x}_{ij})$, where $\boldsymbol{\beta}_j$ is a vector of parameters to be estimated specific to crash type j.

The correlation or joint behavior of random variables c_{i1}, c_{i2},…c_{iM} are explored in the current study by using a copula-based approach. A copula is a mathematical device that identifies dependency among random variables with pre-specified marginal distribution (Bhat and Eluru, 2009) provide a detailed description of the copula approach). In constructing the copula dependency, let us assume that $\Lambda_1(c_{i1})$, $\Lambda_2(c_{i2})$ … $\Lambda_J(c_{iJ})$ are the marginal distribution functions of the random variables c_{i1}, c_{i2},…c_{iM}, respectively; and $\Lambda_{12\dots M}(c_{i1}, c_{i2}, \dots c_{iJ})$ is the *M* variate joint distribution with corresponding marginal distributions. Subsequently, the *M* variate distribution $\Lambda_{12\dots M}(c_{i1}, c_{i2}, \dots c_{iJ})$ can be generated as a joint cumulative probability distribution of uniform [0, 1] marginal variables U_1, U_2 … U_J as below:

$$\begin{aligned}\Lambda_{12\dots M}(c_{i1}, c_{i2}, \dots c_{iJ}) &= Pr(U_1 \le c_{i1},\ U_2 \le c_{i2} \dots, U_M \le c_{iJ}) \\ &= Pr\left[\Lambda_1^{-1}(U_1) \le c_{i1},\ \Lambda_2^{-1}(U_2) \le c_{i2} \dots, \Lambda_M^{-1}(U_M) \le c_{iJ}\right] \\ &= Pr\left[U_1 < \Lambda_1(c_{i1}),\ U_2 < \Lambda_2(c_{i2}) \dots,\ U_M < \Lambda_M(c_{iJ})\right]\end{aligned} \quad (9)$$

The joint distribution (of uniform marginal variable) in equation 9 can be generated by a function $C_{\theta i}(.,.)$ such that:

$$\Lambda_{12\dots M}(c_{i1}, c_{i2}, \dots c_{iJ}) = C_{\theta i}(U_1 = \Lambda_1(c_{i1}), U_2 = \Lambda_2(c_{i2}) \dots,\ U_J = \Lambda_M(c_{iJ})) \quad (10)$$

where, $C_{i\theta}(.,.)$ is a copula function and θ_i is the dependence parameter defining the link between $c_{i1}, c_{i2}, \dots c_{iJ}$. In the case of continuous random variables, the joint density can be derived from partial derivatives. However, in our study, c_{ij} are nonnegative integer valued events. For such count data, following (Cameron et al., 2004), the probability mass function ($\varrho_{i\theta}$) is presented (instead of continuous derivatives) by using finite differences of the copula representation as follows:

$$\varrho_{i\theta}\left(\Lambda_1(c_{i1}), \Lambda_2(c_{i2}) \dots \Lambda_M(c_{iJ})\right)$$

$$= \sum_{a_1=1}^{2} \sum_{a_2=1}^{2} \dots \sum_{a_J=1}^{2} (-1)^{a_1+a_2+\dots\ a_J} \left[C_{i\theta}\left(\Lambda_1(c_{i1}+a_1-2), \Lambda_2(c_{i2}+a_2 - 2) \dots \Lambda_M(c_{iJ}+a_J-2)\,;\ \theta_i\right)\right] \quad (11)$$

The reader would note the probability in Equation 11 is written in terms of 2^J copula evaluations (see (Eluru et al., 2010; Sener et al., 2010) for a similar derivation). The number of computations increases rapidly with the number of dependent variables (J), but this is not much of a problem when the dependent variable number J is 6 or less because of the closed-form structures of the copula function evaluation. Given the above setup, we specify $\Lambda_1(c_{i1})$, $\Lambda_2(c_{i2})$ … $\Lambda_M(c_{iM})$ as the cumulative distribution function (cdf) of the NB formulation. The cdf of NB probability expression (as presented in Equation 1) for c_{ij} can be written as:

$$\Lambda_j(c_{ij}|v_{ij}, \alpha_j) = \sum_{k=0}^{c_{ij}} P_{ij}(c_{ij}|v_{ij}, \alpha_j) \quad (12)$$

Thus, the log-likelihood function (LL) with the joint probability expression in Equation 12 can be written as:

$$LL = \sum_{i=1}^{N} \ln\left(\varrho_{i\theta}\left(\Lambda_1(c_{i1}), \Lambda_2(c_{i2}) \dots \Lambda_M(c_{iJ})\right)\right) \quad (13)$$

In the current empirical study, we employ Archimedean copulas that span the spectrum of different kinds of dependency structures including Frank, Gumbel, Clayton and Joe copulas (see (Bhat and Eluru, 2009) for graphical descriptions of the implied dependency structures). Archimedean copulas, in their multivariate forms, allow only positive associations and equal dependencies among pairs of random variables. It is important to note here that, the study allow the dependency structure to vary across TAZs. Therefore, in the current study, the dependence parameter θ_i is parameterized as a function of observed attributes as follows:

$$\theta_i = fn(\boldsymbol{\rho} \quad \boldsymbol{w}_i) \tag{14}$$

where, $\boldsymbol{w}_i$ is a vector of exogenous variables, $\boldsymbol{\rho}$ is a vector of unknown parameters to be estimated (including a constant). Based on the dependency parameter permissible ranges, alternate parameterization forms for the four Archimedean copulas are considered in our analysis. The parameters are estimated using maximum likelihood approaches. The model estimation routine is coded in GAUSS Matrix Programming software.

3.3.4 Copula-Based Random Parameter Multivariate NB Model

Building on the model structure in 3.3.3, we consider the parameters to vary across the population. For this purpose, v_{ij} (expected number of crashes occurring in TAZ i over a given time period for crash type j) equation from 3.3.3 is updated as follows:

$$v_{ij} = E\left(c_{ij}|\boldsymbol{x}_{ij}\right) = exp\left((\boldsymbol{\beta}_j + \Phi_i)\boldsymbol{x}_{ij}\right) \tag{15}$$

where Φ_i is a vector of unobserved factors moderating the influence of attributes in x_{ij} on the crash count propensity for analysis unit i and crash type j.

In examining the model structure of crash count across different crash types, it is necessary to specify the structure for the unobserved vectors Φ represented by Ψ. In this framework, it is assumed that these elements are drawn from independent normal distributions: $\Psi \sim N(0, v_j^2)$. Thus, conditional on Ψ, the likelihood function for the joint probability can be expressed as:

$$L = \int_{\Psi} \ln\left(\varrho_{i\theta}\left(\Lambda_1(c_{i1}), \Lambda_2(c_{i2}) \ldots \Lambda_M\left(c_{iJ}\right)\right)\right) f(\Psi) d\Psi \tag{16}$$

Finally, the log-likelihood function is:

$$LL = \sum_{i=1}^{N} Ln(L_i) \tag{17}$$

3.4 Empirical Analysis

3.4.1 Model Specification and Overall Measure of Fit

The empirical analysis involved a series of model estimations. At first, four separate independent NB models are estimated for four different crash types to establish a benchmark for comparison. Second, a simulation based RPMNB (Random parameter multivariate NB model) is estimated to examine number of crashes across four different collision types jointly. Third, for the closed-form approach, the empirical analysis involves estimation of count models using four different copula structures (Frank, Clayton, Gumbel and Joe) that restricts the variable effect to be same across the entire TAZs. Fourth, all the copula models (all four) are re-estimated with random parameters across each count dependent variable. Finally, a comparison exercise was undertaken to determine the most suitable model.

The results from the various model systems – convergence log-likelihood, number of parameters and Bayesian Information Criterion (BIC) metric are presented in Table 3.2. The reader would note that for the copula models with and without random parameters four alternative model structures were estimated. From the table, several observations can be mde. First, it is evident that all models perform better than the independent model which illustrates the importance of incorporating for the influence of unobserved factors in examining crash count by different crash types. Second, across copula models, Clayton copula model provides the superior fit compared to other copula models in both classes (without random effect and with random effect). Third, within copula system, models considering random parameters outperform their counterparts that do not consider random parameters. Fourth, comparing the copula model system with the RPMNB model, we observe that in general copula based model systems (both classes with and without random effect) provide improved data fit compared to the RPMNB model (except Joe copula without random effect). Fifth, Random Parameter Clayton Copula (RPCC) provides the best model fit

(lowest BIC value) in accommodating the dependency among crash counts for four crash types. The results illustrate the value of accommodating for unobserved heterogeneity through analytical formulations whenever possible.

3.4.2 Model Estimation Results

This section offers a detailed discussion of the effects of exogenous variables on the crash count component for different crash types. To conserve on space, we will restrict ourselves to the discussion of RPCC model results (however, the estimation results of the RPMNB model are presented in Table 3.4). Table 3.3 summarizes the estimation results for the RPCC where the 2nd, 3rd, 4th and 5th column represents the count component for motorized intersection, motorized road segment, motorized off-road and non-motorized crashes, respectively. The copula parameters are presented in the last row panel of Table 3.3. A positive (negative) sign for a variable in the crash count component of Table 3.3 indicates that an increase in the variable is likely to result in more (less) crashes. For the sake of brevity, model results are discussed for all crash types simultaneously by different variable groups.

3.4.2.1 Roadway Characteristics

Proportion of arterial roads is associated with increased incidence of crash in all crash types except motorized off-road category. The result is expected because off-road crashes are likely to be related with high vehicular speed whereas in arterial roads, speeds are likely to be lower due to higher vehicular volume. The coefficient associated with number of intersections reveals a positive impact on motorized intersection and non-motorized crashes while a negative effect is observed for motorized off-road crashes. This is intuitive as intersections are one of the most hazardous location for both motorists and non-motorists due to complex turning movements (see (Abdel-Aty

et al., 2005; Cai et al., 2016) for similar results). Signal intensity offers a negative sign on off-road crashes indicating a lower likelihood of motorized off-road crash in a TAZ with increased number of signals. As expected, vehicles are likely to drive at a lower speed in the location with higher number of signals and as a result, the risk of motorized off-road crashes might go down. Further, the estimated results show that a TAZ with higher variance in speed limit is likely to experience increased number of motorized intersection, road segment and off-road crashes. On the other hand, the likelihood of these three crash types are lower for zones with higher width of outside shoulder which is perhaps indicating greater safety margins for vehicular maneuvers. With respect to sidewalk width, the variable is found to be significant in non-motorized crash component with a negative impact indicating a lower risk for non-motorists with increased sidewalk width.

3.4.2.2 Land-use Attributes

With regards to land-use attributes, several factors are found to be significant determinants of crash counts for different crash type components. The model estimation results reveal that there are higher likelihoods of motorized intersection, motorized road segment and non-motorized crashes in a TAZ with higher urbanized and office areas. Institutional area is positively associated with motorized intersection and non-motorized crashes. As evident from Table 3.3, we can see that the variable indicating residential area is found to have a negative impact on motorized intersection crashes while a positive association is observed for non-motorized crashes.

3.4.2.3 Built Environment Characteristics

The variable corresponding to built environment characteristics reveals that higher number of restaurants and shopping centers are likely to result in increased number of intersection and road

segment crashes for motorists. With respect to non-motorized crashes, number of restaurants is found to be a significant determinant with a positive impact (see (Yasmin et al., 2018a) for similar result). However, none of the built environment attributes are found to have significant impacts on motorized road segment crashes.

3.4.2.4 Traffic Characteristics

The parameters associated with traffic characteristics offer expected results. With higher VMT, a TAZ is likely to have higher crash incidence for all crash types. Further, we found a significant variability of VMT specific to motorized on-road crashes as indicated by the standard deviation parameter. The distributional parameter indicates that the overall impact of VMT on motorized on-road crashes is always positive (99.99%). Additionally, proportion of heavy vehicles is found to be positively associated with motorized road segment crashes.

3.4.2.5 Sociodemographic Characteristics

With respect to sociodemographic characteristics, the estimates indicate that TAZs with high share of walk and bike commuters are likely to experience more motorized intersection crashes. On the other hand, the parameter for proportion of household with no vehicle reveals a positive association with non-motorized crashes. This is expected because people from households without access to vehicles are more exposed to the traffic as they are restricted to using public transport, walk or bike as their primary mode for their trips. In terms of sociodemographic characteristics, no other variables are found to have significant impacts on motorized road segment and off-road crashes.

3.4.2.6 Spatial Spillover Effect

In terms of spatial spillover effects, office area of the surrounding zones is found to be positively associated with motorized intersection and road segment crashes of the targeted zones. As expected, signal intensity in the neighbouring zones has a positive impact on motorized intersection crash. TAZs surrounded by zones with higher proportion of major road are likely to experience more motorized road segment crashes. Number of commuters by walking and bicycling and proportion of household with zero vehicle in the neighbouring zones have a positive influence on non-motorized crashes. Moreover, we accommodate the variation of the influence of this variable (indicated by the standard deviation in table 3) on non-motorized crashes and found that the overall impact is not always to be positive (61.79% positive). On the other hand, average sidewalk width in the surrounding zones has a negative coefficient indicating a reduction in non-motorized crashes of the targeted zone. However, in terms of motorized off-road crashes, none of the spatial spillover variables are found to have a significant impact.

3.4.2.7 Dependency Effect

The copula parameter representing the dependency effects across different count components by crash types is presented in the last row panel of Table 3.3. As highlighted earlier, in the current analysis, Clayton copula (with random effect) has provided the best model fit in accommodating the dependency among crash counts for four crash types. For the Clayton copula, the dependency is entirely positive, and the coefficient sign and magnitude reflect whether a variable increase or reduces the dependency across dimensions and by how much. The Clayton copula is best suited for strong left tail dependence and weak right tail dependence (see (Eluru et al., 2010) for detail); that is, it is suitable for the case when, after controlling for observed covariates, all four crash types

tend to have a simultaneously high propensity for low crash counts, but not a simultaneously high propensity for high crash counts. Further, as indicated earlier, the dependency is expressed as a function of observed attributes. Several variables are explored and number of intersections is found to have a significant impact on the correlation profile supporting our hypothesis that the dependency profile varies across TAZs. The proposed framework by incorporating for such parameterizations allows us to improve the model estimation results.

3.5 Predictive Performance Evaluation

In order to demonstrate the comparison between RPMNB and random parameter Copula-based frameworks, we evaluate the predictive performance by employing goodness of fit measures including MPB (Mean prediction bias), MAD (mean absolute deviation), MAPE (mean absolute percentage error), RMSE (Root mean square error) and predictive log-likelihood (please see (Bhowmik et al., 2018) for a discussion on estimating these measures). Two types of prediction exercise are undertaken: 1) In-sample prediction for the zones used in model estimation (3,800) and 2) holdout sample prediction for the zones that have been set aside for validation analysis (947). The reader would note that these fit measures quantify the error associated with model predictions and the model with lower value of predictive measures and higher value of predictive log-likelihood will provide better prediction of the observed data. Table 3.5 summarizes the value of these measures for both RPMNB and RPCC models at a disaggregate level. As evident form Table 3.5, we can observe that RPCC outperforms the RPMNB model across most of the (38 out of 42) measures computed. The result clearly highlights the superiority of the proposed approach over the traditional RPMNB framework.

In an effort to further assess the predictive performance of the estimated models, an in-depth comparison for different count events across different crash types are carried out.

Specifically, we predict the crash frequencies across different count alternatives for different crash types estimated from the two models RPMNB and RPCC and compare their performance based on that. For this purpose, 20 data samples with 250 records (TAZs) each are randomly generated from the holdout validation sample consisting of 947 records (TAZs). For these samples, we predict the number of TAZs for different count events (total 5 count categories are considered for each crash types based on the crash count distribution. For example: for intersection crashes, five classes are considered - TAZs with 0, 1-5, 6-20, 21-40 and >40crashes) across different crash types from both models (RPMNB and RPCC) and using these counts, we generate the ratio of predicted to observed counts specific to each level (count events and crash types). For instance, if there are 100 TAZs (out of 250) from data sample 1 experiencing ”0” single non-motorized crash and we predict 70 and 80 TAZs from RPMNB and RPCC model, then the estimated ratio of these models will be 0.7 (70/100) and 0.8 (80/100) respectively. The reader would note that, the estimated ratio corresponds to the value of 1 would imply a perfect prediction. For the ease of presentation, we generate two box plots using all the data samples (total 20 points for every count alternative) specific to each model (RPMNB and RPCC) by each count events across the four crash types. Figure 3.2 represent the ratio statistics for different crash types while in figure 3.3, we present the overall ratio statistics incorporating all the crashes together (total 80 points for each count alternatives). In terms of the crash types, it is very clear (from figure 3.2) that the RPCC offers better prediction relative to the RPMNB especially for the motorized crashes in the current study context. However, for the non-motorized crashes, the RPMNB model performs marginally better. On the other hand, based on the overall crash perspective, the resulting predictive measures estimated for different count alternative further confirm the superiority of the copula approach over the RPMNB model.

The comparison exercise between these two frameworks was further augmented by undertaking a correct classification analysis. Based on observed crash counts for each crash type, we divided all the zones (4,747) into 4 groups based on the quartile for number of crashes. Again, based on the predicted counts from both RPMNB and RPCC model, we create 4 groups of zones similarly and compute the percentage of correctly classified TAZs within each group. Figure 3.1 represents the classification accuracy for both RPMNB and RPCC model by each quartile across different crash type. From Figure 3.1, the reader would note that for motorized intersection crashes, the classification percentage for the RPCC model is 17.4% in the 1st quartile which denotes that out of 1,187 TAZs, around 772 are correctly classified for the 1st quartile. This means, within the first quartile, the RPCC framework is able to classify around 70% (17.4*4) TAZs correctly for intersection crashes. Similarly, we can observe that for almost every crash type, the accuracy rate is higher for the RPCC model (except non-motorized crashes: RPMNB model has slightly better prediction rate in the higher quartiles) relative to RPMNB within each quartile which further reinforces the superiority of the copula model in the current study context.

3.6 Spatial Distribution

To illustrate the applicability of the estimated copula model, we also identify the hot and cold zones by using prediction of the estimated RPCC model. Specifically, we generate the predicted number of crashes by crash type and identify the cold (bottom ten percentile zones with respect to number of crashes) and hot zones (top ten percentile zones with respect to number of crashes). The predicted results for Central Florida for the year 2016 are presented in Figure 3.4. Figure 3.4a to 3.4d represents the hot and cold zone locations for all crash types considered while Figure 3.4e represents the hot and cold zone locations for all crashes (identified based on common hot/cold zones across all crash types). From figure 3.4a to 3.4d, we can observe that Orange and Seminole

county are under more risk for intersection, non-motorized and on-road crashes while the risk of getting involved in off-road crashes is higher in Polk, Osceola and Lake county. On the other hand, Volusia and Brevard county are found to be relatively safe across crash types. For hot and cold zones by all crashes, the results indicate that TAZs with greater risk are dispersed throughout the Central Florida region with visible clustering. This spatial illustration can easily be used to prioritize TAZs based on crash risk across different crash types to enhance road safety.

3.7 Summary

In our research, we compare the performance of the simulation-based framework with closed-form copula-based frameworks. In addition, we build on the closed-form copula based frameworks to incorporate unobserved heterogeneity associated with variable impacts on crash types (random parameters). The proposed model system is compared with the simulation based and analytical multivariate models. The comparison exercise is undertaken with the univariate models following negative binomial model structure. Within the copula framework, we estimate models for four copula structures: (1) Frank, (2) Gumbel, (3) Clayton and (4) Joe which cover a wide range of dependency structures, including radial symmetry and asymmetry, and asymptotic tail independence and dependence. The empirical analysis is based on the traffic analysis zone (TAZ) level crash count data for both motorized and non-motorized crashes from Central Florida for the year 2016. The models were estimated employing a comprehensive set exogenous variable including roadway, built environment, land-use, traffic, socio-demographic characteristics and spatial spillover effects. The model fit measures clearly highlight that the RPCC (random parameter Clayton copula) model outperforms simulation-based RPMNB model. The comparison exercise was further augmented by generating a host of comparison metrics for both estimation sample and hold-out sample. In an effort to further assess the predictive performance of the

estimated models, an in-depth comparison for different count events across different crash types and correct classification analysis are carried out. The estimated results further reinforce the superiority of the RPCC-based multivariate approach. The RPCC based copula model is also employed to generate hot and cold zone categorization of TAZs in the Central Florida region to identify potential vulnerable zones by crash type.

The proposed model results offer insights on important variables affecting crash frequency by crash types (road user and location for the current study context). The macro-level model outcomes can be used to devise safety-conscious decision support tools to facilitate a proactive approach in assessing medium and long-term policy-based countermeasures. Moreover, with the spatial illustration, high risk zones for every crash type can be easily identified and thus help the planners in enhancing safety for these high crash risk zones.

The objective is not without limitations. While the study considers the effect of observed spatial attributes, it would be beneficial to capture the spatial unobserved heterogeneity as well. Moreover, it might be interesting to explore the transferability of models developed for crash type simultaneously by estimating similar models for multiple spatial units across several years.

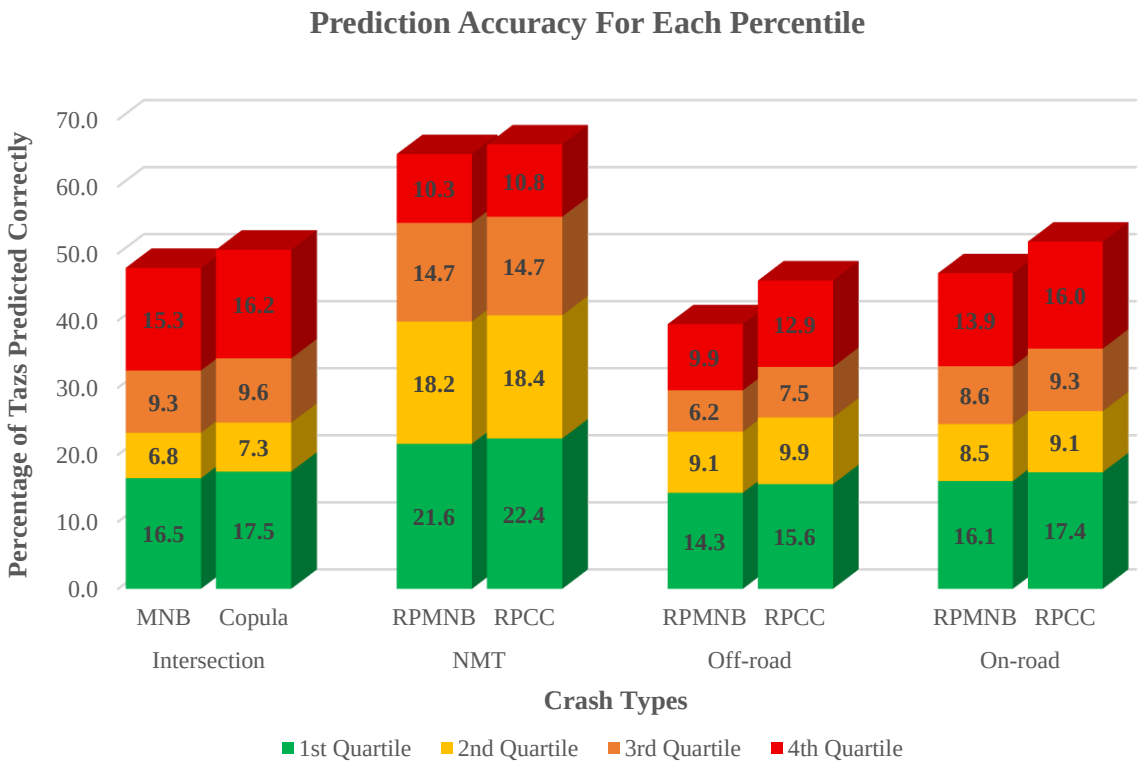

Figure 3.1 Prediction Accuracy for Two Frameworks by Crash type Quartile

Note: *RPMNB=Random parameter multivariate negative binomial model, RPCC =Random Parameter Clayton Copula model

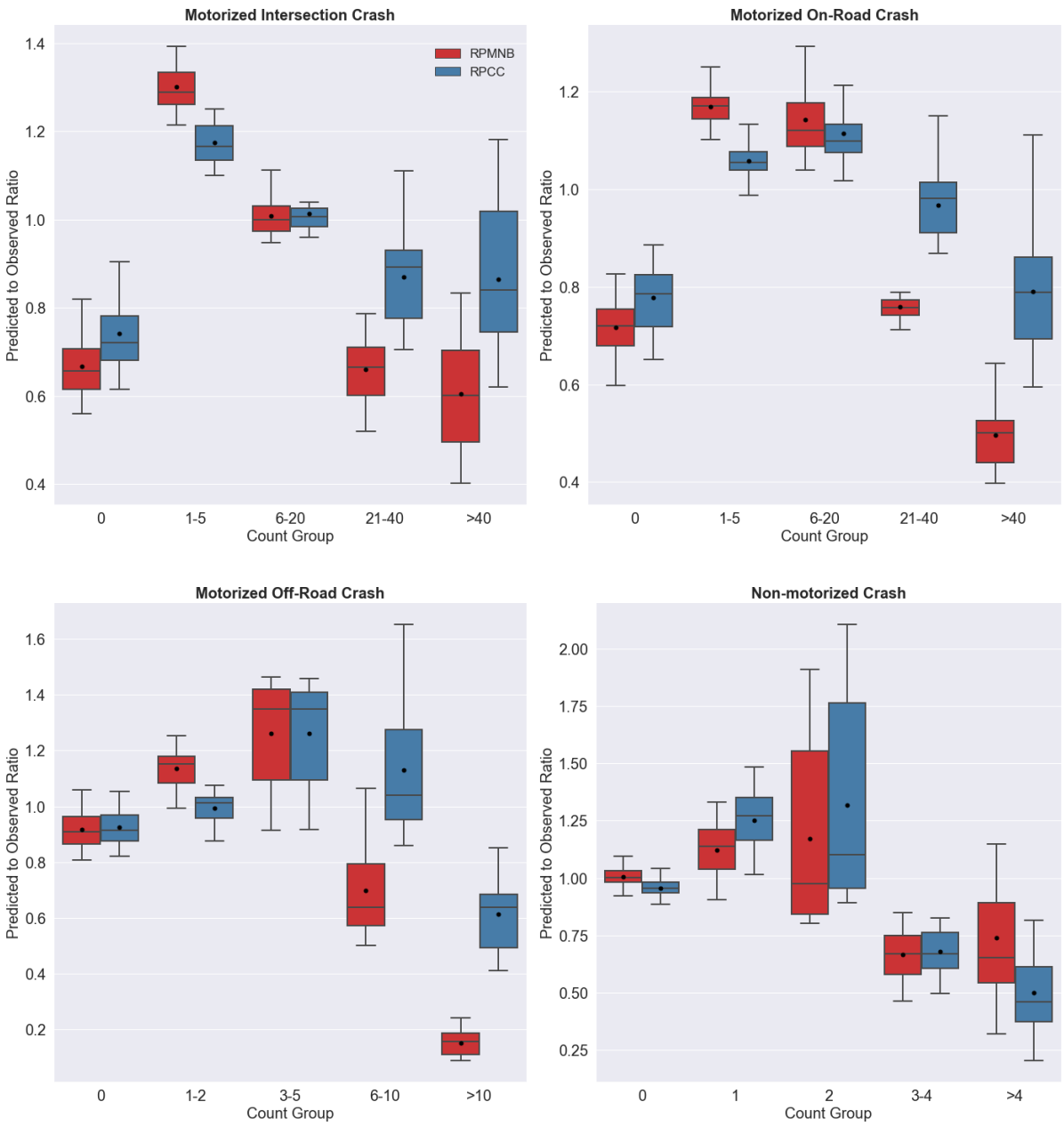

Figure 3.2 Predicted to Observed Ratio for Different Crash Types

Note: *RPMNB=Random parameter multivariate negative binomial model, RPCC =Random Parameter Clayton Copula model

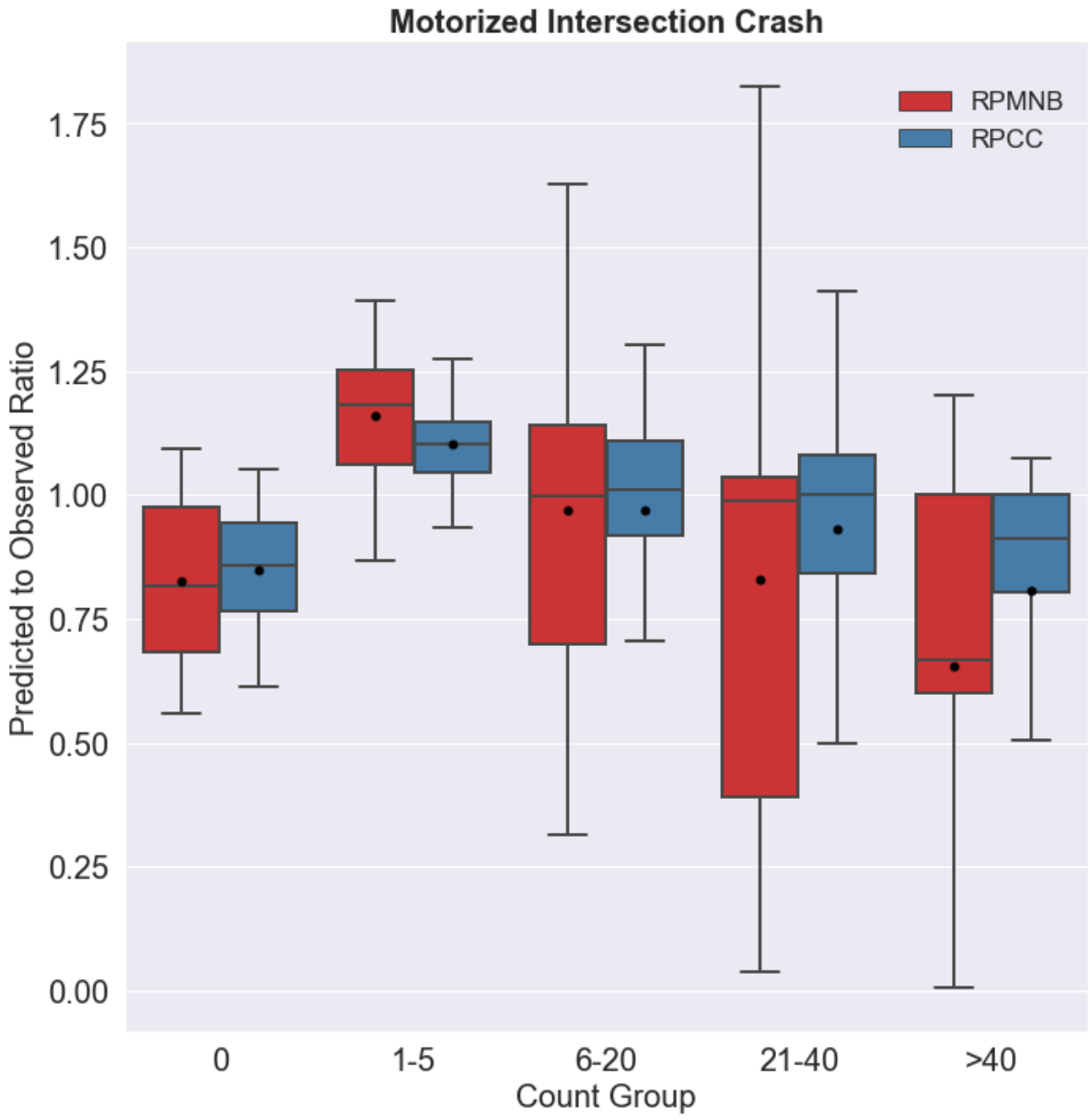

Figure 3.3 Predicted to Observed Ratio for Overall Crashes

Note: *RPMNB=Random parameter multivariate negative binomial model, RPCC =Random Parameter Clayton Copula model

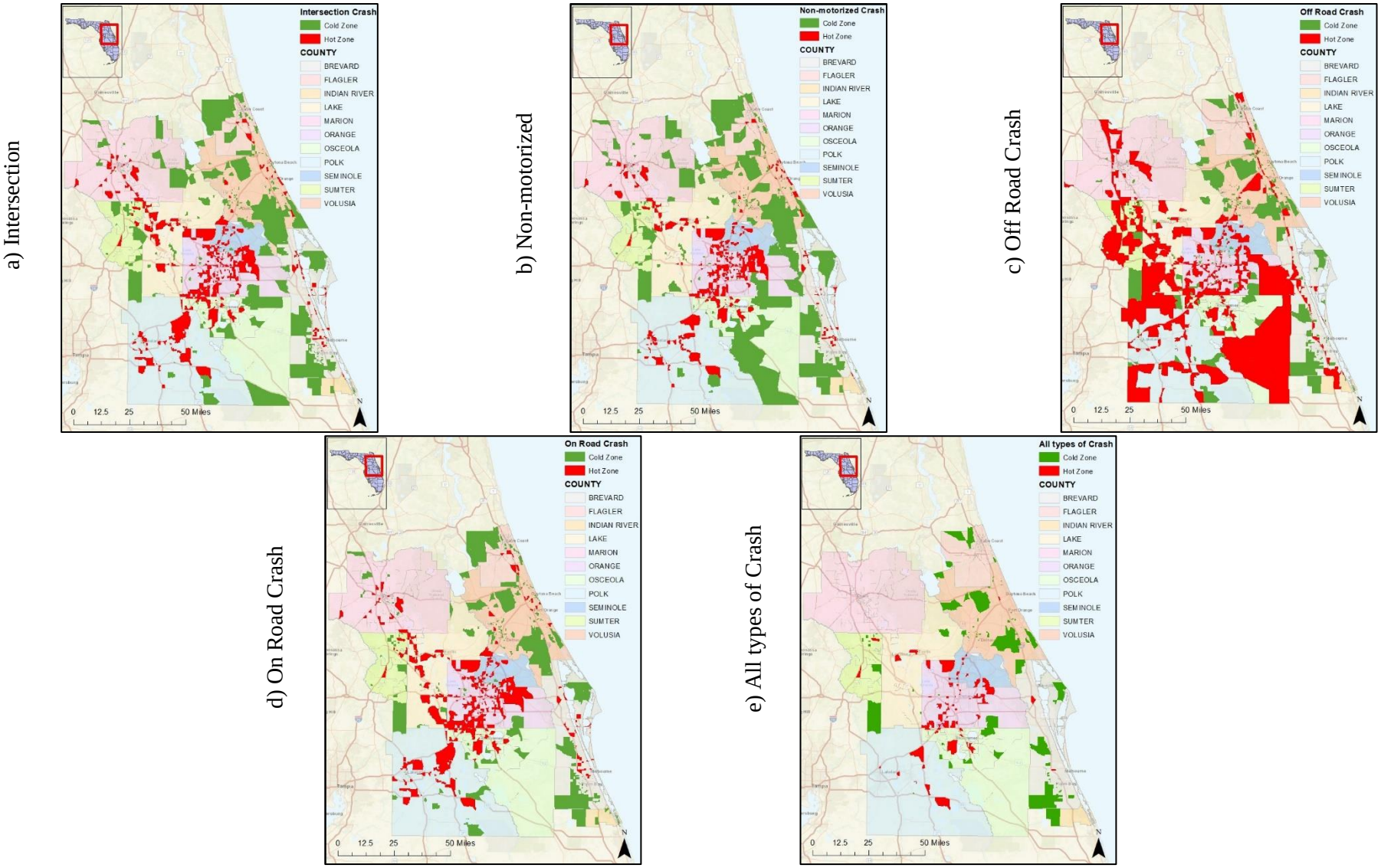

Figure 3.4 Spatial Distribution for Every Crash Types

Table 3.1 Summary of Existing Crash Frequency Studies

Studies	Study Unit	Methodology	Estimation Technique	Dependent Variables Analyzed	Number of Dimension
Simulation-Based Approach					
Count Framework					
(Anastasopoulos et al., 2012)	Micro	Multivariate tobit regression	MSL*	Rates of crashes by severity levels - no-injury, possible injury and injury crashes	3
(Aguero-Valverde, 2013)	Macro	Multivariate Spatial Model	MSL	by severity level - property damage only, possible injury, and injury/fatality	3
(Bhat et al., 2014)	Micro	Random parameters count models	MACML	by intersection control type – No control. Yield sign, stop sign, flashing light, regular signal light	5
(Chiou and Fu, 2013)	Micro	Multinomial-Generalized Poisson (MGP) Withatu1/without Error-Components (EMGP) and Nested Generalized Poisson Models (NGP)	MSL	by severity level - property damage only, possible injury, and injury/fatality by segment length	3
(Li et al., 2013)	Macro	Geographically Weighted Poisson Regression (GWPR)	MSL	Fatal crash only	1
(Wang and Kockelman, 2013)	Macro	Poisson-based multivariate conditional auto-regressive (CAR) framework	MCMC	Pedestrian Crash Counts by walk miles travelled (WMT)	1
(Ye et al., 2013)	Micro	Joint Poisson regression model	MSL	by severity level - property damage only, possible injury, and injury/fatality	3
(Yu and Abdel-Aty, 2013)	Micro	Bayesian bivariate Poisson-lognormal model and a Bayesian hierarchical Poisson model	MCMC	by multi-vehicle and single vehicle crash	2
(Zou et al., 2014)	Micro	Finite-mixture/latent-class and Markov switching models	MSL	by segment length	11
(Barua et al., 2014)	Micro	Multivariate Poisson lognormal model	MCMC	by crash severity – no injury and injury/fatal crashes	2
(Dong et al., 2014)	Micro	Multivariate random-parameters zero-inflated negative binomial model	MCMC	by vehicles involved – car only crash, car-truck crash and truck only crash	3
(Chiou et al., 2014)	Micro	Multinomial-Generalized Poisson With Error-Components (EMGP) - spatial error-EMGP and spatial exogenous-EMGP	MSL	by severity level - property damage only, possible injury, and injury/fatality by segment length	3

(Chiou and Fu, 2015)	Micro	Multinomial generalized Poisson model with error components and spatiotemporal dependence (ST-EMGP)	MSL	by severity level - property damage only, possible injury, and injury/fatality	3
(Lee et al., 2015)	Macro	Multivariate Poisson Lognormal Conditional Autoregressive Model	MSL	by modes - motor vehicle, bicycle, and pedestrian	3
(Zhan et al., 2015)	Macro	Multivariate Poisson-lognormal model	MCMC	by severity levels – fatal and severe injury crashes Crash frequency by crash severity – no injury, possible injury and evident injury	2, 3
(Aguero-Valverde et al. 2016)	Micro	Multivariate Poisson log-normal spatial model	MCMC	by crash types – same direction, opposite direction, angle and hit-fixed object crashes	4
(Anastasopoulos, 2016)	Micro	Random parameter multivariate tobit model, Multivariate zero-inflated negative binomial model	MSL	by severity type – PDO, injury and fatality	3
(Barua et al., 2016)	Micro	Bayesian multivariate random parameters spatial model	MCMC	by severity levels – no injury and injury/fatal crashes	2
(Dong et al., 2016)	Micro	Random parameter bivariate zero-inflated negative binomial model	MCMC	by severity – disabling injury and non-disabling injury	2
(Mothafer et al., 2016)	Micro	Multivariate Poisson Gamma Mixture Count Model (MVPGM)	MSL	by crash types – rear end, sideswipe, fixed object and other crash types on freeway section	4
(Serhiyenko et al., 2016)	Micro	Multivariate Poisson Lognormal model	MCMC	by crash type – single vehicle, same direction and opposite direction crashes	3
(Zeng et al., 2016)	Macro	Neural Networks Model	MCMC	by severity level on road segments - fatality or serious injury and slight injury	2
(Chen et al., 2017)	Micro	Multivariate Random Parameters Negative Binomial Approach	MSL	by severity level - property damage only, possible injury, and injury/fatality by pavement conditions – Excellent, Good, Good-Fair, Fair and Poor.	3, 5
(Cheng et al., 2017)	Micro	Multivariate Poisson lognormal temporal and spatial models	MCMC	by crash type - Rear-end, Head-on, Side-swipe, Broad-side, Hit object, and Other crashes	6
(Heydari et al., 2017)	Micro	Bayesian latent class flexible mixture multivariate model	MCMC	by crash type – pedestrian and bicycle crashes	2
(Huang et al., 2017)	Micro	Multivariate Poisson log-normal regression model	MCMC	by transportation Modes (motor vehicle, bicycle and pedestrian crashes) at urban intersections.	3
(Wang et al., 2017)	Micro	Integrated Nested Laplace Approximation Multivariate Poisson Lognormal model	MCMC	by crash types –same-direction, intersection-direction, opposite direction and single vehicle crashes and by severity outcomes – no injury,	4, 3

				possible/non-incapacitating injury and fatal/incapacitating injury crashes	
(Zeng et al., 2017)	Micro	Multivariate random parameter tobit model	MCMC	by severity levels – slight injury crash and killed/seriously injured crashes	2
(Cheng et al., 2018)	Macro	Multivariate Space-Time Models with Different Temporal Trends and Spatiotemporal Interactions	MCMC	by collisions modes - motor vehicle, pedestrian, bicycle, and motorcycle	4
Fractional Split Framework (proportion of crashes)					
(Bhowmik et al., 2018)	Macro	Joint Negative Binomial-Multinomial Logit Fractional Split (NB-MNLFS) Model	QMCSL	by collision type - rear-end, head-on, angular, left-turn, right-turn, off-road, rollover, sideswipe, other collision type	10
(Lee et al., 2018)	Macro	Mixed Fractional Split Multinomial Logit Modeling Approach	QMCSL	by vehicle type	8
(Bhowmik et al., 2018)	Macro	Joint Negative Binomial-Ordered Logit Fractional Split (NB-OLFS) Model	QMCSL	by crash severity - (1) proportion of no injury crashes, (2) proportion of minor injury crashes, (3) proportion of incapacitating injury crashes and (4) proportion of fatal crashes	4
Closed-Form Approach (count)					
(Narayanamoorthy et al., 2013)	Macro	Spatial Multivariate Count Model	CML	by severity level – Possible injury, non-incapacitating injury, incapacitating injury and fatal injury	4
(Nashad et al., 2016)	Macro	Copula based bivariate negative binomial model	ML	by crash type – pedestrian and bicycle crashes	2
(Yasmin et al., 2018b)	Macro	Copula based multivariate approach	ML	By road user group – car, light truck, other motorized (truck, bus and other vehicles) and non-motorized (pedestrian and bicyclist)	4

Note: *MSL= Maximum simulated likelihood approach, MCMC= Markov Chain Monte Carlo approach, MACML=maximum approximate composite marginal likelihood, QMCSL= Qausi monte carlo simulated likelihood approach, ML= Maximum likelihood approach, CMT=Composite marginal likelihood approach.

Table 3.2 Summary of Statistical Data Fit from Different Model Systems

Model (Sample Size = 3,800)		Log-Likelihood	No. of Parameter	AIC	BIC
Independent Model		-33108.777	51.000	66319.554	66637.934
RPMNB		-32541.376	54.000	65190.752	65527.861
Copula Without random effect	*Frank*	-32330.666	53.000	64767.332	65098.198
	Clayton	-32285.560	53.000	64677.120	65007.986
	Gumbel	-32477.992	52.000	65059.984	65384.607
	Joe	-32609.282	52.000	65322.564	65647.187
Copula With random effect	*Frank*	-32324.966	54.000	64757.932	65095.041
	Clayton	-32269.296	55.000	64648.592	64991.944
	Gumbel	-32437.408	53.000	64980.816	65311.682
	Joe	-32345.828	54.000	64799.656	65136.765

Table 3.3 Random Parameter Clayton Copula (RPCC) Model Estimation Results

Variables (N=3800)	Motorized Intersection Crashes		Motorized On-Road Crashes		Motorized Off-Road Crashes		Non-motorized Crashes	
	Estimate	T-stat	Estimate	T-stat	Estimate	T-stat	Estimate	T-stat
Constant	-0.403	-8.614	-0.986	-18.434	-0.795	-17.644	-2.823	-35.121
Roadway Characteristics								
Proportion of arterial roads	0.134	5.268	0.118	4.306	-0.309	-7.366	0.224	5.608
Number of intersections	0.302	13.873	--	--	-0.070	-6.363	0.249	10.132
Signal Intensity	--	--	--	--	-0.842	-5.635	--	--
Variance of speed limit	0.030	4.338	0.065	7.219	0.056	7.554	--	--
Average width of outside shoulder	-0.256	-9.248	-0.330	-10.574	-0.122	-5.684	--	--
Average sidewalk width							-0.140	-4.854
Land-use Attributes								
Urban rea	0.142	16.194	0.107	13.656	--	--	0.140	11.697
Office area	0.158	13.206	0.107	10.725	--	--	0.101	8.925
Institutional area	0.052	5.808	--	--	--	--	0.066	5.325
Residential area	-0.076	-12.069	--	--	--	--	0.025	5.933
Built Environment Characteristics								
Number of restaurants	0.230	13.599	0.255	12.551	--	--	0.245	13.638
Number of shopping centers	--	--	0.049	6.623	--	--	--	--
Traffic Characteristics								
VMT	0.057	8.281	0.161	23.760	0.198	29.882	0.031	5.887
Standard Deviation			0.018	4.304				
Proportion of heavy vehicles	--	--	--	--	2.023	6.955	--	--
Socio-demographic Characteristics								
Non-motorist commuter	0.036	4.701	--	--	--	--	--	--
Proportion of HH with no vehicles	--	--	--	--	--	--	2.060	6.811
Spatial Effects								
Office area	0.100	8.771	0.176	14.987	--	--	--	--

Signal intensity	1.868	6.761	--		--	--	--	--
proportion of major road	--	--	0.450	8.625	--	--	--	--
Proportion of HH with no vehicle	--	--	--	--	--	--	2.253	7.579
Non-motorist commuter	--	--	--	--	--	--	0.034	10.025
Standard Deviation							0.148	17.317
Average sidewalk width	--	--	--	--	--	--	-0.133	-4.820
Over-dispersion	0.755	34.710	0.841	31.889	0.724	25.168	0.059	5.671
Copula Parameter	Estimate				T-stat			
Constant	0.824				31.432			
Number of intersections	-0.015				-6.632			
Log-Likelihood *(No. of parameters):* -32,269.30 (55); **AIC**: 64,648.59; **BIC:** 64,991.94								

Table 3.4 Random Parameter Multivariate NB (RPMNB) Model Estimation Results

Variables (N=3800)	Motorized Intersection Crashes		Motorized On-Road Crashes		Motorized Off-Road Crashes		Non-motorized Crashes	
	Estimate	T-stat	Estimate	T-stat	Estimate	T-stat	Estimate	T-stat
Constant	-1.086	-12.170	-1.541	-14.687	-1.488	-25.072	-3.477	-20.121
Roadway Characteristics								
Proportion of arterial roads	0.151	2.666	0.113	1.843	-0.307	-5.102	0.216	2.872
Number of intersections	0.319	11.358	--	--	--	--	0.335	7.628
Signal Intensity	--	--	--	--	-0.996	-3.951	--	--
Standard Deviation	--	--	--	--	0.848	2.034	--	--
Variance of speed limit	0.033	2.762	0.061	4.667	0.052	4.041	--	--
Average width of outside shoulder	-0.262	-6.013	-0.395	-8.520	-0.159	-3.749	--	--
Average sidewalk width	--	--	--	--	--	--	-0.198	-3.071
Land-use Attributes								
Urban rea	0.151	12.720	0.105	9.080	--	--	0.153	7.542
Office area	0.173	10.445	0.088	5.108	--	--	0.146	7.048
Institutional area	0.075	5.248	--	--	--	--	0.089	4.544
Residential area	-0.072	-7.290	--	--	--	--	0.027	1.680
Built Environment Characteristics								
Number of restaurants	0.260	11.101	0.257	7.986	--	--	0.268	11.636
Standard Deviation	--	--	0.096	2.211	--	--	--	--
Number of shopping centers	--	--	0.063	2.933	--	--	--	--
Traffic Characteristics								
VMT	0.071	6.478	0.213	21.065	0.232	24.954	0.038	2.395
Proportion of heavy vehicles	--	--	--	--	2.545	5.604	--	--
Socio-demographic Characteristics								
Non-motorist commuter	0.074	4.644	--	--	--	--	--	--
Proportion of HH with no vehicles	--	--	--	--	--	--	1.730	3.245
Spatial Effects								

Office area	0.120	7.459	0.164	9.678	--	--	--	--
Signal intensity	1.696	5.039	--	--	--	--	--	--
proportion of major road	--	--	0.479	6.393	--	--	--	--
Proportion of HH with no vehicle	--	--	--	--	--	--	1.016	1.409
Non-motorist commuter	--	--	--	--	--	--	0.142	6.184
Average sidewalk width	--	--	--	--	--	--	-0.243	-2.660
Over-dispersion	0.304	11.647	0.427	16.618	0.254	8.706	0.035	1.990
Correlation								
Correlation 1	0.686	30.723	--	--	--	--	0.686	30.723
Correlation 2	--	--	0.735	39.370	0.735	39.370	--	--
Log-Likelihood *(No. of parameters):* -32,541.38 (54); **AIC**: 65,190.75; **BIC**: 65,527.86								

Table 3.5 Prediction Performance Evaluation for Two Frameworks

Data	Crash Type	MPB		MAD		MAPE		RMSE		Predicted BIC	
		RPMNB*	RPCC	RPMNB	RPCC	RPMNB	RPCC	RPMNB	RPCC	RPMNB	RPCC
In-Sample Data	Motorized Intersection	1.648	0.756	8.558	6.272	1.380	1.185	21.817	12.867	65,527.86	64,991.94
	Motorized On-road	2.445	1.104	12.049	9.022	1.334	1.558	55.214	24.249		
	Motorized Off-road	0.032	-0.079	2.257	1.859	0.216	0.050	3.708	2.977		
	Non-Motorized	0.046	-0.004	0.804	0.756	0.178	0.219	1.632	1.266		
	Across observation	4.170	1.777	23.668	17.910	3.108	3.012	59.506	27.642		
Validation Data	Motorized Intersection	2.026	0.587	10.042	6.977	2.655	1.250	35.937	16.989	20,904.03	16,864.40
	Motorized On-road	1.155	0.839	12.219	9.077	1.882	1.299	36.179	24.494		
	Motorized Off-road	-0.073	-0.139	2.286	1.930	0.322	0.026	3.945	3.332		
	Non-Motorized	0.071	0.033	0.852	0.818	0.056	0.230	1.987	1.560		
	Across observation	3.179	1.320	25.400	18.801	4.915	2.805	51.185	30.035		

Note: *RPMNB=Random parameter multivariate negative binomial model, RPCC =Random Parameter Clayton copula model

*Model with underline gives better measure

Table 3.6 Independent NB Model Results

Variables (N=3800)	Motorized Intersection Crashes		Motorized On-Road Crashes		Motorized Off-Road Crashes		Non-motorized Crashes	
	Estimate	T-stat	Estimate	T-stat	Estimate	T-stat	Estimate	T-stat
Constant	-0.699	-7.782	-1.468	-13.634	-1.116	-9.995	-3.237	-21.637
Roadway Characteristics								
Proportion of arterial roads	0.153	3.069	0.106	1.904	-0.356	-6.286	0.209	2.925
Number of intersections	0.288	9.613	--	--	-0.064	-2.116	0.270	5.934
Signal Intensity	--	--	--	--	-0.660	-2.601	--	--
Variance of speed limit	0.030	2.731	0.060	5.180	0.052	4.360	--	--
Average width of outside shoulder	-0.231	-6.080	-0.352	-8.390	-0.144	-3.158	--	--
Average sidewalk width	--	--	--	--	--	--	-0.146	-2.473
Land-use Attributes								
Urban rea	0.147	14.254	0.123	11.101	--	--	0.151	8.355
Office area	0.164	10.688	0.118	6.886	--	--	0.121	5.976
Institutional area	0.068	4.666	--	--	--	--	0.083	4.182
Residential area	-0.074	-7.067	--	--	--	--	0.037	2.143
Built Environment Characteristics								
Number of restaurants	0.265	11.844	0.268	9.446	--	--	0.249	10.799
Number of shopping centers	--	--	0.057	1.903	--	--	--	--
Traffic Characteristics								
VMT	0.064	5.677	0.179	17.401	0.237	14.564	0.039	2.446
Proportion of heavy vehicles	--	--	--	--	1.772	3.679	--	--
Socio-demographic Characteristics								
Non-motorist commuter	0.075	4.730	--	--	--	--	--	--
Proportion of HH with no vehicles	--	--	--	--	--	--	1.860	3.287
Spatial Effects								
Office area	0.113	5.933	0.206	10.926	--	--	--	--
Signal intensity	2.017	4.291	--	--	--	--	--	--

proportion of major road	--	--	0.594	6.968	--	--	--	--
Proportion of HH with no vehicle	--	--	--	--	--	--	2.105	2.910
Non-motorist commuter	--	--	--	--	--	--	0.164	7.268
Average sidewalk width	--	--	--	--	--	--	-0.287	-3.221
Over-dispersion	0.757	31.611	0.921	33.970	0.766	18.113	0.452	9.788
Log-Likelihood *(No. of parameters)*	-10909.91 (15)		-11367.74 (12)		-7103.68 (9)		-3727.44 (15)	
Log-Likelihood *(No. of parameters):* -33,108.79 (51); **AIC**: 66,319.58; **BIC**: 66,637.96								

CHAPTER 4: PANEL MIXED APPROACH TO MODELING CRASH FREQUENCY BY CRASH TYPES

[2]In the United States, road traffic crashes have resulted in nearly 40,000 fatalities in 2016 (NHTSA, 2017). In addition to the alarmingly high number of fatalities, there are multiple worrying trends within these numbers. The increase in the number of fatalities year over year for 2015 and 2016 represent the two largest year over year increases over the last three decades. Further, in 2016, the percentage of non-motorized road user fatalities as a proportion of total fatalities have increased. These trends clearly highlight the challenges associated with addressing the enormous consequences of road traffic crashes. Thus, it is not surprising that safety researchers are working toward devising appropriate remedial solutions for reducing the number and consequence of traffic crashes. A major tool employed in the literature to develop counter measures is the application of econometric models for crash frequency and crash severity. Crash frequency models explore the relationship between various attributes and crash occurrences (Yan et al., 2009; Geedipally et al., 2010; Jonathan et al., 2016) while crash severity models, conditional on crash occurrence, examine attributes affecting crash consequences (Abdelwahab and Abdel-Aty, 2002; Milton et al., 2008; Wang and Abdel-Aty, 2008; Eluru et al., 2010). The current research effort contributes to literature on crash frequency analysis by suggesting an alternative and mathematically simpler approach for analyzing multiple crash frequency variables for the same study unit.

[2] Bhowmik, T., Yasmin, S., & Eluru, N. (2019). Do we need multivariate modeling approaches to model crash frequency by crash types? A panel mixed approach to modeling crash frequency by crash types. Analytic Methods in Accident Research, 24, 100107.

4.1 Earlier Research

Several research efforts have developed crash frequency models in safety literature. The various crash frequency dimensions explored in existing literature include total crashes, crashes by severity, crashes by crash type and crashes by vehicle type for a spatial unit over a given time period (Ye et al., 2009, 2013; Lee et al., 2015; Wang et al., 2017; Yasmin et al., 2018). Earlier research efforts typically adopted a univariate framework to study a single crash frequency variable (such as total crashes) or multiple crash frequency variables (such as crash frequency by injury severity). While univariate approaches are adequate to accommodate for the influence of observed factors, they are not appropriate to account for the common unobserved factors affecting the multiple dependent variables for the same observational unit (see (Mannering et al., 2016) for a detailed review). Toward addressing this limitation, several research efforts have developed frameworks that accommodate for the influence of these common unobserved factors (Anastasopoulos, 2016; Mannering et al., 2016; Nashad et al., 2016). These approaches typically estimate the univariate models for crash frequency and bundle these univariate models into a multivariate version. The univariate models could take the form of a negative binomial or a log-normal formulation (or other variants). The bundling process can be achieved through simulation-based approaches within the classical regime using maximum simulated likelihood approaches or in the Bayesian regime using Markov Chain Monte Carlo (MCMC) methods (Anastasopoulos et al., 2012; Aguero-Valverde, 2013; Wang and Kockelman, 2013; Barua et al., 2014; Dong et al., 2014). In safety literature, a number of model structures have been adopted within the simulation-based multivariate framework including multivariate Poisson regression model, multivariate Poisson lognormal model, multinomial-generalized Poisson model, multivariate Poisson lognormal spatial and/or temporal model, flexible Bayesian semiparametric approach and multivariate random-parameters zero-inflated negative binomial model.

For some specific cases, analytically closed form bundling approaches have also been proposed. These approaches rely on developing multivariate distributions (or approximations of multivariate distributions) with analytical closed form probability expressions that obviate the need for simulation. These model frameworks are estimated employing maximum likelihood or composite maximum likelihood approaches (Wang et al., 2015; Nashad et al., 2016; Yasmin et al., 2018). In safety literature the analytical frameworks adopted include copula-based bivariate negative binomial (NB) model, copula-based multivariate NB model, copula-based ordered logit model and composite maximum likelihood based crash frequency and severity models.

4.2 Current Study

Our proposed research attempts to contribute to simulation-based multivariate approaches by altering how the multiple dependent variables are analyzed. Prior to presenting our alternative approach, challenges with the current simulation-based multivariate approaches in estimating observed and unobserved variable effects are discussed. In multivariate approaches, a separate crash propensity equation is adopted for each crash type. Thus, if there are *D* dependent variables and *K* independent variables, the order of observed parameters estimated in the model structure is of the order of *D***K*. With increasing number of dimensions (*D)*, the number of parameters to be estimated increase rapidly. Thus, in models with $D > 3$, the number of parameters to be estimated are prohibitively high. For example, consider a case of crash frequency for four crash types at an intersection (rear-end, side-swipe, angle and non-motorized). In the univariate models, for each of the crash types, Annual Average Daily Traffic (AADT) is likely to have a statistically significant impact. So, the typical multivariate model estimates 4 parameters for AADT. However, it is possible that the impact of AADT on side-swipe and angle crashes is not statistically different. Testing this is not straightforward in the multivariate model structure. The analyst will need to

modify the model estimation code to restrict the parameters across the side-swipe and angle univariate models to be the same. Subsequently, the restricted model version data fit must be compared with the data fit of the unrestricted version using log-likelihood ratio (LR) test. Based on the result, the analyst can conclude if AADT does offer different impacts for side-swipe and angle crash profiles. Given the additional burden of these steps, the models employed in safety literature typically ignore if the variable impacts are really different across crash type propensities. The result is an ill-specified model structure with too many parameters. To be sure, the model estimates thus obtained are not incorrect. However, the estimation process could become inefficient particularly when sample sizes for crash frequency are small (<1000). The sample sizes for micro-level analysis can typically vary from 200-500 and the number of total parameters estimated has an impact of model estimation efficiency.

In simulation-based multivariate approaches, the influence of unobserved factors is typically accommodated as random effects and correlation parameters across dimensions. The random effects accommodate for the influence of unobserved factors affecting crash propensity within the dimension. The correlation parameters account for the influence of unobserved factors affecting multiple dependent variables. These effects require simulation for parameter estimation. The complexity of the model estimation is dependent on the number of unobserved parameters estimated. With higher dimensions, the model estimation infrastructure can get computationally demanding (while not unmanageable with latest computing power).

In our research, we propose to address these challenges by recasting the multivariate crash frequency modeling problem as a pooled univariate crash frequency (with unobserved heterogeneity accommodated) analysis problem. To elaborate, instead of considering the crash frequency by crash type as a multivariate distribution, we represent it as repeated measures of crash

frequency while recognizing that each repetition represents a different crash type. Thus, in this process we cast a multivariate distribution as a univariate distribution with repeated measures. The recasting will offer multiple advantages. First, the recasting allows us to employ a simple panel random parameter based univariate model code for model estimation. The panel model is substantially easier to program and estimate compared to the multivariate version. Second, instead of estimating crash propensity equations by crash type, a single crash propensity equation that completely generalizes the separate crash propensity equations can be estimated. The consideration of a single crash propensity equation allows the analyst to estimate a base effect for each independent variable and then estimate deviations for different crash types. If the deviation variable for a crash type is statistically insignificant based on the t-statistic the parameter does not exhibit differential sensitivity for the base crash type and crash type for which the deviation was computed. Thus, through this recasting, we are able to replace the parameter by parameter LR test based analysis (discussed earlier) to a simple t-statistic evaluation. Through this approach, the analyst can estimate a parsimonious model without substantial effort and with less computational burden. The reader would note that the multivariate model and the recasted panel univariate model will provide identical data fit with the same number of parameters but with different representation of the parameter effects. Third, the estimation process can use the same infrastructure to estimate random effects and correlation parameters in the proposed pooled model. The only additional burden is associated with creating appropriate variables during data preparation to represent correlation structures. The reader would note that the proposed approach provides exactly the same mathematical formulation by leveraging the panel model structure of the pooled data (with as many records per observation unit as crash types). Such a recasting is only possible in our context because all the univariate dependent variables are assumed to follow the same mathematical

structure. If the simulation-based multivariate model has multiple model structures, then our approach can be customized but will become cumbersome. However, the adoption of different mathematical structures is not common for crash frequency analysis multivariate model contexts.

In summary, the proposed research presents an alternative formulation to analyze multiple crash frequency variables by recasting a multivariate distributional problem as a repeated measure univariate problem. Methodologically, the study presents a first of its kind approach in safety literature to simplify current modeling infrastructure for multivariate analysis. The recasting allows us to estimate parsimonious model systems thus improving parameter estimation efficiency. Further, by simplifying the specification process, it is likely to reduce computational time for estimating parameters associated with unobserved factors. Empirically, the research contributes to our understanding of analyzing zonal level crashes for both motorized and non-motorized road user group while considering different crash types within the motorized category including rear-end, angular, sideswipe, all single vehicle and other multiple vehicle crashes. We employ a panel mixed negative binomial model (PMNB) for examining crash count by different crash types as well as incorporating the presence of unobserved heterogeneity across crash types. The analysis is conducted using the zonal level crash records from Central Florida for the year 2016 considering a comprehensive set of exogenous variables. Further, the study evaluates the performance of the proposed approach by undertaking a comparison exercise with the traditional random parameter multivariate negative binomial model.

The rest of the chapter is organized as follows: The next section presents the methodological framework adopted in the analysis while the 4.4 section provides a detailed description of the model findings. Comparison exercise are discussed in section 4.5 followed by the summary in the last section.

4.3 Econometric Framework

In this section, we briefly provide the details of the model frameworks employed in our study.

4.3.1 Random Parameter Multivariate NB Model

The focus of random parameter multivariate NB (referred as multivariate NB model in the following sections for simplicity) model is to examine number of crashes across different crash types jointly. In our current study context, we consider six different crash types (Five within motorized category: rear-end, angular, sideswipe, all single vehicle and other multiple vehicle crashes; and non-motorized crashes). Thus, in estimating multivariate NB model, we examine six different NB models for six different crash types simultaneously. Let us assume that i ($i = 1,2,3, \dots, N, N = 3{,}815$) be the index for TAZ. Let j be the index representing different crash type, where ($j = 1,2, \dots, J, J = 6$), the index j may take the values of rear-end (j =1), angular (j =2), sideswipe (j =3), all single vehicle (j =4) crashes, other multiple vehicle (j =5) , and non-motorized (j =6) crashes. Using these notations, the equation system for modeling crash count across different crash type j in the usual negative binomial (NB) formulation can be written as:

$$P(c_{ij}|\mu_{ij}, \alpha_j) = \frac{\Gamma\left(c_{ij} + \frac{1}{\alpha_j}\right)}{\Gamma(c_{ij}+1)\Gamma\left(\frac{1}{\alpha_j}\right)}\left(\frac{1}{1+\alpha_j\mu_{ij}}\right)^{\frac{1}{\alpha_j}}\left(1-\frac{1}{1+\alpha\mu_{ij}}\right)^{c_{ij}} \quad (18)$$

where, c_{ij} be the index for crash counts specific to crash type j occurring over a period of time in TAZ i. $P(c_{ij})$ is the probability that TAZ i has c_{ij} number of crashes for crash type j. $\Gamma(\cdot)$ is the gamma function, α_j is NB over dispersion parameter and μ_{ij} is the expected number of crashes occurring in TAZ i over a given time period for crash type j. Further, we can express μ_{ij} as a function of explanatory variables by using a log-link function as follows:

$$\mu_{ij} = E(c_{ij}|\boldsymbol{z}_{ij}) = exp((\boldsymbol{\delta}_j + \boldsymbol{\zeta}_{ij})\boldsymbol{z}_{ij} + \varepsilon_{ij} + \eta_{ij}) \quad (19)$$

where, $\boldsymbol{z}_{ij}$ is a vector of explanatory variables associated with TAZ i and crash type j. $\boldsymbol{\delta}_j$ is a vector of coefficients to be estimated. $\boldsymbol{\zeta}_{ij}$ is a vector of unobserved factors on crash count propensity associated with crash type j for TAZ i and its associated zonal characteristics, assumed to be a realization from standard normal distribution: $\boldsymbol{\zeta}_{ij} \sim N(0, \boldsymbol{\pi}_j{}^2)$. ε_{ij} is a gamma distributed error term with mean 1 and variance α_j. η_{ij} captures unobserved factors that simultaneously impact number of crashes across different crash types for TAZ i. Here it is important to note that the unobserved heterogeneity between total number of crashes across different crash types can vary across TAZs. Therefore, in the current study, the correlation parameter η_{ij} is parameterized as a function of observed attributes as follows:

$$\eta_{ij} = \boldsymbol{\gamma_j} \boldsymbol{s}_{ij} \quad (20)$$

where, $\boldsymbol{s}_{ij}$ is a vector of exogenous variables, $\boldsymbol{\gamma_j}$ is a vector of unknown parameters to be estimated (including a constant). In the current analysis, the multivariate NB model only allows for a positive correlation for total number of crashes across different crash types.

In examining the model structure of crash count across different crash types, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\zeta}$ and $\boldsymbol{\gamma}$ represented by $\boldsymbol{\Omega}$. In this framework, it is assumed that these elements are drawn from independent normal distributions: $\Omega \sim N(0, (\boldsymbol{\pi}_j{}^{\boldsymbol{2}}, \boldsymbol{\sigma}_j^2))$. Thus, conditional on $\boldsymbol{\Omega}$, the likelihood function for the joint probability can be expressed as:

$$L_i = \int_{\boldsymbol{\Omega}} \prod_{j=1}^{J} \left(P(c_{ij})\right) f(\boldsymbol{\Omega}) d\boldsymbol{\Omega} \quad (21)$$

Finally, the log-likelihood function is:

$$LL = \sum_i Ln(L_i) \tag{22}$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 22. The parameters to be estimated in the multivariate NB model are: $\boldsymbol{\delta}_j$, α_j, $\boldsymbol{\pi}_j$, and $\boldsymbol{\sigma_j}$.

4.3.2 Panel Mixed NB Model

The focus of our study is to estimate a panel mixed univariate NB modeling framework. As highlighted earlier, we alter the dataset by taking all six types of crashes as repeated measures (same TAZ is repeated 6 times) of crash frequency in a univariate NB formulation while recognizing that each repetition represents a different crash type. The econometric framework of the proposed approach is presented in this section. Let's assume i $(i = 1,2,3, \dots, N, N = 3{,}815)$ be an index to represent observation unit and $r(r = 1,2, \dots, R, R = 6)$ be an index for different crash type at observation unit i. Then the probability equation of the NB formulation can be rewritten as follow:

$$P(y_{ir}|v_{ir}, \lambda') = \frac{\Gamma\left(y_{ir} + \frac{1}{\lambda'}\right)}{\Gamma(y_{ir}+1)\Gamma\left(\frac{1}{\lambda'}\right)}\left(\frac{1}{1+\lambda' v_{ir}}\right)^{\frac{1}{\lambda'}}\left(1 - \frac{1}{1+\lambda' v_{ir}}\right)^{y_{ir}} \tag{23}$$

where, y_{ir} be the index for crash counts occurring over a period of time in observation unit i and crash type r. $P(y_{ir})$ is the probability that unit i has y_{ir} number of crashes for crash type r. λ' is NB over dispersion parameter and v_{ir} is the expected number of crashes occurring in i over a given time period for crash type r. Similar to the multivariate structure, v_{ir} an be expressed as a function of explanatory variables using a log-link function as follows:

$$v_{ir} = E(y_{ir}|\boldsymbol{x}_{ir}) = exp((\boldsymbol{\beta} + \boldsymbol{\theta}_i + \boldsymbol{\varrho}_{ir})\boldsymbol{x}_{ir} + \varepsilon_{ir}) \tag{24}$$

where, $\boldsymbol{x}_{ir}$ is a vector of explanatory variables associated with observations i for crash type r. $\boldsymbol{\beta}$ is a vector of coefficients to be estimated. $\boldsymbol{\theta}_i$ is a vector of unobserved factors moderating the influence of attributes in $\boldsymbol{x}_{ir}$ on the crash count propensity for analysis unit i, $\boldsymbol{\varrho}_{ir}$ is a vector of unobserved effects specific to crash type r. ε_{ir} is a gamma distributed error term with mean 1 and variance λ'. In estimating the model, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\theta}, \boldsymbol{\varrho}$ represented by Ψ. In this framework, it is assumed that these elements are drawn from independent normal distribution: $\Psi \sim N(0, (\boldsymbol{\pi}'^{\mathbf{2}}, \boldsymbol{\Phi}^{\mathbf{2}}))$.

This $\boldsymbol{\varrho}_{ir}$ will be same across crash types in our case and thus the unobserved heterogeneity across crash types will be captured (same as η_{ij} in the multivariate NB structure). Moreover, $\boldsymbol{\theta}_i$ term will capture the random effect across observations (same as $\boldsymbol{\delta}_j$ in the multivariate structure). The reader would note that, in the multivariate NB model, we can accommodate correlation and attribute variability across different crash type. In the proposed approach, we can do the same by introducing variables specific to crash types (interaction term between crash types and variables). Thus, conditional on Ψ, the likelihood function across TAZ can be expressed as

$$L_i = (\int_{\Psi} \prod_{r=1}^{R} \big(P(y_{ir})\big) f(\Psi) d\Psi \tag{25}$$

Finally, the log-likelihood function is:

$$LL = \sum_i Ln(L_i) \tag{26}$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 26.

4.4 Empirical Analysis

4.4.1 Estimation process

To assist the reader with the model estimation process, we provide a discussion of the various intermediate steps in the estimation process. First, we estimate the traditional multivariate NB model with separate propensity equations for all crash types (Model 1; Table 4.1). Subsequently, we estimate an equivalent panel model with the exact same specification (Model 2; Table 4.2). Then, this specification was employed to drop deviation effects that were insignificant (Model 3; Table 4.3). Finally, we present the net effect of each exogenous variable in the crash propensity equation for representing the model in a similar fashion as model 1 (Model 4; Table 4.4). To facilitate the comparison, let us focus on the variance of speed variable in Models 1, 2, 3 and 4. In model 1, the variable variance of speed has 5 distinct parameters. In Model 2, the same variable has 1 base effect (rear-end serve as the base) and 4 deviation terms. In Model 3, the insignificant deviation terms were dropped to arrive at 2 distinct parameters: 1 base effect (here rear-end, angular, all single vehicle and other multiple vehicle serve as the base) and 1 deviation term for sideswipe crashes. The estimated base effect is 0.032 and the deviations across crash types are: rear-end 0.000, angular 0.000, Sideswipe 0.044, all single vehicle 0.000 and other multiple vehicle 0.000. Finally, in Model 4, we compute the net effect of the variable for each crash type by taking the summation of base effect and deviation corresponds to specific crash types. So, the effect of variance of speed variable for rear-end, angular, all single vehicle and other multiple vehicle would be: 0.032+0.000 = 0.032; and for sideswipe crash, the effect would be 0.032+0.044=0.076.

The reader would note, for simplicity in comparison, we do not add unobserved parameters in the models provided in 5.1 to 5.3.

4.4.2 Model Specification and Overall Measure of Fit

The empirical analysis involves estimation of count models from two approaches: 1) traditional approach - we estimated two models including Independent NB model (separate NB models for 6 different crash types) and Random Parameter Multivariate NB model (RPMNB); and 2) proposed approach - two models are estimated including Independent Panel NB model (counterpart of Independent NB model in the traditional approach) and Panel Mixed NB (PMNB) model (counterpart of RPMNB in the traditional approach).The reader would note that the model estimation in the proposed approach is informed from the traditional approach models (particularly for the independent models). To elaborate, observing the model specifications in the independent models, we identify potential parameters that can be restricted to be the same across various crash types and test that restriction in our proposed model system. Subsequently, we estimate a base effect for each exogenous variable that is common across crash types and then, we estimate the deviation for each crash type relative to the base effect. Given we have 6 total crash types, we typically can estimate 5 deviations from the base effect. The t-statistic of the estimated parameters will provide evidence if the deviation term offers a statistically significant difference from the base effect. If the deviation variable for a crash type is statistically insignificant based on the t-statistic, the parameter does not exhibit differential sensitivity for the base crash type and crash type for which the deviation was computed. The reader would note that for some exogenous variables, the overall parameters estimated for an exogenous variable could vary from 0 (i.e. the variable has no impact across crash types) to 6 (i.e. the variable has a statistically distinct effect for every crash type). Typically, models estimated within the panel formulation have fewer parameters. To facilitate the reader's understanding of the overall model estimation, Appendix A provides details of the intermediate steps in the estimation process.

The log-likelihood values at convergence for the final estimated models are: For traditional approach, (a) Independent NB model (89 parameters) is -44,791.54, and (b) RPMNB model (92 parameters) is -43,597.82; and for proposed approach, (a) Independent Panel NB model (58 parameters) is –44,808.32, and (b) PMNB model (61 parameters) is -43,622.57. We also compute the Bayesian Information Criterion (BIC) (lower is better) for these four models. For the traditional models, the corresponding BIC values are 90,317.02 (Independent NB) and 87,954.34 (RPMNB) respectively. On the other hand, for the proposed frameworks, the BIC values are as follows: 90,094.95 (Independent Panel NB), and 87,748.19 (PMNB model). Based on the BIC values, two observations can be made. First, models accommodating unobserved effects perform better than their corresponding independent models (in both traditional and proposed regimes) highlighting the importance of accommodating for unobserved heterogeneity in examining crash count by different crash types. Second, our proposed approach provides superior fit compared to its' counterparts in the traditional frameworks (Independent Panel NB vs Independent NB and PMNB vs RPMNB) when accounting for penalty for additional parameters. Thus, our proposed approach allows us to estimate parsimonious model systems with more efficient parameter estimation.

4.5 Model Estimation Result

This section presents a detailed discussion of the factors affecting crash count components across different crash types. Table 4.4 presents the model estimation results for the proposed panel mixed NB model. The estimation results of the multivariate NB model are presented in Table 4.5 for comparison. For the sake of brevity, we do not discuss these parameter estimates.

As discussed before, in presenting our model results, we have selected a representation that provides results similar to the traditional model approach i.e. present the net effect of each exogenous variable in the crash propensity equation. For example, consider the constants estimated

in the various crash type propensity equations. The proposed estimated the base effect as -1.074 and the deviations across crash type as – rear-end 0.000, angular -0.716, Sideswipe -0.907, All single vehicle 2.137, Other multiple vehicle -1.172, and Non-motorized -2.109. The reader would note that the "rear-end" crash type served as the base. The model results presented compute the net effect for each crash type. For non-motorized crash type this would be computed as -1.704 (base) + -2.109 (non-motorized deviation) = -3.841. The consolidation of parameters in this manner allows an easy comparison with the traditional approach. The consolidation of parameters in this manner allows an easy comparison with the traditional approach. At the same time, to highlight the gains in parameters if any, we identify the number of parameters estimated across the crash types (range between 1 and 6). In cases where the deviation for a crash type was insignificant, the reader would notice a common coefficient across 2 or more crash types. The number of distinct parameters estimated provides a guide to the improvement in model estimation attained by the proposed model structure. For instance, the variable length of divided roads offers an important comparison across the two models (see Table 4.4 and 4.5). In our proposed model, we estimated a single parameter across 5 crash types while the same variable results in five distinct parameters across 5 crash types in the traditional multivariate model. The variable impact illustrates how our proposed approach allows for parsimonious specification while not compromising on model explanatory power. Finally, the reader would note that for some exogenous variables, a common base effect might not be statistically significant. In such cases, the exogenous variable is considered by crash type to test for the variable impact.

A positive (negative) sign for a variable in the crash count component of Table 4.4 indicates that an increase in the variable is likely to result in more (less) crashes.

4.5.1.1 Crash Specific Constants

The crash specific constants represent the intercept of crash propensity after adding the various exogenous variables and do not have any substantive interpretation.

4.5.1.2 Roadway Attributes

The parameter associated with proportion of arterial roads offers a positive impact (with same magnitude) on crash count propensity for rear-end, angular, sideswipe and non-motorized crashes indicating a higher likelihood of crashes with increased proportion of arterial roads in a TAZ. On the other hand, with respect to all single vehicle crashes, the impact is negative revealing a reduced incidence of all single vehicle crashes with higher proportion of arterial roads. This is intuitive as off-road and rollover crashes (these are combined in all single vehicle crashes) are likely to be associated with high vehicular speed and on arterial roads drivers are likely to drive at lower operating speeds. Number of intersections are found to positively influence angular, other multiple vehicle and non-motorized crashes indicating a higher likelihood of crash occurrence for these three crash types in a zone with increased number of intersections. It is also found that the impact is not statistically different for angular and non-motorized crashes. The results are in line with earlier research specific to angular and non-motorized crashes (Abdel-Aty and Wang, 2006; Reynolds et al., 2009). In terms of variance of speed, the estimated result shows that a TAZ with higher variance in speed limit is likely to result in higher crash risk across all crash types except non-motorized crashes. Among these effects, the magnitude of impact is larger for sideswipe crashes and remains the same across other four crash types

In terms of length of divided roads, the variable is found to have the same positive effect on all crash types except non-motorized crashes. Signal intensity in the zone reveals a negative

association with sideswipe and all single vehicle specific crashes indicating a reduced occurrence of sideswipe and all single vehicle crashes in a zone with higher number of signals. This is expected because, vehicles are likely to drive at a lower speed in the location with higher number of signals and as a result, the risk of motorized off-road crashes reduces. Average outside shoulder width has a negative influence on crash risk propensity for rear-end, angular, sideswipe and other multiple vehicle crashes which is perhaps indicating greater safety margins for vehicular maneuverability.

The estimated results show that a TAZ with higher proportion of roads over 55mph speed limit is likely to experience increased number of rear-end, sideswipe and all single vehicle crashes while a negative effect is observed for angular and non-motorized crashes. Further, we found that proportion of road over 55mph has significant variability specific to angular crashes as indicated by the standard deviation parameter. The reader would note that the distributional parameter indicates that the overall impact of the variable on angular crashes is likely to be negative (80%). With respect to sidewalk width, the variable is found to be significant in rear-end crash component with a positive impact while a negative association is observed for the non-motorized crashes. The results are contrary to some of the earlier studies (Aguero-Valverde and Jovanis, 2006; Cai et al., 2016; Dong et al., 2014). However, there is a reasonable explanation for the effects identified. Increasing sidewalk width is a surrogate for non-motorized activity in the zone. The presence of non-motorists can potentially increase rear-end crashes at as vehicles might stop abruptly to allow for non-motorist movement increasing rear-end crash risk. Also, the presence of a wider side walk provides additional margin of safety for non-motorists from colliding with a motorized vehicle and thus results in reduced risk for non-motorized users in the zone.

4.5.1.3 Traffic Characteristics

As expected, the coefficient associated with VMT offers a positive impact on the crash risk component of angular, sideswipe, other multiple vehicle and non-motorized crashes while the likelihood of all single vehicle crashes will go down with higher VMT. VMT mainly reflects the exposure measure for traffic volume and therefore, with increased VMT, the probability of getting involved in a crash is likely to be higher. However, with increased traffic volume, the likelihood of speeding is lower which eventually results in reduced number of all single vehicle crashes. Truck VMT is found to positively influence the rear-end and all single vehicle crash propensity indicating a higher risk of getting involved in rear-end and all single vehicle specific crashes with increased proportion of trucks on the road.

4.5.1.4 Land-use Attributes

From Table 4.4, we can observe that TAZs with higher urbanized and office areas are likely to experience more crashes specific to all crash types. This is expected as urban area serves as an additional surrogate for exposure for traffic. Moreover, the impact of urban area specific to rear-end crash is of higher magnitude relative to other crash types signifying that rear-end crash is a prominent safety issue in urban areas. Institutional areas are associated with increased crash risks for rear-end, angular, other multiple vehicle and non-motorized crash. The variable also illustrated the advantages of our proposed approach. Specifically, in our proposed framework, we estimate a total of two parameters for the variable. However, in the traditional multivariate structure, four distinct parameters were estimated. Residential area has a significant negative impact for rear-end, angular and sideswipe crashes.

4.5.1.5 Built Environment Characteristics

In terms of built environment attributes, we considered a number of variables, among which only number of restaurants and number of shopping centers have significant impact on zonal level crash risks. The coefficient associated with number of restaurants reveals the higher likelihood of crash propensity of all crash types with increased number of restaurants in a TAZ. On the other hand, a zone with higher number of shopping centers is likely to experience an increased number of rear-end and angular crashes relative to other zones.

4.5.1.6 Sociodemographic Characteristics

With respect to sociodemographic characteristics, population density – another surrogate for exposure – is positively associated with increased likelihood of crash risk for all crash types. We can also observe that the parameter associated with the number of non-motorist commuters in the TAZ reveals a higher probability of crash risk for rear-end, sideswipe and non-motorized crashes in the TAZ. In fact, the reader would note that the magnitude of these impacts is same across the three crash types in the current study context. Further, the coefficient specific to proportion of households without vehicle indicates that the variable is negatively associated with rear-end and sideswipe (motorized) crashes but has a positive impact on non-motorist road user group. The result is expected as people from households without access to personal vehicles experience higher exposure for non-motorized crashes as they are restricted to using public transport, walk or bike as their primary mode of transportation.

4.5.1.7 Unobserved Heterogeneity

The final set of variables in Table 4.4 correspond to the unobserved heterogeneity across zones. The reader would not that, in estimating the model, we found two common unobserved components[3] including (1) common unobserved factors affecting rear-end and non-motorized crashes and (2) common unobserved factors affecting angular, sideswipe and all single vehicle crashes. These parameter estimates lend support to the presence of unobserved heterogeneity across different crash type.

4.6 Model Comparison Exercise

4.6.1 Predictive Performance

In an effort to assess the predictive performance of the estimated models, we compute several goodness fit of measures at disaggregate level including MPB (Mean prediction bias), MAD (mean absolute deviation), MAPE (mean absolute percentage error), RMSE (Root mean square error) and predictive BIC (please see (Bhowmik et al., 2018) for a discussion on estimating these measures). Specifically, we employ these measure on two datasets: 1) in-sample dataset: for the records used in the model estimation (sample size = 3,815 TAZs) and 2) holdout sample: records that are set aside for validation analysis (sample size = 932 TAZs). The reader would note that model with lower value of predictive measures and BIC will reflect better performance in terms of prediction and statistical fit relative to the observed data. Table 4.6 presents the values of these measures for Random parameter multivariate NB and Panel mixed NB models for both in-sample and holdout-sample measures. From Table 4.6, we can observe that the performance of the two

[3]The same correlation structure was revealed from the traditional multivariate model structure (as shown in Table 4.5).

models across various prediction measures are quite similar even though there is a large difference in the number of parameters between the two specifications (92 vs 61). Further, RPMNB model performs marginally better than the proposed framework for the deviation measures with respect to angular, sideswipe, other multiple vehicle and non-motorized crashes while in terms of rear-end and all single vehicle crashes, the proposed approach offers better performance (for both in-sample and holdout samples). These deviation measures do not consider the difference in number of parameters across the two models. The BIC measure that penalizes additional parameters clearly shows that the proposed panel model structure offers improved statistical fit. In summary, the resulting goodness of fit measures clearly highlight the comparable performance offered by the proposed framework compared to the commonly used RPMNB model even with substantially fewer parameters.

To further evaluate the predictive performance of the estimated models, we carried out a comparison exercise between the random parameter negative binomial model and panel mixed NB model by predicting the crash frequencies across different count events for different crash types. For this purpose, 20 data samples with 250 records (TAZs) each, are randomly generated from the holdout validation sample consisting of 932 records (TAZs). For these samples, we predict the number of TAZs from both models (RPMNB and PMNB) for different count events across different crash types. These counts are employed to generate the ratio of predicted and observed counts specific to each level (count groups and crash types). A value of 1 for the ratio would imply a prefect prediction. For example, if there are 100 TAZs with 0 rear end crashes in data sample 1 and we predict 60 and 50 TAZs from RPMNB and PMNB model respectively, then the estimated ratio of these models will be 0.6 (60/100) and 0.5 (50/100) respectively. For both models, two box plots are generated using all the data samples (for every count event, there are 20 points) by each

count group and crash type. Figure 4.1 to 4.3 represents the ratio statistics for different crash types. From Figure 4.1, we can see that while the models might under-predict or over-predict crash counts, the performance of the two models are quite similar. Thus, one can conclude that the proposed approach has offered equivalent predictions relative to the multivariate NB model despite with substantially fewer model parameters (31 less parameters to be precise).

4.6.2 Elasticity Effect

The parameters of the exogenous variables in Table 4.4 and 4.5 do not directly provide the exact magnitude of the effects of variables on the zonal level crash counts across different crash types. However, it might be possible that the effects (exact magnitude) of some attributes could differ considerably across the two frameworks. To evaluate this, we compute aggregate level elasticity effects for both PMNB and RPMNB models. For this purpose, we identify a subset of exogenous variables including proportion of arterial roads, length of divided roads, proportion of roads over 55mph, institutional areas and number of non-motorist commuters. In our study, we investigate the effect as percentage change in the expected zonal level crash counts in response to the increase of the explanatory variable by 10% (see Eluru and Bhat, 2007 for a discussion on the methodology for computing elasticities). The numbers in Figure 4.4 can be interpreted as the percentage change in the expected crash counts (increase for positive sign and decrease for negative sign) due to the change in the exogenous variable for different crash types. For instance, the elasticity estimates generated from the proposed PMNB (RPMNB) model for proportion of arterial roads variable in rear-end crashes indicates that the expected mean rear-end crash will increase by 0.656% (1.038%) for an 10% increase in the proportion of arterial roads.

Several observations can be made based on the elasticity effects presented in Figure 4.4. First, in general, we do not observe any large differences in the elasticity effects of the two models

across different crash types. From the five variables considered for our elasticity exercise, a substantial number of the effects (14 out of 22) offer very little differences. Second, the PMNB model with fewer parameters is able to represent the substantial differences in the elasticity effects for the same variable across different crash types. For instance, the elasticity effect for length of divided roads variable is different across the five crash types despite estimating a single parameter (same impact in magnitude) for the variable across the five crash types. Third, for some variables, we found substantial differences in the elasticity effects across the two frameworks for different crash types. For example, in case of rear-end crashes, the proposed PMNB model predicts an 0.65% increase in the expected mean for 10% increase in the proportion of roads over 55mph while we found an increase of 0.92% from the RPMNB model. Such differences could be attributed to the non-linearity embedded within the two model structures estimated with similar data fit. In summary, the proposed framework allows for a parsimonious specification without compromising the model explanatory power and provides similar performance (most of the times) as the most traditional multivariate NB model.

4.7 Summary

In our current research effort, a simple random parameter based univariate model code was employed to analyze zonal level crash counts for different crash types including rear-end, angular, sideswipe, all single vehicle, other multiple vehicle and non-motorized crashes. The empirical analysis was based on the traffic analysis zone (TAZ) level crash count data from Central Florida for the year 2016. A host of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics were considered in the current research effort. A comprehensive comparison of the proposed model with the most commonly used multivariate negative binomial (NB) model was conducted. The comparison exercise based on the BIC value

clearly highlighted the superiority of the proposed approach over the traditional multivariate formulation in terms of data fit. The comparison exercise was further augmented by generating several predictive measures for both estimation and holdout samples. Based on the resulting fit measures, the study concludes that the proposed formulation has offered equivalent predictions relative to the most traditional multivariate NB model even though there is a significant difference in the number of parameters within these two frameworks (61 vs 92). Further, we compute aggregate level elasticity effects for both PMNB and RPMNB models to quantify whether the effect of variables significantly differs across the two frameworks. For this purpose, we identify a subset of exogenous variable including proportion of arterial roads, length of divided roads, proportion of roads over 55mph, institutional areas and number of non-motorist commuters. The elasticity results clearly indicate that for most of the variables, the effects are quite similar for both models across different crash types. However, for some variables, we found some significant and substantial differences in the elasticity effects across the two frameworks for some crash types. Such differences could be attributed to the non-linearity embedded within the two model structures estimated with similar data fit.

The current research effort contributes to literature on crash frequency analysis by suggesting an alternative and mathematically simpler approach for analyzing multiple crash frequency variables for the same study unit. Specifically, the proposed framework while simplifying the model estimation process, allows for parsimonious specification without compromising the model explanatory power and provides similar performance (predictions) as the currently employed multivariate NB model. In conclusion, the aim of the proposed scheme is to augment the inventory of crash frequency models with an alternative formulation and serves as a

viable approach to reduce the parameter explosion that is common within a multivariate NB model with large number of dependent variable dimensions.

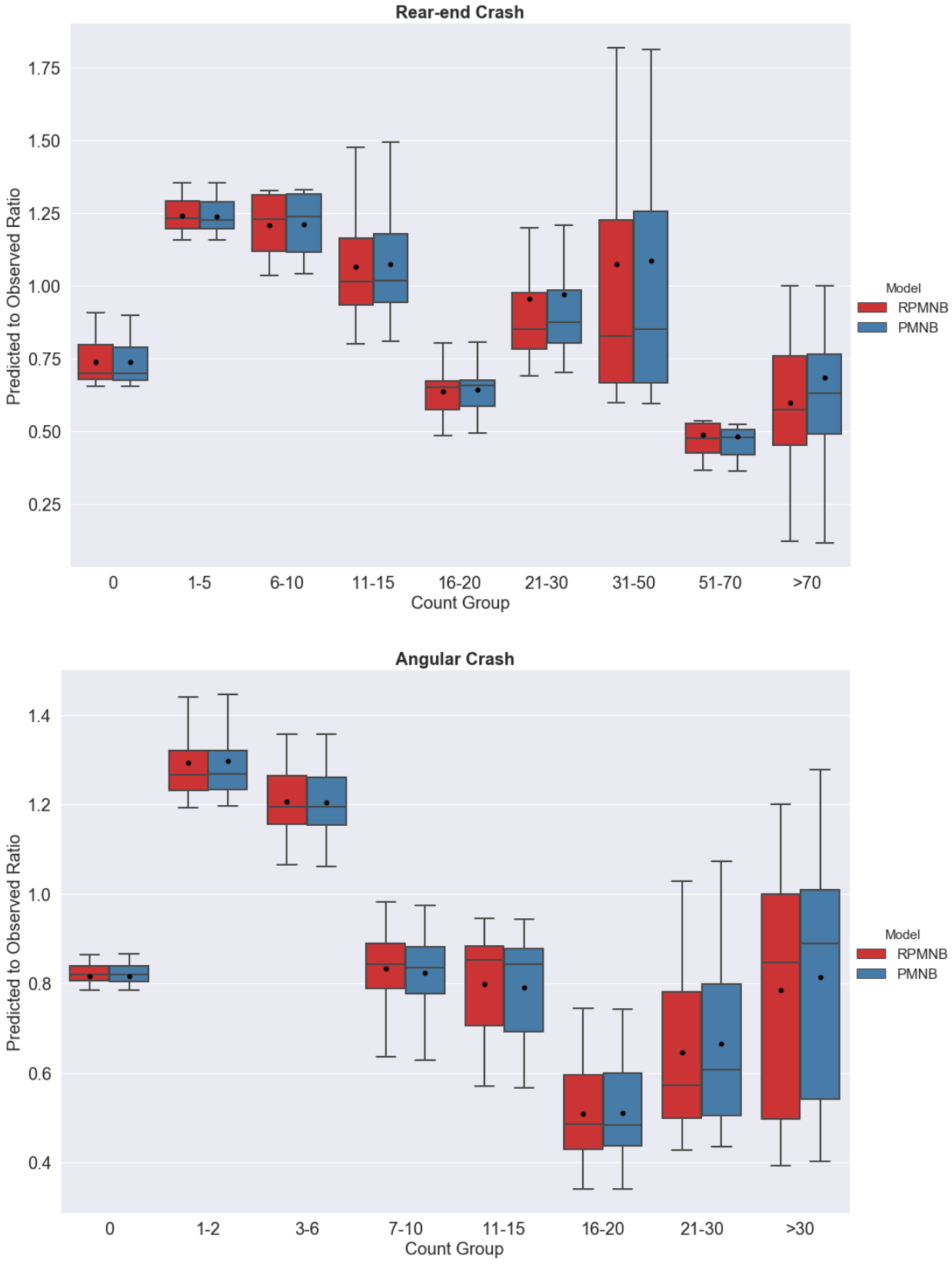

Figure 4.1 Predicted to Observed Ratio for Rear-end and Angular Crashes.

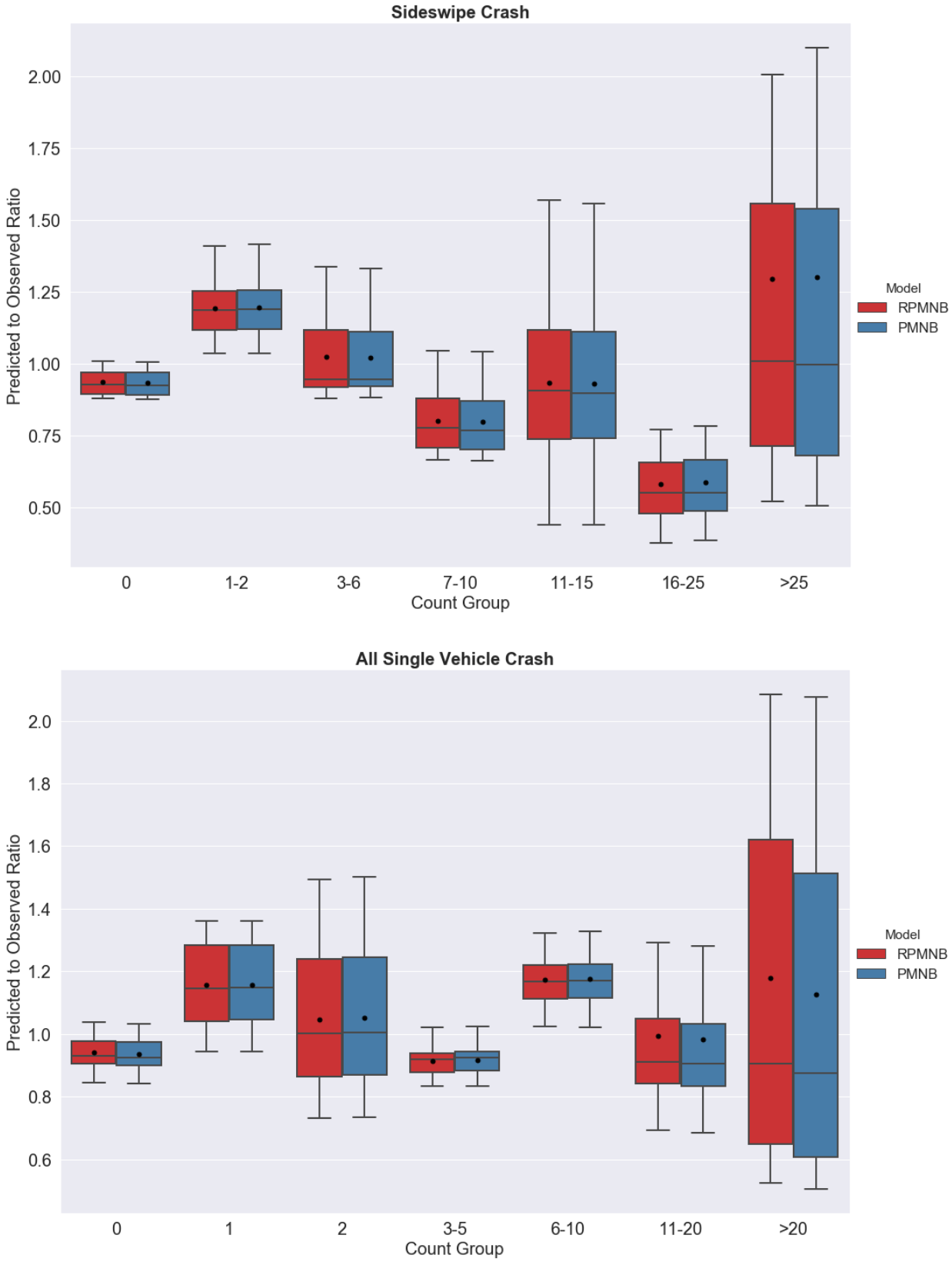

Figure 4.2 Predicted to Observed Ratio for Sideswipe and All Single Vehicle Crashes.

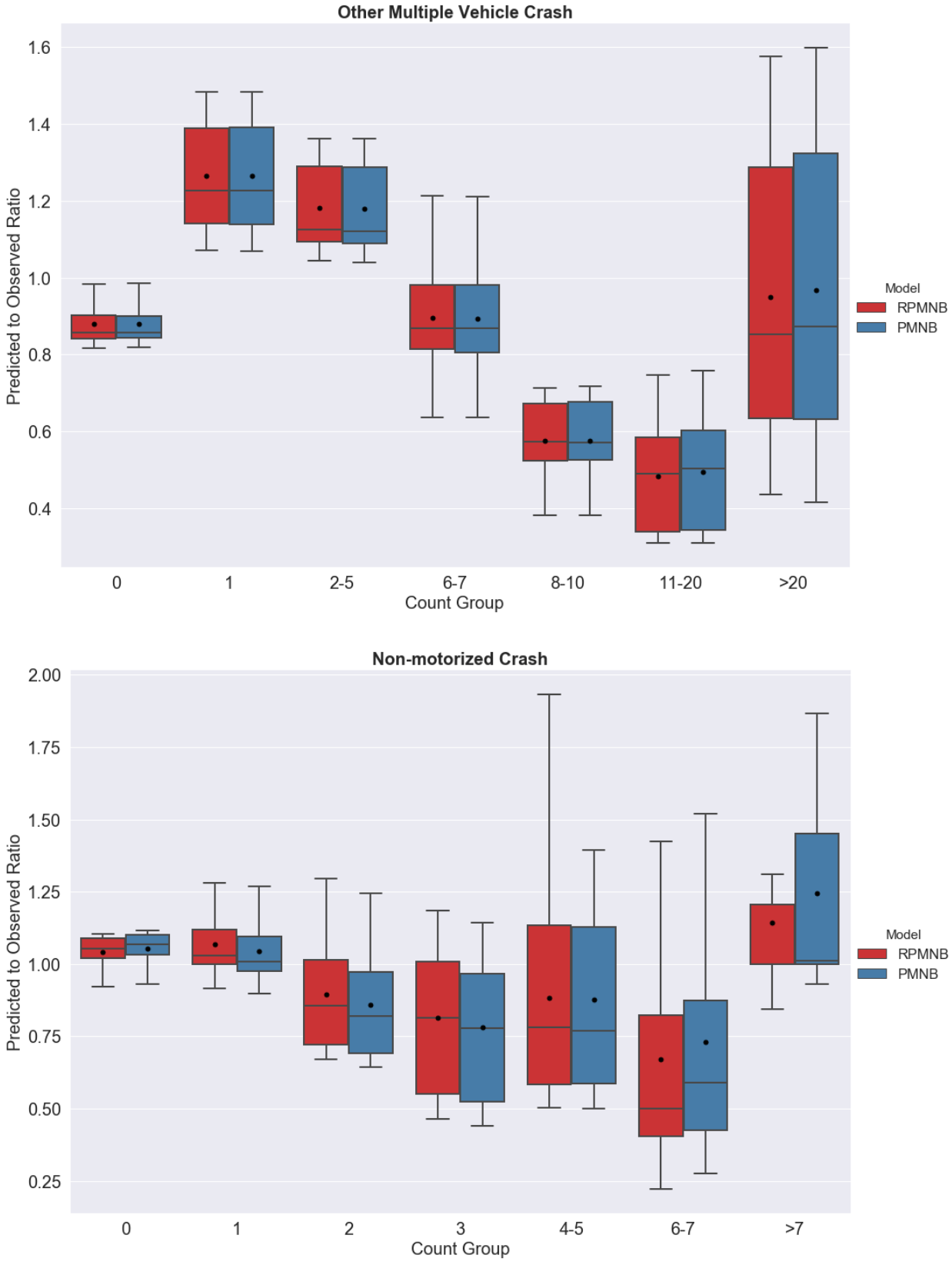

Figure 4.3 Predicted to Observed Ratio for Other Multiple Vehicle and Non-motorized Crashes.

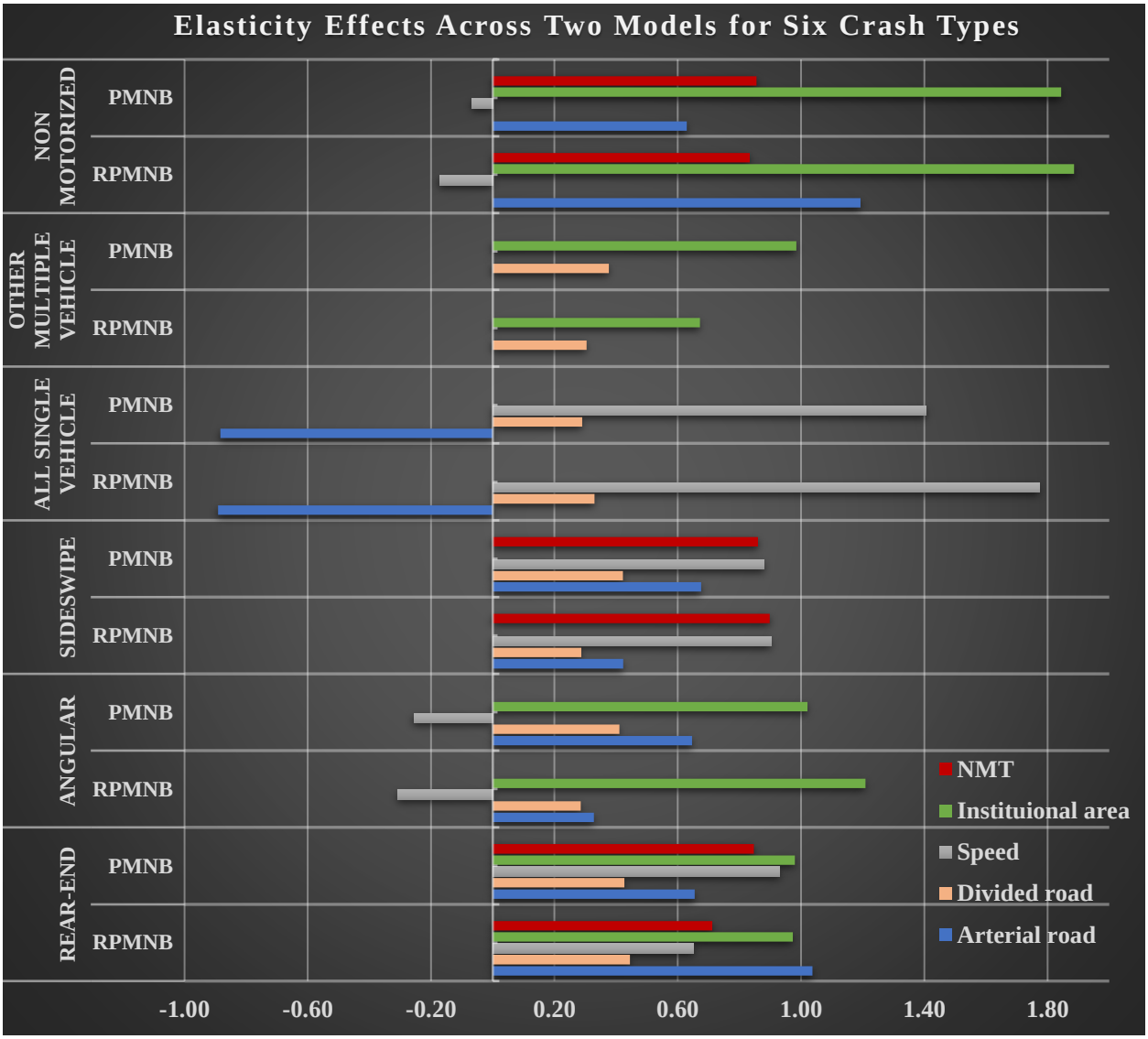

Figure 4.4 Elasticity Effects Across Two Models (PMNB and RPMNB) for Six Crash Types

Table 4.1 Model 1: Traditional Multivariate Model with Distinct Propensity Equations

Variables[4]	Rear End		Angular		Sideswipe		All single vehicle		Other multiple vehicle		Non-motorized	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Constant	-0.770	-10.181	-1.301	-14.573	-2.161	-16.951	-0.612	-7.364	-1.261	-14.781	-3.391	-25.776
Roadway Characteristics												
Proportion of arterial roads	0.232	5.769	0.114	2.082	0.123	1.770	-0.232	-4.805	--	--	0.265	3.757
Number of intersections	--	--	0.235	7.037	--	--	--	--	0.087	3.147	0.243	5.239
Variance of speed	0.037	3.560	0.040	3.263	0.075	5.153	0.021	1.971	0.032	2.830	--	--
Length of divided road	0.478	2.427	0.357	1.978	0.361	1.668	0.512	3.590	0.458	2.528	--	--
Signal intensity	--	--	--	--	-0.753	-3.330	-0.632	-2.704		--	--	--
Average outside shoulder width	-0.420	-7.493	-0.135	-3.078	-0.321	-5.840	--	--	-0.072	-1.891	--	--
Road length over 55mph	0.900	7.911	-0.424	-2.509	1.165	6.711	1.245	10.174	--	--	-0.469	-1.923
Sidewalk width	0.104	3.859	--	--	--	--	--	--	--	--	-0.071	-2.874
Traffic Characteristic												
VMT	--	--	0.060	4.504	0.187	12.395	-0.111	-4.127	0.094	7.979	0.061	3.300
Truck VMT	0.183	20.085	--	--	--	--	0.325	11.169	--	--	--	--
Land-use attributes												
Urban area	0.169	17.080	0.117	8.812	0.132	7.968	0.063	6.286	0.082	6.304	0.158	7.629
Office area	0.201	15.566	0.226	14.091	0.221	10.625	0.087	6.602	0.157	9.691	0.158	7.414
Institutional area	0.046	3.342	0.079	4.996	--	--	--	--	0.054	3.892	0.113	5.683
Residential area	-0.064	-7.251	-0.025	-2.128	-0.103	-6.621	--	--				
Built environment characteristic												
No. of restaurant	0.226	8.275	0.222	8.124	0.318	11.882	0.102	6.062	0.292	11.228	0.212	9.009

[4] Please see Table 2.3 for variable definitions and units

No. of shopping center	0.074	2.842	0.067	1.721	--	--	--	--	--	--	--	--
Socio-demographic characteristics												
Population density	0.148	15.432	0.127	16.045	0.129	11.311	0.027	3.675	0.105	14.110	0.126	11.010
Non-motorist commuter	0.037	2.096	--	--	0.055	2.381	--	--	--	--	0.041	1.770
Proportion of household without vehicle	-0.463	-1.683	--	--	-0.646	-1.871	--	--	--	--	2.508	6.609
Over dispersion	0.943	36.926	0.729	23.693	0.946	20.026	0.491	20.471	0.557	18.801	0.427	9.019
Total number of parameters = 89, Log-likelihood: -44,791.53; AIC: 89,761.07; BIC:90,317.02												

Table 4.2 Model 2: Panel Model with Same Specification as Model 1

Variables (Base in Overall Crash Risk Component)	Overall Crash Risk	Deviation					
		Rear End (1)	Angular (2)	Sideswipe (3)	All single Vehicle (4)	Other Multiple Vehicle (5)	Non-motorized (6)
	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)
Constant	-3.390 (-22.411)	2.617 (13.917)	2.091 (11.546)	1.229 (6.025)	2.778 (15.951)	2.129 (11.989)	--
Roadway Characteristics							
Proportion of arterial roads (1)	0.232 (3.969)	--[*]	-0.118 (-1.287)	-0.110 (-0.927)	-0.464 (-5.913)	N/I[**]	0.033 (0.336)
Number of intersections (6)	0.243 (5.228)	N/I	-0.008 (-0.126)	N/I	N/I	-0.156 (-2.541)	--
Variance of speed (1)	0.037 (2.633)	--	0.003 (0.159)	0.038 (1.639)	-0.016 (-0.890)	0.005 (0.260)	N/I
Length of divided roads (1)	0.476 (1.723)	--	-0.117 (-0.278)	-0.112 (-0.235)	0.034 (0.093)	-0.016 (-0.046)	N/I
Signal intensity (3)	-0.752 (-2.821)	N/I	N/I	--	0.122 (0.368)	N/I	N/I
Average outside shoulder width (1)	-0.421 (-6.389)	--	0.286 (3.513)	0.100 (1.040)	N/I	0.349 (4.038)	N/I
Roads length over 55mph (1)	0.903 (5.407)	--	-1.328 (-5.063)	0.261 (0.862)	0.343 (1.179)	N/I	-1.367 (-4.487)
Sidewalk width (1)	0.105 (3.629)	--	N/I	N/I	N/I	N/I	-0.176 (-6.962)
Traffic Characteristic							
VMT (2)	0.060 (5.128)	N/I	--	0.128 (6.975)	-0.170 (-6.620)	0.035 (1.911)	-0.002 (-0.071)
Truck VMT (1)	0.183 (15.736)	--	N/I	N/I	0.142 (4.699)	N/I	N/I
Land-use Attributes							
Urban area (1)	0.170 (13.060)	--	-0.053 (-2.620)	-0.038 (-1.746)	-0.106 (-5.839)	-0.087 (-5.139)	-0.012 (-0.418)
Office area (1)	0.201 (10.952)	--	0.025 (1.327)	0.020 (0.833)	-0.114 (-4.775)	-0.044 (-1.811)	-0.044 (-1.612)

Institutional area (1)	0.046 (2.675)	--	0.034 (1.156)	N/I	N/I	0.008 (0.330)	0.068 (2.318)
Residential area (1)	-0.063 (-5.321)	--	0.038 (2.098)	-0.040 (-1.212)	N/I	N/I	N/I
Built Environment Characteristic							
No. of restaurant (1)	0.226 (5.106)	--	0.003 (0.044)	0.092 (1.651)	-0.124 (-2.424)	0.067 (1.758)	-0.014 (-0.279)
No of shopping center (1)	0.074 (1.802)	--	-0.007 (-0.105)	N/I	N/I	N/I	N/I
Socio-demographic Characteristics							
Population density (1)	0.148 (10.789)	--	-0.021 (-1.107)	-0.019 (-0.750)	-0.121 (-7.447)	-0.043 (-2.534)	-0.022 (-1.121)
Non-motorist commuters (1)	0.036 (2.494)	--	N/I	0.019 (0.401)	N/I	N/I	0.005 (0.145)
Proportion of household without vehicle (1)	-0.437 (-1.730)	--	N/I	-0.213 (-0.267)	N/I	N/I	2.935 (4.606)
Over dispersion	--	0.943 (32.137)	0.729 (24.060)	0.946 (21.171)	0.491 (23.583)	0.557 (21.641)	0.427 (9.987)
Total number of parameters = 89, Log-likelihood: -44,791.53; AIC: 89,761.07; BIC:90,317.02							

Table 4.3 Model 3: Parsimonious Model Specification Dropping Insignificant Variables from Model 2

Variables (Base in Overall Crash Risk Component)	Overall Crash Risk	Deviation					
		Rear End (1)	Angular (2)	Sideswipe (3)	All single Vehicle (4)	Other Multiple Vehicle (5)	Non-motorized (6)
	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)
Constant	-3.448 (-33.249)	2.699 (24.465)	2.113 (16.499)	1.245 (8.111)	2.867 (22.165)	2.187 (15.938)	--
Roadway Characteristics							
Proportion of arterial roads (1-3,6)	0.179 (9.674)	--[*]	--	--	-0.403 (-7.413)	N/I[**]	--
Number of intersections (2,6)	0.242 (9.471)	N/I	--	N/I	N/I	-0.159 (-3.273)	--
Variance of speed (1,2,4,5)	0.032 (7.566)	--	--	0.044 (2.389)	--	--	N/I
Length of divided roads (1-5)	0.451 (9.257)	--	--	--	--	--	N/I
Signal intensity (3-4)	-0.685 (-6.538)	N/I	N/I	--	--	N/I	N/I
Average outside shoulder width (1,3)	-0.351 (10.229)	--	0.223 (3.895)	--	N/I	0.278 (4.712)	N/I
Roads length over 55mph (1,3,4)	1.109 (21.579)	--	-1.489 (-9.475)	--	--	N/I	-1.494 (-4.487)
Sidewalk width (1)	0.076 (3.088)	--	N/I	N/I	N/I	N/I	-0.151 (-6.958)
Traffic Characteristic							
VMT (2,6)	0.060 (7.079)	N/I	--	0.126 (8.932)	-0.175 (-7.508)	0.035 (2.332)	--
Truck VMT (1)	0.186 (18.694)	--	N/I	N/I	0.141 (5.173)	N/I	N/I
Land-use Attributes							
Urban area (1,6)	0.174 (17.160)	--	-0.056 (-3.283)	-0.049 (-2.531)	-0.113 (-8.161)	-0.092 (-6.200)	--
Office area (1-3)	0.216 (31.945)	--	--	--	-0.132 (-8.670)	-0.061 (-3.337)	-0.061 (-2.784)

Institutional area (1,2,5)	0.063 (9.772)	--	--	N/I	N/I	--	0.068 (2.318)
Residential area (1,3)	-0.079 (-13.480)	--	0.059 (4.234)	--	N/I	N/I	N/I
Built Environment Characteristic							
No. of restaurant (1,2,6)	0.219 (14.863)	--	--	0.099 (2.336)	-0.117 (-4.296)	0.074 (2.498)	--
No of shopping center (1,2)	0.076 (6.523)	--	--	N/I	N/I	N/I	N/I
Socio-demographic Characteristics							
Population density (1,2,3,6)	0.134 (34.649)	--	--	--	-0.109 (-11.870)	-0.029 (-2.412)	--
Non-motorist commuters (1,3,6)	0.043 (5.163)	--	N/I	--	N/I	N/I	--
Proportion of household without vehicle (1,3)	-0.476 (-2.444)	--	N/I	--	N/I	N/I	3.044 (6.321)
Over dispersion	--	0.948 (32.462)	0.731 (24.381)	0.951 (21.434)	0.490 (23.749)	0.557 (21.905)	0.433 (10.217)
Total number of parameters = 58, Log-likelihood: -44,808.32; AIC: 89,732.64; BIC:90,094.95							

Table 4.4 Panel Mixed NB Model (PMNB) Estimation Results

Variables	No. of Param*	Rear End	Angular	Sideswipe	All single vehicle	Other multiple vehicle	Non-motorized
		Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)
Constant	6	-1.074 (-12.165)	-1.790 (-22.389)	-2.697 (-25.745)	-0.560 (-8.412)	-1.732 (-21.046)	-3.841 (-34.651)
Roadway Characteristics							
Proportion of arterial roads	2	0.134 (5.545)	0.134 (5.545)	0.134 (5.545)	-0.230 (-5.266)	--	0.134 (5.545)
Number of intersections	2	--[5]	0.305 (13.006)	--	--	0.173 (6.503)	0.305 (13.006)
Variance of speed	2	0.031 (6.845)	0.031 (6.845)	0.072 (6.379)	0.031 (6.845)	0.031 (6.845)	--
Length of divided roads	1	0.456 (6.346)	0.456 (6.346)	0.456 (6.346)	0.456 (6.346)	0.456 (6.346)	--
Signal intensity	1	--	--	-0.585 (-5.613)	-0.585 (-5.613)	--	--
Average outside shoulder width	3	-0.386 (-11.366)	-0.166 (-4.576)	-0.386 (-11.366)	--	-0.099 (-2.633)	--
Road length over 55mph	3	1.039 (21.596)	-0.516 (-4.090)	1.039 (21.596)	1.039 (21.596)	--	-0.139 (-1.717)
Standard deviation	1	--	0.622 (3.040)	--	--	--	--
Sidewalk width	2	0.089 (3.401)	--	--	--	--	-0.085 (-4.136)
Traffic Characteristic							
VMT	4	--	0.065 (7.910)	0.211 (21.727)	-0.118 (-5.491)	0.087 (9.454)	0.065 (7.910)
Truck VMT	2	0.209 (18.563)	--	--	0.332 (13.182)	--	--
Land-use attributes							
Urban area	5	0.173 (13.355)	0.125 (9.322)	0.134 (8.675)	0.060 (7.964)	0.092 (7.916)	0.173 (13.345)
Office area	4	0.234 (30.359)	0.234 (30.359)	0.234 (30.359)	0.083 (6.931)	0.169 (13.301)	0.161 (8.505)
Institutional area	2	0.063 (7.291)	0.063 (7.291)	--	--	0.063 (7.291)	0.109 (5.942)
Residential area	2	-0.085 (-14.223)	-0.023 (-2.668)	-0.085 (-14.223)	--	--	--
Built environment characteristic							

[5] -- = attribute insignificant at 90% significance level

No. of restaurants	4	0.241 (19.756)	0.241 (19.756)	0.301 (17.306)	0.101 (5.017)	0.265 (19.626)	0.241 (19.756)
No of shopping center	1	0.022 (1.932)	0.022 (1.932)	--	--	--	--
Socio-demographic characteristics							
Population density	3	0.142 (32.333)	0.142 (32.333)	0.142 (32.333)	0.023 (2.944)	0.118 (16.551)	0.142 (32.333)
Non-motorist commuter	1	0.042 (4.013)	--	0.042 (4.013)	--	--	0.042 (4.013)
Proportion of households without vehicle	2	-0.760 (-3.938)	--	-0.760 (-3.938)	--	--	2.447 (6.409)
Over dispersion	6	0.523 (25.262)	0.184 (10.107)	0.291 (11.621)	0.490 (23.805)	0.098 (6.059)	0.055 (1.837)
Unobserved Heterogeneity							
Correlation 1	1	0.672 (27.686)	--	--	--	--	0.672 (27.686)
Correlation 2	1	--	0.771 (50.059)	0.771 (50.059)	0.771 (50.059)	--	--
Total number of parameters = 61, Log-likelihood: -43,622.58; AIC: 87,367.14; BIC:87,748.19							

Note: *No. of Parm = Number of parameters estimated for the corresponding variable. So, 6 means, the effect of that specific variable is estimated for all six crash types

Table 4.5 Random Parameter Multivariate NB (RPMNB) Model Estimation Results

Variables	No. of Parm*	Rear End	Angular	Sideswipe	All single vehicle	Other multiple vehicle	Non-motorized
		Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)
Constant	6	-1.069 (-9.246)	-1.763 (-18.966)	-2.663 (-21.251)	-0.612 (8.595)	-1.738 (-19.770)	-3.722 (-23.811)
Roadway Characteristics							
Proportion of arterial roads	5	0.206 (3.332)	0.070 (1.777)	0.085 (1.998)	-0.232 (-5.503)	--	0.248 (3.205)
Number of intersections	3	--	0.291 (9.732)	--	--	0.176 (5.906)	0.318 (7.114)
Variance of speed	5	0.031 (2.127)	0.040 (2.932)	0.075 (4.630)	0.021 (2.016)	0.034 (2.532)	--
Length of divided road	5	0.454 (1.757)	0.332 (1.942)	0.320 (1.707)	0.512 (2.688)	0.376 (1.725)	--
Signal intensity	2	--	--	-0.489 (-2.324)	-0.632 (-4.721)	--	--
Average outside shoulder width	4	-0.489 (-6.398)	-0.167 (-3.720)	-0.341 (-6.050)	--	-0.087 (-1.927)	--
Road length over 55mph	5	0.814 (5.138)	-0.608 (-4.131)	1.038 (6.416)	1.245 (12.224)	--	-0.366 (-1.752)
Standard deviation	1	--	0.681 (3.459)	--	--	--	--
Sidewalk width	2	0.135 (4.174)	--	--	--	--	-0.072 (-2.798)
Traffic Characteristic							
VMT	5	--	0.070 (6.053)	0.209 (16.874)	-0.111 (-4.974)	0.086 (7.450)	0.053 (3.111)
Truck VMT	2	0.202 (14.464)	--	--	0.325 (12.257)	--	--
Land-use attributes							
Urban area	6	0.168 (11.399)	0.124 (8.870)	0.136 (8.079)	0.063 (8.193)	0.094 (7.674)	0.160 (7.631)
Office area	6	0.212 (10.241)	0.243 (13.821)	0.244 (11.481)	0.087 (7.031)	0.177 (10.097)	0.168 (7.715)
Institutional area	4	0.062 (3.580)	0.074 (4.853)	--	--	0.044 (3.060)	0.111 (5.566)
Residential area	3	-0.074 (-5.909)	-0.031 (-3.196)	-0.101 (-8.061)	--	--	--
Built environment characteristic							
No. of restaurant	6	0.246 (6.002)	0.254 (8.792)	0.299 (10.372)	0.102 (4.703)	0.265 (11.693)	0.219 (8.116)
No of shopping center	2	0.041 (1.859)	0.021 (1.721)	--	--	--	--
Socio-demographic characteristics							

Population density	6	0.246 (10.682)	0.133 (12.254)	0.144 (10.490)	0.027 (3.243)	0.114 (10.808)	0.128 (10.153)
Non-motorist commuter	3	0.034 (1.883)	--	0.044 (2.152)	--	--	0.042 (1.752)
Proportion of household without vehicle	3	-0.674 (1-.748)	--	-1.143 (-3.077)	--	--	2.491 (6.084)
Over dispersion	6	0.522 (24.872)	0.179 (9.849)	0.291 (11.565)	0.491 (23.614)	0.098 (5.921)	0.033 (2.152)
Unobserved Heterogeneity							
Correlation 1	1	0.669 (27.067)	--	--	--	--	0.669 (27.067)
Correlation 2	1	--	0.772 (48.990)	0.772 (48.990)	0.772 (48.990)	--	--
Total number of parameters= 92, Log-likelihood: -43,597.82; AIC: 87,379.64; BIC:87,954.34							

Note: *No. of Parm = Number of parameters estimated for the corresponding variable. So, 6 means, the effect of that specific variable is estimated for all six crash types

Table 4.6 Predictive Performance Measure of Two Models

Dataset	Crash Type	MPB		MAD		MAPE		RMSE		Predictive BIC	
		RPMNB*	PMNB	RPMNB*	PMNB	RPMNB	PMNB	RPMNB	PMNB	RPMNB	PMNB
In-Sample Measures (3,815 TAZs)	Rear-end	3.340	2.787	9.395	8.884	2.676	2.584	53.823	35.848	87,954.34	87,748.19
	Angular	0.878	0.942	3.321	3.386	0.882	1.205	10.627	13.044		
	Sideswipe	0.661	0.654	2.555	2.553	0.764	0.753	10.612	10.852		
	All single vehicle	0.025	0.007	2.197	2.189	0.228	0.217	3.508	3.502		
	Other multiple vehicle	0.486	0.492	2.253	2.258	0.579	0.502	6.120	6.190		
	Non-Motorized	0.063	0.076	0.699	0.712	0.056	0.107	1.388	1.607		
	Total	5.454	4.956	20.421	19.983	5.185	5.367	56.339	40.325		
Hold-out sample Measures (932 TAZs)	Rear-end	5.546	4.691	10.927	10.102	2.583	2.932	71.879	56.098	21,868.31	21,66173
	Angular	1.402	1.449	3.623	3.669	0.723	0.774	13.666	14.955		
	Sideswipe	1.352	1.353	3.056	3.063	0.915	1.029	17.978	18.597		
	All single vehicle	0.098	0.080	2.138	2.119	0.200	0.219	3.452	3.415		
	Other multiple vehicle	0.659	0.682	2.575	2.603	0.777	0.282	9.351	9.860		
	Non-Motorized	0.136	0.158	0.748	0.768	0.124	0.069	1.552	1.896		
	Total	9.193	8.414	23.066	22.325	5.323	5.306	76.015	61.879		

Note: *RPMNB=Random parameter multivariate negative binomial model, PMNB= Panel mixed negative binomial model

*Model with underline gives better measure

CHAPTER 5: ECONOMETRIC APPROACH FOR MODELING CRASH COUNTS BY CRASH TYPE AND SEVERITY

The traditional modeling framework for crash frequency analysis is the univariate frequency model such as Poisson, Negative binomial or the Poisson-Lognormal model (see (Bhowmik et al., 2018; Lord and Mannering, 2010) for a detailed review of these studies). In these studies, for an observation unit, the modeling variable of interest is typically the total number of crashes. The approach of aggregating all crashes into a single dependent variable can result in aggregation bias and a loss of information available in the dataset. For instance, consider two zones with 5 observed crashes in the analysis period. For zone 1, the 5 crashes include 5 head-on crashes while for zone 2, the 5 crashes include 4 rear-end crashes and 1 vehicle pedestrian crash. While the crash distribution by crash type across the two zones is quite distinct, an approach focusing on total crashes will consider both zones as having identical dependent variables. The aggregation would make it quite cumbersome to accurately estimate the impact of independent variables on total crashes. For example, in zone 1, geometric design inadequacies might be the reason for head-on crashes while in zone 2, the presence of a significant number of signalized urban intersections might be the reason for rear-end and pedestrian crashes. A single *total crash* model will not be able to parse these distinctions accurately. Hence, it is not surprising that in recent years, safety researchers have focused on disaggregating the data by various attributes such as crash typology (such as head-on or rear-end), injury severity (such as crashes by no injury or crashes by severe injury) and crash location (such as intersection versus non-intersection).

The proposed disaggregation of the crash frequency variable increases the complexity of the modeling effort and presents many additional challenges. The number of dependent variables

of interest increase based on the attribute levels of interest. For analyzing these multiple dependent variables, multiple univariate models with frequency by attribute levels (such as crashes by crash type) will need to be estimated. While developing multiple univariate crash frequency models will account for the influence of independent variables, these models ignore that the multiple crash frequency variables for a traffic analysis zone (TAZ) are potentially correlated. For example, for zonal level crash frequency analysis, it is possible that several characteristics specific to the zone such as driver behavior, geometric design and build quality (possibly of higher or lower quality relative to the other zones) and traffic signal design objectives might influence different crash counts by crash type (such as head-on, rear-end). Thus, any modeling approach to analyze the multiple crash frequency variables need to explicitly account for the presence of these common factors that are most often unobserved. Ignoring for the presence of such unobserved heterogeneity in model development will result in inaccurate and biased model estimates (see Liu and Sharma, 2018; Mannering et al., 2016; Zeng et al., 2018 for an extensive discussion). The most common approach employed to address the potential unobserved heterogeneity in safety literature is the development of multivariate crash frequency models.

5.1 Earlier Research

A summary of earlier research efforts investigating crash frequencies by crash type and severity level are presented in Table 5.1 with information on the spatial unit (aggregation level), the region (covered area, for example state or city), crash unit (type of crash considered), number of dimensions examined (of the dependent variable), methodological framework employed, and different categories of exogenous variables considered in the analysis. The following observations can be made from Table 5.1. First, the most prevalent mechanism to analyze crash count by different levels are multivariate count regression approaches. Second, several spatial units are

considered both at macro and micro level for analyzing the crash counts by type and injury severity including segments and intersections (for micro level); and census block and traffic analysis zone (for macro level). Third, the methodological frameworks adopted in these studies include Negative binomial, Poisson regression, Multivariate Poisson-lognormal, Multivariate Negative Binomial, Multinomial Generalized Poisson and Integrated Nested Laplace Approximation. Fourth, with respect to exogenous variables, the overall findings from earlier research effort are consistent. The various factors identified that influence crash severities include - (1) roadway characteristics such as shoulder width, arterial road length; (2) land-use characteristics such as urban land use and land use mix; (3) built environment characteristics such as number of access points (number of restaurant, entertainment center); (4) traffic characteristics such as Average Annual Daily Traffic (AADT) and truck volume; (5) socio-demographic characteristics such as population density and people by different age group; and (6) weather variables such as precipitation rate. Fifth, the highest number of dependent variables considered in multivariate models is 8. Finally, none of the studies[6] examined the crash counts of different crash types and their corresponding severity outcomes in an integrated framework at the planning level.

5.2 Current Study

In multivariate count regression approaches described above, the impact of exogenous variables is quantified through the propensity component of count models. In accommodating the influence of unobserved effects, in general, these approaches partition the error components as a common term

[6] One study (Yasmin et al., 2016) investigated the crash severity proportions considering different crash types, while developing separate models for different crash types. However, the study did not model the crash frequencies by crash type in the joint modeling approach.

and an independent term across dependent variables (see (Mannering et al., 2016) for a detailed discussion of various methodologies). The approaches rely either on Maximum Simulated Likelihood (MSL) or Markov Chain Monte Carlo (MCMC) approach in the Bayesian realm for model estimation. MSL and MCMC methods provide substantial flexibility in accommodating for unobserved heterogeneity.

While several research efforts have developed multivariate crash frequency models for a small number of dimensions (such as 5); there is limited adoption of multivariate approaches for count variables in the presence of larger number of dependent variables (say greater than 15). For example, consider the development of crash frequency models by crash type (say N types) and severity level (say K levels). In the currently employed approaches, the number of crash propensity equations to be estimated will be N*K. While the estimation of $N * K$ univariate model systems is repetitive, it is still feasible. However, accommodating for unobserved heterogeneity with a large number of dependent variables is substantially challenging. The probability evaluation with high dimensional integrals is potentially affected by several challenges including - requirements of generating high dimensionality of random numbers, empirical identification issues due to relatively flat objective functions in larger dimensions and longer computational run times. Furthermore, the stability of the variance-covariance matrix is often sensitive to model specification and number of simulation draws.

The proposed research is geared toward addressing the dimensionality challenge in the traditional multivariate crash frequency models. In doing so, the proposed research builds on recent developments in crash frequency analysis along multiple directions. First, we draw on our recent work employing fractional split modeling approach for crash frequency analysis. In a fractional split approach, as opposed to modeling the count events, count proportions by different attributes

(such as injury severity, crash type or vehicle type) for a study unit are examined. Yasmin et al. (Bhowmik et al., 2018) employed a joint Negative Binomial-Ordered Logit Fractional Split (NB-OLFS) model using zonal level crash records to tie the total crash count and severity in a single joint system. The authors concluded that the proposed approach is more appealing relative to the traditional multivariate models for multiple reasons: 1) it is computationally less burdensome as it requires the estimation of only two equations irrespective of the number of crash severity levels; 2) the fractional split approach directly relates a single exogenous variable to count proportions of all attribute levels simultaneously. On the contrary, in the traditional multivariate models, the observed variables in different count propensity equations do not interact across different dimensions; and 3) the ordered fractional split framework recognize the inherent ordering for the severity levels which is ignored in the traditional multivariate models. Building on this fractional split approach, the proposed research develops a joint system for analysing crash frequency by crash type (N) and severity level (K) with ($N * K$)dependent variables per observation as follows: The NB count model is employed to incorporate the frequency by the crash type dimension and the fractional model is employed to analyze crash severity within each crash type dimension. Thus, instead of modeling $N * K$ dependent variables with $N * K$propensity equations (and integration of unobserved factors of the same order), we reduce the dimensionality to $N * 2$. At this stage, if the analyst is considering L_1 observed variables and L_2 unobserved parameters, the model estimation complexity has reduced to $N * 2 * (L_1 + L_2)$ from $N * K * (L_1 + L_2)$.

Second, we draw on another recent work that recasts the multivariate distributional problem (for multiple crash frequency dependent variables) as a repeated measure univariate problem (see (Bhowmik et al., 2019b) for detail). For example, crash frequency by crash type is represented as a repeated measure of crash frequency variable recognizing that each repetition

represents a different crash type instead of considering it as a multivariate distribution. The recasting process allows for the estimation of a parsimonious model system by allowing for an improved specification testing of variable impacts across different crash types (see Bhowmik et al., 2019a for detail). Using this consideration, the proposed model system enhances the efficiency of estimation through a single crash frequency model and a single crash proportion model, while also allowing for parameter effects to vary across different crash types through crash type specific deviation terms. Building on this study design, the $N * 2 * (L_1 + L_2)$ could potentially be reduced to $2 * (L_1 + L_2)$. Of course, we envision that the exact number of parameters to be estimated will lie somewhere in the range between $2 * (L_1 + L_2)$ and $N * 2 * (L_1 + L_2)$. The reduction in parameters especially for unobserved factors will contribute to substantial improvements in model efficiency and computational times.

Finally, in earlier fractional split modeling efforts, the severity variable is analyzed using the traditional ordered outcome structure. However, as illustrated in existing literature (see Eluru and Yasmin, 2015; Fountas and Anastasopoulos, 2017; Xin et al., 2017; Bhowmik et al., 2019b for detail), adopting a generalized ordered framework that relaxes the restrictive assumptions of the ordered outcome model (also referred to as parallel lines assumption) by allowing the threshold parameters to vary in response to observational attributes would be more representative.

In summary, the current study contributes to safety literature both methodologically and empirically by proposing a joint econometric approach for examining the count events as well as the severity outcome for different crash types. Methodologically, we employ a Joint Panel mixed Negative Binomial- Generalized Ordered Probit Fractional Spilt (PM-NB-GOPFS) model where the first component (NB) will accommodate for crash frequency by crash type and the second component (GOPFS) will study the fraction of severity outcome for different crash types. The

model is estimated using zonal level crash count, crash type and severity data for both motorized and non-motorized crashes. The crash data is extracted for the year 2016 from Central Florida region of the USA. The dimension of the dependent variables analysed is 24 [$(6 * 4)$ from 6 crash types (rear-end, angular, sideswipe, head-on, single vehicle and non-motorist crash) and 4 severity levels (severe (fatal and incapacitating as one category), non-incapacitating, possible injury and property damage). Empirically, the proposed approach allows for flexible consideration of crashes by crash types and severity levels within a single framework. Further, the proposed model results offer insights on important variables affecting crash frequency and severity for different crash types. Moreover, the macro-level model outcomes can be used to devise safety-conscious decision support tools to facilitate proactive consideration in assessing medium and long-term policy-based countermeasures.

The rest of the chapter is organized as follows: the next section presents the methodological framework adopted in the analysis while the section 5.4 provides a detailed description of the model findings. Finally, the predictive performance evaluation of the proposed framework is discussed in section 5.5 followed by the concluding remarks in the last section.

5.3 Econometric Framework

In this section, we provide details of the Panel mixed Negative Binomial - Generalized Ordered Probit Fractional Spilt (PM-NB-GOPFS) model employed in our study.

5.3.1 Count Model Structure

The focus of our study is to recast the multivariate NB count model as a panel mixed univariate NB modeling framework. For this purpose, we consider the six types of crashes as repeated measures (same TAZ is repeated 6 times) of crash frequency in a univariate NB formulation while

recognizing that each repetition represents a different crash type. The econometric framework of the proposed approach is presented in this section. Let's assume i $(i = 1,2,3, \dots, N;\ N = 3{,}815)$ be an index to represent observation unit (TAZs) and r $(r = 1,2, \dots, R;\ R = 6)$ be an index for different crash type and k $(k = 1,2,3, \dots, K; K = 4)$ be the index to represent injury severity categories at observation unit i. Then the probability equation of the NB formulation can be rewritten as follow:

$$P(c_{ir}|v_{ir}, \lambda') = \frac{\Gamma\left(c_{ir} + \frac{1}{\lambda'}\right)}{\Gamma(c_{ir}+1)\Gamma\left(\frac{1}{\lambda'}\right)}\left(\frac{1}{1+\lambda' v_{ir}}\right)^{\frac{1}{\lambda'}}\left(1 - \frac{1}{1+\lambda' v_{ir}}\right)^{c_{ir}} \quad (27)$$

where, c_{ir} be the index for crash counts occurring over a period of time in observation unit i and crash type r. $P(c_{ir})$ is the probability that unit i has c_{ir} number of crashes for crash type r. λ' is NB over dispersion parameter and v_{ir} is the expected number of crashes occurring in i over a given time period for crash type r. In equation 27, we can express v_{ir} as a function of explanatory variables using a log-link function as follows:

$$v_{ir} = E(c_{ir}|\boldsymbol{x}_{ir}) = exp((\tau + \Phi_i + \varrho_{ir} + \eta_{irk})x_{ir} + \varepsilon_{ir}) \quad (28)$$

where, x_{ir} is a vector of explanatory variables associated with observations i for crash type r. $\boldsymbol{\tau}$ is a vector of coefficients to be estimated. Φ_i is a vector of unobserved factors moderating the influence of attributes in x_{ir} on the crash count propensity for analysis unit i, ϱ_{ir} is a vector of unobserved effects specific to crash type r. This $\boldsymbol{\varrho}_{ir}$ will be same across crash types in our case and thus the unobserved heterogeneity across crash types will be captured. ε_{ir} is a gamma distributed error term with mean 1 and variance λ'. η_{irk} captures unobserved factors that simultaneously impact number of crashes by crash type and proportion of crashes by severity for different crash types for unit i.

5.3.2 Severity Model Structure

In the joint model framework, the modeling of crash proportions by severity levels across different crash types is undertaken using the Generalized Ordered Probit Fractional Split (GOPFS) model. In the ordered outcome framework, the actual injury severity proportions (y_{irk}) are assumed to be associated with an underlying continuous latent variable (y_{ir}^{*}). The latent propensity equation is typically specified as the following linear function:

$$y_{ir}^{*} = (\alpha_r + \gamma_{ir} + \delta_{ir} \pm \eta_{irk}) z_{ir} + \xi_{irk} \quad (29)$$

This latent propensity y_{ir}^{*} is mapped to the actual severity proportion categories y_{ik} by the ψ_r thresholds (ψ_{r0} =-∞ and ψr_k= ∞). z_{ir} is a vector of attributes that influences the propensity associated with crash severities. α_r is a corresponding vector of mean effects specific to r, and γ_{ir} is a vector of unobserved factors on severity proportion propensity for TAZ i and its associated zonal characteristics assumed to be a realization from standard normal distribution: $\boldsymbol{\rho} \sim N(0, \boldsymbol{\sigma}^2)$. δ_{ir} is a vector of unobserved effects specific to crash type r. This δ_{ir} will be same across severity proportions in any TAZ and thus the unobserved heterogeneity across the severity proportions will be captured. ξ_{irk} is an idiosyncratic random error term assumed to be identically and independently standard normal distributed across TAZ i. η_{irk} term generates the correlation between equations for total number of crashes and crash proportions by severity levels for different crash type.

The GOPFS model relaxes the constant threshold across observation to provide a flexible form of the OPFS model. The basic idea of the GOPFS is to represent the threshold parameters as a linear function of exogenous variables. Thus, the thresholds are expressed as:

$$\psi_{rk} = fn(s_{irk}) \quad (30)$$

where, s_{irk} is a set of exogenous variables (including a constant) associated with k th threshold. Further, to ensure the accepted ordering of observed crash severity proportion

$(-\infty < \psi_{r1} < \psi_{r2} < \ \ldots\ldots\ldots < \psi_{rK-1} < +\infty)$, we employ the following parametric form as employed by Eluru et al.(Eluru et al., 2008):

$$\psi_{rk} = \psi_{r,k-1} + exp((\beta_{rk} + \theta_{irk} + \varsigma_{ir} \pm \eta_{irk})s_{irk}) \quad (31)$$

where, β_{rk} is a vector of parameters to be estimated. θ_{irk} is another vector of unobserved factors moderating the influence of attributes in s_{irk} on the severity proportions for analysis unit i , crash type r and injury severity category k. ς_{ir} is a vector of unobserved effects specific to crash type r. This ς_{ir} will be same across the threshold parameters (upper severity categories) in any TAZ and thus the unobserved heterogeneity across the threshold parameters will be captured.

To estimate the model presented in equation 29, we assume that:

$$E(y_{irk}|Z_{irk}) = H_{irk}(\alpha_r, \psi_{rk}, \delta_{ir}, \theta_{irk}),\ 0 \le H_{irk} \le 1, \sum_{rk=1}^{rK} H_{irk} = 1 \quad (32)$$

where H_{irk} in our model takes the generalized ordered probit probability form for the severity category k specific to crash type r. Given these relationships across different parameters, the resulting probability for the GOPFS model takes the following form:

$$P_{irk} = G\ [(\psi_{rk} - \{(\alpha + \gamma_i + \delta_{irk} \pm \eta_{irk})z_{ir}\}\ ] \quad - \quad G\ [(\psi_{r,k-1} - \{(\alpha + \gamma_i + \delta_{irk} \pm \eta_{irk})z_{ir}\}] \quad (33)$$

where, $G(\cdot)$ is the standard normal cumulative distribution function (Eluru et al., 2013; Papke, 1996). The proposed model ensures that the proportion for each severity category is between 0 and 1 (including the limits). The $\pm$ sign in front of η_{irk} in equation 33 indicates that the correlation in unobserved individual factors between total crashes and crash proportions by severity levels for different crash types may be positive or negative.

5.3.3 Correlation Structure

In the current research effort, several unobserved factors are considered. At the observation level (TAZ), we consider influence of common unobserved factors across crash frequency (Φ_i) and crash severity (γ_{ir}, θ_{irk}). In addition to this, a number of correlation terms are tested including: 1) common unobserved factors simultaneously affecting crash counts of different crash types ($\boldsymbol{\varrho}_{ir}$); 2) common unobserved factors simultaneously affecting crash severity proportions of different crash types ($\delta_{ir}, \varsigma_{ir}$); and 3) common unobserved factors that simultaneously impact crash counts and severity proportions by different crash types (η_{irk}). A discussion of these correlation structures are presented below:

$$\mathbb{A} = \begin{array}{c} \\ \text{Crash Types (1,2,...J, J=6)} \\ \\ \text{Crash Severity (1..K, K=4)} \end{array} \begin{array}{cc} \text{Crash Types (1,2,...J, J=6)} & \text{Crash Severity (1..K, K=4)} \\ \left[\begin{array}{cc} \mathbb{A}_1 & \mathbb{A}_3 \\ \varrho_{ir} & \eta_{irk} \end{array}\right. & \left.\begin{array}{c} \\ \\ \end{array}\right] \\ \left[\begin{array}{cc} \mathbb{A}_3 & \mathbb{A}_2 \\ \eta_{irk} & \delta_{ir} \end{array}\right. & \left.\begin{array}{c} \\ \\ \end{array}\right] \end{array} \tag{34}$$

Equation 34 provides the overall structure of the correlation matrix. The order of the correlation matrix is provided by the total number of crash type and crash severity levels (N+K). To better elaborate on the structure, we discuss the three main components of the matrix. The top left part represents the correlation matrix for the crash type only:

$$\mathbb{A}_1 = \begin{array}{c} \\ \text{Crash Types (1,2,...J, J=6)} \end{array} \begin{array}{c} \text{Crash Types (1,2,...J, J=6)} \\ \begin{bmatrix} & \varrho_1 & \varrho_2 & \cdots & \varrho_J \\ \varrho_1 & 1 & я_{12} & \ldots\ldots & я_{1J} \\ \varrho_2 & я_{12} & 1 & \ldots\ldots & я_{2J} \\ \cdots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \varrho_J & я_{1J} & я_{2J} & \ldots\ldots & 1 \end{bmatrix} \end{array} \tag{35}$$

As described in Equation 35, these terms represent the correlation between crash types. For example, the correlation parameter $я_{12}$ in equation 35 captures the common unobserved factors affecting the crash counts of crash type 1 and crash type 2 (which is rear-end and angular for the current study context) simultaneously while $я_{2J}$ represents the potential correlation between crash type 2 and crash type J.

Equation 36 represents the lower right part of the correlation matrix in equation 34 that accommodates for the common unobserved heterogeneity across the crash severity proportions.

$$\mathbb{A}_2 = \begin{array}{c} \\ \text{Crash Severity (1..K, K=4)} \end{array} \begin{array}{c} \text{Crash Severity (1..K, K=4)} \\ \begin{bmatrix} & \delta_1 & \ldots & \delta_K \\ \delta_1 & 1 & \ldots\ldots & ц_{1K} \\ \ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \delta_K & ц_{1K} & \ldots\ldots & 1 \end{bmatrix} \end{array} \tag{36}$$

To elaborate, the correlation parameter $ц_{1K"}$ captures the presence of common unobserved factors between the crash proportion of severity category 1 and K (which is no and severe injury for the current analysis).

Equation 37, representing the bottom left or top right parts in the correlation matrix from equation 34 captures the potential correlation between a crash type and its' corresponding severity proportion.

$$\mathbb{A}_3 = \begin{array}{c} \\ \textit{Crash Types and severities} \end{array} \begin{array}{c} \textit{Crash Types and severities} \\ \begin{bmatrix} & \varrho_1 & \varrho_2 & \cdots & \varrho_J \\ \delta_1 & \acute{r}_{11} & \acute{r}_{21} & \cdots\cdots & \acute{r}_{1J} \\ \cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \delta_K & \acute{r}_{1K} & \acute{r}_{2K} & \cdots\cdots & \acute{r}_{JK} \end{bmatrix} \end{array} \tag{37}$$

Specifically, the correlation parameter $\acute{r}_{11}$ captures the presence of potential correlation between the crash counts of crash type 1 and crash proportion of severity category 1 It is important to note that the correlation structure presented is applicable to each independent variable examined in the model (including constants). This indicates that potentially (N+K)* (N+K)/2 elements can be estimated for each variable. While theoretically this is possible, it is important to conduct the estimation judiciously to avoid identification issues.

It is also useful to note that the correlation parameters in the $\mathbb{A}_3$ matrix can be positive or negative. For instance, let us consider the correlation parameters in the last row from equation 37. Here, a positive sign implies that TAZs with higher number of crashes are intrinsically more likely to incur higher proportions for severe crashes specific to any crash types. On the other hand, negative sign implies that for any types of crash, TAZs with higher number of crashes intrinsically incur lower proportions for severe crashes. To determine the appropriate sign one can empirically test the models with both $'+'$ and $'-'$ signs independently. The model structure that offers the superior data fit is considered as the final model.

5.3.4 Joint (NB-GOPFS) Model Estimation

In estimating the model, it is necessary to specify the structure for the unobserved vectors $\Phi, \varrho, \gamma \; and \; \delta$ represented by Ω. In this study, it is assumed that these elements are drawn from independent normal distribution: $\Omega \sim N(0, (\pi^2, \sigma^2 \; \nu^2))$. Thus, conditional on Ω, the likelihood function for the joint probability can be expressed as:

$$L_i = \int_{\Omega} \prod_{r=1}^{R} \left[\left(P(c_{ir}) \right) \times \prod_{k=1}^{K} (P_{irk})^{\varpi_{ir} d_{irk}} \right] d\Omega \tag{38}$$

where, ϖ_{ir} is a dummy with $\varpi_{ir} = 1$ if TAZ i has at least one crash specific to crash type r over the study period and 0 otherwise. d_{irk} is the proportion of crashes in severity category k for each crash types. Finally, the log-likelihood function is:

$$LL = \sum_{i} Ln(L_i) \tag{39}$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 39. The parameters to be estimated in the model are: $\Phi, \varrho, \gamma \; \delta$, α, τ, β, ψ, π, σ and ν. To estimate the proposed model, we apply Quasi-Monte Carlo simulation techniques based on the scrambled Halton sequence to approximate this integral in the likelihood function and maximize the logarithm of the resulting simulated likelihood function across individuals (see (Bhat, 2001; Eluru et al., 2008) for examples of Quasi-Monte Carlo approaches in literature).The model estimation routine is coded in GAUSS Matrix Programming software (Aptech, 2015)

5.4 Empirical Analysis

5.4.1 Model Specification and Overall Measure of Fit

The number of TAZs in the study area is 4,747. Among these zones, 3,815 TAZs are randomly selected for model estimation and the records from other 932 TAZs are set aside for validation

purposes. Thus, the estimation sample has 22,890 (3,815*6) records and the validation sample has 5,592 (932*6) data records. The empirical analysis involved a series of model estimations. First, we estimated separate independent models (NB and GOPFS models) to establish a benchmark for comparison. Second, we proposed a parsimonious model structure using the same independent model system (NB and GOPFS) while restricting the parameters across different crash types considered. To elaborate, observing the model specifications in the independent models (NB and GOPFS), we identify potential parameters that can be restricted to be the same across various crash types and test that restriction (both NB and GOPFS dimension) in our proposed model system (see (Bhowmik et al., 2019b) for more details). Third, within our proposed system, we consider the unobserved heterogeneity in the joint model estimation. In summary, we estimated three different models in the current research effort including: 1) Independent NB-GOPFS model; 2) Panel NB-GOPFS model without unobserved component parameters and 3) Joint Panel NB-GOPFS model with unobserved heterogeneity. The log-likelihood values at convergence for these estimated models are: a) Independent NB-GOPFS (with 131 parameters) is -51,904.45 (b) Panel NB-GOPFS model without unobserved component (with 100 parameters) is -51,912.92.11 and (c) Joint Panel NB-GOPFS model with unobserved heterogeneity (with 105 parameters) is -50,945.82. We also compute the Bayesian Information Criterion (BIC) (lower is better) for these three frameworks to determine the best model. The corresponding BIC values for the three models are as follows: 105,123.93 (independent NB-GOPFS model), 104,650.50 (panel NB-GOPFS model) and 102,757.53 (joint panel NB-GOPFS model). Based on the BIC values, two observations can be made. First, the proposed framework that accounts for penalty for additional parameters provide improved data fit compared to the traditional model (independent NB-GOPFS model). This supports our hypothesis that the impact of some variables may not differ across the crash types

and through the proposed structure (recasting), we can have a parsimonious model system with improved parameter efficiency. Second, models considering unobserved heterogeneity outperforms the respective independent models which underscores the importance of accommodating for such unobserved effects in examining crash frequencies and severities at the planning level for different crash types.

5.4.2 Model Estimation Results

This section offers a detailed discussion of exogenous variable effects on the crash count as well as the severity outcome for different crash types. In discussing the model results, for the sake of brevity, we will restrict ourselves to the discussion of the joint panel model (NB-GOPFS) only (see table 5.4 and 5.5 for the results of independent NB-GOPFS model). For the ease of presentation, we first present an intuitive discussion of crash count component (Table 5.2) followed by the discussion of the severity component (Table 5.3) for different crash types.

5.4.2.1 Count Component

The coefficients in Table 5.2 represent the effect of exogenous variables on the frequency component of each crash type. The reader would note that, the variables in the crash count component of Table 5.2 with positive (negative) sign indicates that an increase in the variable is likely to result in more (less) crashes. In the subsequent sections, we provide a discussion of model results for different crash types by variable groups. The reader would note that Table 5.2 identifies the number of parameters estimated for each variable from a possible set of six (one effect for each crash type).

5.4.2.1.1Roadway Characteristics

The results regarding the impact of proportion of arterial roads reveal that a TAZ with higher proportion of arterial road is more likely to experience increased incidence of rear-end, angular and non-motorized crashes while the number of single vehicle crashes reduces. Single vehicle crashes (rollover and off-road) usually occur on high speed roads. On arterial roads, there is likely to be higher traffic interactions reducing operating speed and thus contributing to fewer single vehicle crashes. At the same time, the increased traffic interactions result in higher number of rear-end and angular crashes. It is also important to note that the influence of arterial roads is not different for rear-end, angular and non-motorized crashes i.e. a single parameter is adequate to accommodate for the impact of the variable. Traditional approaches in frequency modeling would have estimated three separate parameters while in our model, we estimate a single parameter. This is an example of how the proposed framework allows us to obtain a parsimonious specification (see (Bhowmik et al., 2019b) for similar results). Consistent with earlier research, the current analysis also found that the intersection variable is positively associated with angular and non-motorized crashes (Reynolds et al., 2009; Xuesong et al., 2006). Interestingly, the number of intersections variable has a positive coefficient for head-on crashes. While the result might seem counter-intuitive, a possible reason could be that vehicles turning left at an intersection stop at the outside lane that is closest to the oncoming traffic and as a consequence, the possibility of getting hit by the opposing traffic is likely to increase (see (Hosseinpour et al., 2014) for similar effect). The variable corresponding to signal intensity offers interesting insights. While an increase in the variable is positively associated with rear-end and non-motorized crashes, a negative relation is observed for sideswipe and single vehicle crashes. The trend is intuitive as the density of traffic intersections increases the potential conflicts between vehicles to vehicles and vehicles to non-

motorists. At the same time, these conflicts result in lower operating speed thus reducing single vehicle crashes.

The parameter associated with proportion of road over or equal to 55 mph speed limit exhibits contrasting impact on crash occurrence across crash types. The estimated results show that TAZs having higher percentage of roads over 55mph speed limit results in increased incidence of rear-end, sideswipe and single vehicle crashes while the likelihood of angular, head-on and non-motorized crash reduces. Within the positive effects, the parameter for single vehicle crashes has a higher magnitude (Yu and Abdel-Aty, 2013). Moreover, we found that the impact of the proportion of road over 55mph has significant variability on angular crashes (indicated by the standard deviation parameter) which implies that the overall impact is most likely to be negative (96%). Further, variance of speed is also found to be significant in rear-end, angular and sideswipe crash count component with a positive impact. In terms of proportion of road with separate median, the variable is found to have the same positive effect on rear-end, angular and sideswipe crashes whereas a negative coefficient is observed for head-on crashes. Roads with separated median, such as with guardrail, restricts a vehicle from entering the opposing direction. On the other hand, vehicles hitting the guardrail have a higher likelihood of colliding with the vehicles in the same direction. Hence, the result is expected. As found in previous studies (Bhowmik et al., 2018; Geedipally et al., 2010), average outside shoulder width reveals a negative association with all motorized crash types. Outside shoulder width in a road reflects the extra margin of safety for vehicular maneuvers and thus reduce the potential of all kinds of motorized crashes. With respect to sidewalk width, a number of earlier research concluded that increased sidewalk width is associated with higher pedestrian activity and as a result, they are more exposed to crashes. In our current study, we found an opposing (negative) effect of average sidewalk width for non-motorized

crashes. However, there is a reasonable explanation for the effect identified. First, the reader would note that we consider the non-motorist activity separately in the model framework (will be discussed in the following sections) and second, increased sidewalk width will provide additional safety to the non-motorist from colliding with a motorized vehicle.

5.4.2.1.2 Traffic Characteristics

The parameters associated with traffic characteristics highlight intuitive trends. Positive coefficient of VMT clearly underscores the higher propensity of angular, sideswipe, head-on and non-motorized crashes with increased VMT. VMT variable serves as a surrogate for exposure for traffic volume and therefore, with higher exposure, the likelihood of getting involved in a crash increases. On the other hand, zones with increased exposure to truck volume are likely to have a higher risk of getting involved in rear-end and single vehicle crashes, consistent with earlier research findings (Geedipally et al., 2010).

5.4.2.1.3 Land-use Attributes

With respect to land-use attributes, several factors exert significant impact on crash count components across crash types. The coefficient corresponding to urban area indicates that zones with higher urbanized area are likely to have increased crash risk for five of the six crash types (except single vehicle crashes). Similarly, office area in a zone is also found to be positively associated with rear-end, sideswipe and non-motorized crashes. These two variables basically reflect presence of higher vehicular and non-motorist interactions and in turn, higher exposure for both road user groups. Further, the result in Table 5.2 reveals a reduced propensity for sideswipe and single vehicle crashes with higher residential area.

5.4.2.1.4 Built Environment Attributes

In terms of built environment attributes, several variables have been explored out of which only number of restaurants and shopping centers are found to be related with zonal level crash risks. As is evident from Table 5.2, we can observe that both number of restaurants and shopping centers have positive influence on rear-end and sideswipe crashes, perhaps indicating a higher density of traffic volume for these areas. With respect to non-motorized crashes, number of restaurants is found to be a significant determinant with a positive impact (see (Yasmin et al., 2018a) for similar result).

5.4.2.1.5 Socio-demographic Characteristics

For socio-demographic attributes, we consider the number of non-motorists (walk/bike) and transit commuters in a zone serving as additional exposure measures for the crash risk model. The estimated result shows that higher number of pedestrians, bike and transit commuters, intuitively, increases the crash risk for rear-end and non-motorized crashes. Moreover, the coefficient specific to non-motorist commuters indicates that the variable is positively associated with angular and sideswipe crashes.

5.4.2.2 Severity Component

The coefficients in Table 5.3 represent the effect of exogenous variables on the injury severity proportion across different crash types. In the propensity, a positive (negative) coefficient corresponds to increased (decreased) proportion for severe injury categories specific to each crash type. When the threshold parameter is positive (negative), the result implies that the threshold is bound to increase (decrease). The estimation results are discussed by variable groups in the

following sections. The reader would note that Table 5.3 identifies the number of parameters estimated for each variable from a possible set of six (one effect for each crash type).

5.4.2.2.1 Roadway Characteristics

The variable specific to arterial road indicates that the likelihood of more severe crashes (proportions) increases with increasing share (length) of arterial road in a zone, particularly for rear-end, angular and single vehicle crashes. Further, we found an effect of arterial road on threshold value for rear-end crashes which provide a sense of how the probability of injury in specific injury categories is affected relative to the case of fixed thresholds. The negative coefficient of the variable on the threshold value highlights the higher proportions of serious injury (non-incapacitating or severe) crashes for rear-end crashes with increased length of arterial roads. Moreover, it can be seen from Table 5.3 that crashes on local road tends to be less severe for head-on and non-motorized crash types. The reduced likelihood of severe crashes for these two crash types perhaps can be attributed to reduced driving speed on local roads.

With increased number of intersections in a zone, the possibility of being involved in a severe crash decreases, particularly for head-on and single vehicle crashes. Similarly, we find that higher number of traffic signals in a zone reduce the possibility of higher injury risks for angular, sideswipe and head-on crashes. The results associated with both of these variables (intersection and signal) is potentially an indication that denser and signalized zones have a lower vehicle operating speed reducing crash consequences. Similar to the crash count components, the impacts of intersection and traffic signal do not differ across crash types; thus, we only estimate two parameters across the entire 4 dimensions (4 crash types) in the fractional split component.

Wider shoulder in a road provides additional safety margin for vehicular maneuverability and as expected, variables associated with it are found to have a negative influence on crash

severity outcome. While an increase in average insider shoulder width decreases the possibility of severe crashes for head-on crash, the likelihood of higher injury risk for rear-end crashes reduces with wider outside shoulder width. In terms of roadway attributes, one of the most important variables is speed and consistent with previous research, we also find speed to be an important contributing factor for severe crashes for different crash types. Specifically, zones with higher proportion of road over 55mph speed limit are more likely to experience higher proportion of severe crashes for five of the six crash types (except sideswipe crashes). Further the negative sign of threshold demarcating the non-incapacitating and severe injury proportion indicates higher likelihood of severe crash proportion for rear-end and non-motorized crashes with increased share of high speed (>55mph) road in a zone. Finally, the parameter associated with proportion of road with poor pavement condition reflects the higher injury risk propensity for sideswipe crashes.

5.4.2.2.2 Traffic Characteristics

Traffic congestion and truck VMT are found to have significant impact on crash proportions by severity levels for different crash types. As is evident from Table 5.3, we can observe that roads are typically safer in a congested traffic environment. In particular, the likelihood of severe crash proportion for rear-end and angular crashes are lower in a congested traffic environment (>85^{th} percentile traffic) compared to the uncongested condition (<=85^{th} percentile traffic).

Further, the impact of the variable on the threshold value for angular crashes implies a lower propensity of severe crash proportions in a gridlock situation. Moreover, the estimated result reveals a positive association between the truck VMT and the crash severity proportion, specifically for sideswipe and head-on crashes.

5.4.2.2.3 Land-use Attributes

With respect to land use attributes, urban area in a zone contributes negatively to injury severity propensity for sideswipe, head-on and single vehicle crashes, presumably because of the slower traffic on roadways in an urbanized environment. Further, the estimated results show that crash severity proportions are negatively associated with higher land use mix in a zone, particularly for rear-end and angular crashes.

5.4.2.2.4 Built Environment Attributes

In terms of built environment attributes, several factors are considered including number of commercial, recreation, restaurants and shopping centers. Interestingly, all of these reveal negative associations with the crash severity proportions across different crash types, perhaps indicating that with higher traffic density vehicle operating speed is likely to be lower and thus crash consequences are possibly less severe. For instance, consistent with previous findings (Yasmin et al., 2018a), number of commercial centers reduce the higher injury risk propensity for non-motorized crashes. Similarly, in the presence of higher number of recreational centers in a zone, a lower proportion of severe crash outcomes for single vehicle crashes is observed. Further, the GOPFS model results reveals that higher number of restaurants are associated with lower likelihood of severe crash proportions for single vehicle crashes, as indicated by the negative coefficient. The positive coefficient of the variable on the threshold value further reflects the lower probability of severe crash proportions. Finally, the variable corresponding to shopping centers results in lower likelihood of severity outcome, particularly for angular, sideswipe and head-on crashes (same impact). We also found a positive effect of the variable on the threshold which

further implies the lower possibility of higher injury risk for sideswipe crashes with increased number of shopping centers in a zone.

5.4.2.2.5 Socio-demographic Characteristics

The results for the effect of socio-demographic characteristics indicate that non-motorists are less prone to high injury risk with increased number of commuters in a zone (see (Yasmin et al., 2018a) for similar results). The likelihood of being involved in a severe crash is higher for increasing share of motor vehicle commuters, particularly for angular crashes. Previous studies (Pai and Saleh, 2008) also confirm the findings. Further, as found in previous studies (Quddus, 2008), the estimated results suggests that zones with more older people are associated with fewer severe crash proportion for non-motorized crashes. The coefficient specific to proportion of households without vehicle indicates a positive influence on severity outcome for non-motorized crashes indicating a higher propensity of more severe crash proportion for non-motorized crashes (for similar results, see (Quddus, 2008)).

5.4.2.3 Unobserved Effects

The final set of variables in both Tables (4 and 5) correspond to the correlation matrix (unobserved heterogeneity) in the joint model. As discussed earlier, in the current research effort, a number of correlation effects are tested including: 1) common unobserved factors affecting crash counts of different crash types simultaneously; 2) common unobserved factors affecting crash severity proportions of different crash types simultaneously and 3) common unobserved factors that simultaneously impact crash counts and severity proportions by different crash types. Within the crash count component, we found two common unobserved components including (1) common

unobserved factors affecting rear-end and non-motorized crashes and (2) common unobserved factors affecting other multi vehicular crashes (angular, sideswipe and head-on).

On the other hand, with respect to common factors between two components (count and proportions), the correlation could be either positive or negative as shown in Equation 33 of methodology section. In fact, the positive or negative sign can change by unobserved factor. In our analysis, we found the negative sign offers better fit for common correlation between total crash counts and threshold between proportion of no and possible injuries for non-motorized crashes. This indicates that a zone with higher number of non-motorized crashes are more likely to incur lower proportions of no injury crashes. On the other hand, a positive common correlation is found between the total number of head-on crashes and the corresponding threshold between proportion of non-incapacitating and severe crashes which implies that zones with higher number of head-on crashes intrinsically are more likely to incur higher proportions for serious crashes. Overall, the results clearly support our hypothesis that common unobserved factors influence the two components (crash counts and severity proportion).

5.5 Predictive Performance Evaluation

In order to demonstrate the applicability of the proposed joint (count and severity by crash type) model, a prediction exercise was undertaken using the final model parameter estimates. In doing so, we employ mean absolute deviation (MAD) and mean absolute percentage error (MAPE) which quantifies the error associated with model prediction and the measure is computed on two datasets including: 1) model estimation sample with 3,815 TAZs and 2) hold out sample (validation sample) with 932 TAZs to ensure that the statistical results obtained above are not a manifestation of over fitting to data.

One of the major advantage of the proposed framework is that in a single econometric framework, we can predict a number of dimensions including total crash counts, total crash counts by crash types, crash proportions for each severity level, crash counts for each severity level and finally, proportions and counts of crashes for each crash type by severity. In evaluating the predictive performance, we compute the errors (MAD and MAPE) across all the aforementioned dimensions. Specifically, we compute MAD at a disaggregate level by generating measures at the study unit level (TAZ) and compute the average measures across all units (total crash, crash type and severity). Other than total crash counts and crash count by crash type, we generate crash counts by severity levels for different crash types using the following equation:

$$E(\boldsymbol{P}_{irk}) = \mu_{ir} * \Lambda(y_{irk} = k) \quad (40)$$

where, μ_{ir} is the expected number of crashes for crash type r in TAZ i; $\Lambda(y_{irk} = k)$ is the predicted proportion of severity corresponding to crash type r and TAZ i; and $E(\boldsymbol{P}_{irk})$ is the expected number of crashes by injury severity k for crash type r in TAZ i. Finally, we compute MAD as:

$$\text{MAD} = \ mean\ |\hat{y}_i - y_i| \quad (41)$$

where, $\hat{y}_i$ and y_i are the predicted and observed, number of crashes occurring over a period of time in a TAZ i (corresponds to different dimension: total crash, crash type, severity etc). Figure 5.1 and 5.2 presents the value of MAD for estimation and validation sample, respectively.

On the other hand, we employ MAPE measures at an aggregate level where we estimate the number and proportion of crashes for corresponding dimension and predict the TAZ shares for different count and proportion alternatives and compared it with the observed shares. For example, let us consider the crash counts by crash type where we predict the number of crashes for each

crash type at an individual level (observation) and then we estimate how many TAZs have 0,1….250 crashes. Finally, we compute the MAPE as:

$$\text{MAPE} = \frac{1}{n}\sum_{n=1}^{N}\left|\frac{\hat{y}_n - y_n}{y_n}\right| \quad (42)$$

where, $\hat{y}_n$ and y_n are the predicted and observed, number of TAZs (corresponds to different dimension) for different count alternative n. Figure 5.3 and 5.4 presents the value of MAPE for estimation and validation sample, respectively.

In terms of MAD, we found that both datasets (from figure 5.1 and 5.2) offer similar predictive performance which highlights the applicability of the proposed joint framework by eliminating the overfitting issue. Further, out of all crash alternatives, the prediction accuracy is quite poor for no injury crashes followed by rear-end crashes relative to other crash types, crash severities and total crash counts. With respect to MAPE measures, the following observations can be made from the values presented in Figures 5.3 and 5.4. First, the predictive performance of the two datasets (estimation and validation sample) are quite similar. Second, in terms of the total crash counts, the predicted share of TAZs for different count alternatives are reasonably close to the observed share for both dataset with an error of 0.9% (both dataset) respectively. The reader would note we converted the numbers in the figures to percentage for discussion. Third, with respect to different severity levels, the model performs better for the lower categories (up to possible injury) while a slightly higher error rate (about 3%) is observed in the upper classes (category 3 and 4). Fourth, the MAPE values corresponds to crash types offer interesting insights. While we observe a lower accuracy for rear-end crashes in both datasets (12.4 % and 11.4 % respectively), the model performs adequately for other crash types with a maximum of 6.4% error rate for angular crash in the estimation sample. Finally, within each crash type, the MAPE values

for each severity fractions are quite reasonable without any significant trend highlighting the appropriateness of the proposed model.

In summary, the prediction results clearly indicate that the joint model for crash counts and severity proportions by crash type performs adequately for both datasets (in-sample and validation sample) under consideration.

5.6 Summary

In our current research effort, we employed a Panel mixed Negative Binomial- Generalized Ordered Probit Fractional Spilt (PM-NB-GOPFS) model where the first component (NB) accommodated for crash frequency by crash type and the later component (GOPFS) studied the fraction of severity outcome for different crash types. The empirical analysis was conducted using the zonal level crash count data for the year 2016 from Central Florida while considering a comprehensive set of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics. The empirical analysis involved a series of model estimations including: 1) Independent NB-GOPFS model; 2) Panel NB-GOPFS model without unobserved component parameters; and 3) Joint Panel NB-GOPFS model with unobserved heterogeneity. The comparison exercise, based on the Bayesian Information Criterion (BIC)value highlighted the superiority of the proposed framework that accounts for penalty for additional parameters (model 2 and 3) and within the proposed approach, the model considering unobserved heterogeneity (model 3) outperformed its' counterpart (model 2).

The analysis was further augmented by undertaking a prediction exercise using the final model parameter estimates. One of the major advantage of the proposed framework is that in a single econometric framework, we can predict several dimensions including total crash counts, total crash counts by crash types, crash proportions for each severity level, crash counts for each

severity level and finally, proportions and counts of crashes for each crash type by severity. In evaluating the predictive performance, we compute the errors (MAD and MAPE) across all the aforementioned dimensions. Specifically, we compute MAD at a disaggregate level by generating measures at the study unit level (TAZ). On the other hand, MAPE measures are generated at an aggregate level where we estimate the number and proportion of crashes for corresponding dimension (crash types, severities) and predict the TAZ shares for different count and proportion alternatives and compared it with the observed shares. The prediction results clearly indicated that the joint model for crash counts and severity proportions by crash type performed adequately (for both in-sample and validation samples) under consideration.

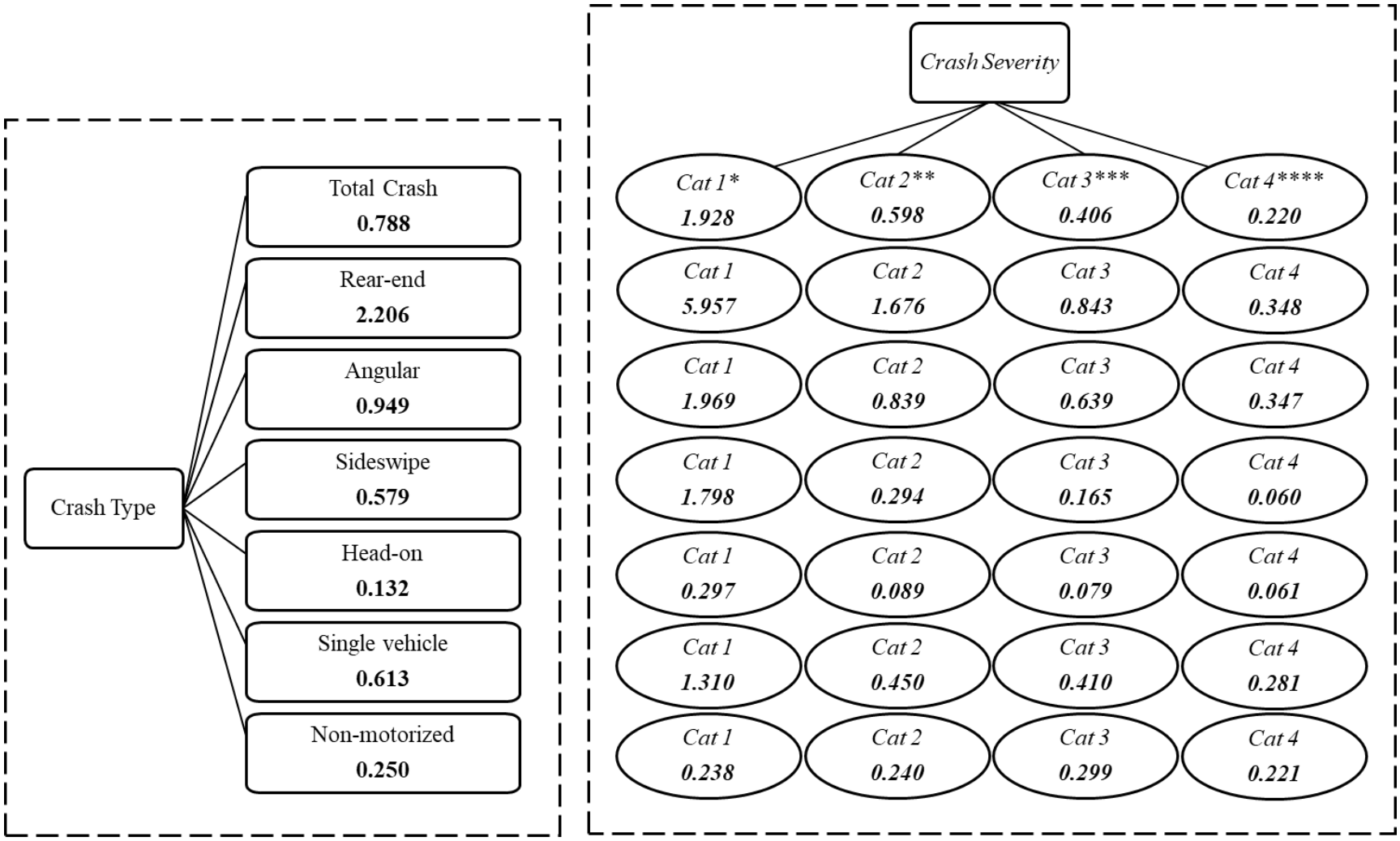

Figure 5.1 MAD Tree for Estimation Sample (3,815 TAZs)

*Cat 1 = proportion of no injury; **Cat 2= proportion of possible injury; ***Cat 3= proportion of non-incapacitating injury, ****Cat 4= proportion of severe injury

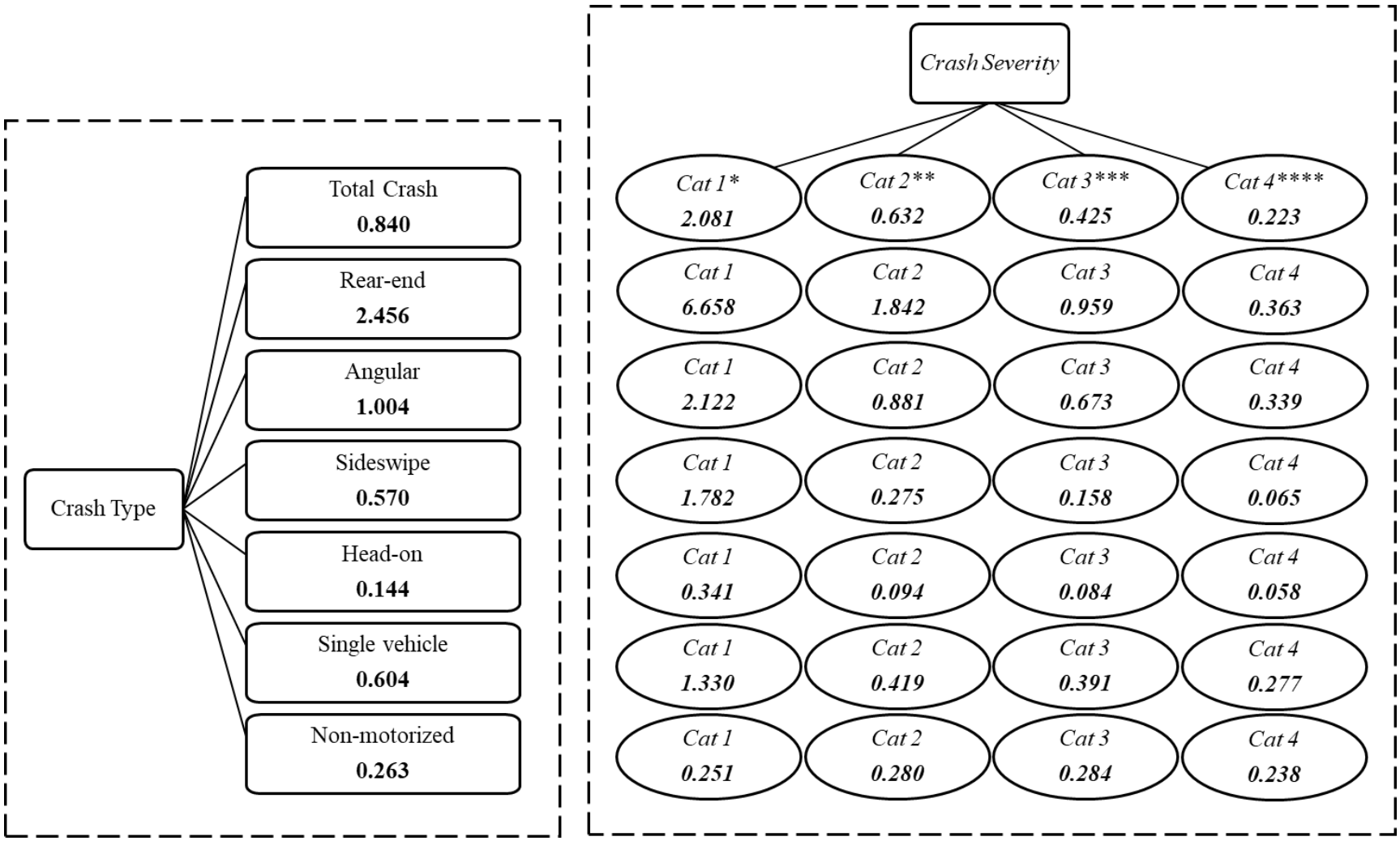

Figure 5.2 MAD Tree for Validation Sample (932 TAZs)

*Cat 1 = proportion of no injury; **Cat 2= proportion of possible injury; ***Cat 3= proportion of non-incapacitating injury, ****Cat 4= proportion of severe injury

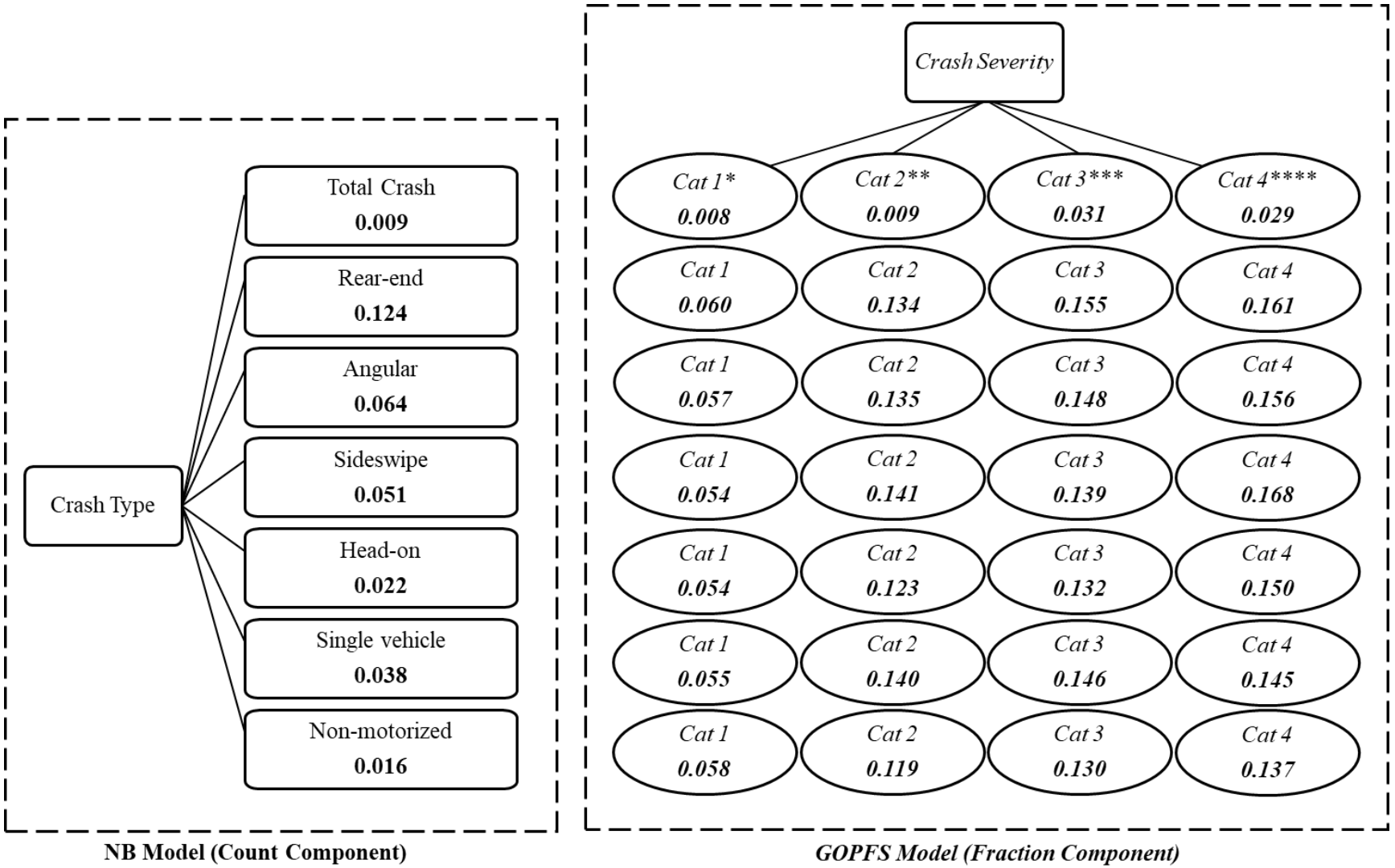

Figure 5.3 MAPE Tree for Estimation Sample (3,815 TAZs)

*Cat 1 = proportion of no injury; **Cat 2= proportion of possible injury; ***Cat 3= proportion of non-incapacitating injury, ****Cat 4= proportion of severe injury

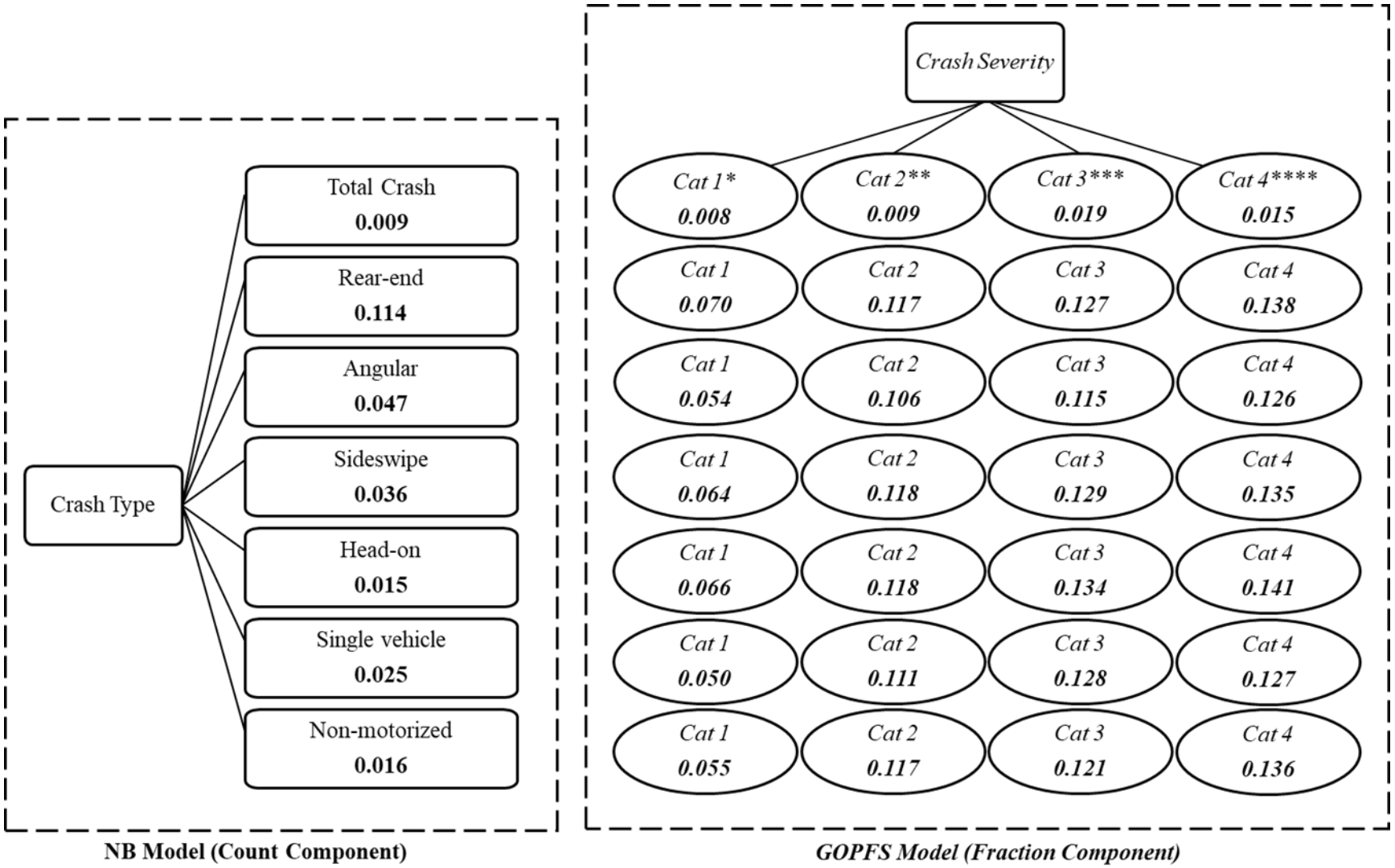

Figure 5.4 MAPE Tree for Validation Sample (932 TAZs)

*Cat 1 = proportion of no injury; **Cat 2= proportion of possible injury; ***Cat 3= proportion of non-incapacitating injury, ****Cat 4= proportion of severe injury

Table 5.1 Summary of Existing Aggregate Level Multivariate Crash Type and Severity Studies

Studies	Spatial Unit	Region	Crash unit	Number of Levels Explored	Methodological Approach	Independent Variables Considered					
						Roadway Infrastructure	*Land-use*	*Built Environment*	*Traffic Characteristics*	*Socio-demographic*	*Weather*
Crash Type Studies											
(Ye et al., 2009)	Intersections (micro)	County (Georgia)	Any Crash	7 (angle, head-on, rear-end, sideswipe: same and opposite direction, pedestrian)	Multivariate Poisson regression model	√	--	--	√	--	--
(El-Basyouny et al., 2014)	Citywide (Macro)	City (Edmonton)	Motorized crash	7 (FTC*, FOTS**, SSV***, left turn, ILC****, parked vehicle and off-road)	Multivariate Poisson-lognormal model	--	--	--	--	--	√
(Mothafer et al., 2016)	Highway segments (Micro)	State (Washington)	Motorized crash	4 (rear-end, sideswipe, fixed object and others)	Multivariate Poisson gamma mixture count model	√	--	--	√	--	--
(Jonathan et al., 2016)	Road segments (Micro)	County (Pennsylvania)	Any Crash	4 (same direction, opposite direction, angular, fixed object)	Multivariate Poisson-lognormal spatial model	√	--	--	√	--	--
(Serhiyenko et al., 2016)	Highway segments (Micro)	State (Connecticut)	Motorized crash	3 (same direction, opposite direction, single vehicle crash)	Multivariate Poisson-lognormal model	√	√	--	√	--	--
(Cheng et al., 2017)	Intersections (micro)	City (California)	Motorized crash	6 (rear-end, head-on, sideswipe, broad side, hit object crash, others)	Multivariate Poisson-lognormal model	√	--	--	√	--	--
(Wang et al., 2017)	Road segments, intersections (Micro)	State (Minnesota, Washington)	Any Crash	4 (same direction, intersecting direction, opposite direction, single vehicle crashes)	Multivariate Poisson-lognormal model	√	--	--	√	--	--
(Bhowmik et al., 2018)	STAZ (Macro)	State (Florida)	Motorized crash	8 (rear-end, angular, sideswipe, head-on, single vehicle, off-	Multivariate negative binomial model,	√	√	√	√	--	--

				road, rollover and others)	multinomial fractional split model						
(Alarifi et al., 2018)	Road segments, intersections (Micro)	County (Florida)	Any Crash	6 (same direction, angular, opposite direction, non-motorized, single vehicle and others)	Bayesian multivariate hierarchical spatial joint model	√	--	--	√	--	--
Crash Severity Studies											
(Narayanamoorthy et al., 2013)	Census tract (Macro)	Region (Manhattan)	Non-motorized Crash	4 (possible injury, non-incapacitating injury, incapacitating injury and fatal injury)	Generalized ordered-response model with Composite Maximum Likelihood	√	√	√	--	√	--
(Li et al., 2013)	County (Macro)	State (California)	Any Crash	1 (fatal crash)	Geographically Weighted Poisson Regression (GWPR)	√	--	--	√	√	--
(Ye et al., 2013)	Freeway segment (Micro)	State (Washington)	Any Crash	3 (PDO, possible injury, injury/ fatality)	Joint Poisson regression model	√	--	--	√	--	√
(Barua et al., 2014)	Road segment (Micro)	City (Richmond, Vancouver)	Any Crash	2 (no injury and injury/fatal crashes)	Multivariate Poisson lognormal model	√	√	√	√	--	--
(Chiou et al., 2014)	Freeway segment (Micro)	State (Taiwan)	Motorized Crash	3 (PDO, possible injury, injury/ fatality)	Multinomial Generalized Poisson with error components	√	--	√	√	--	√
(Chiou and Fu, 2015)	Freeway segment (Micro)	State (Taiwan)	Motorized Crash	3 (PDO, possible injury, injury/ fatality)	Multinomial generalized Poisson with spatiotemporal error components	√	--	√	√	--	√
(Zhan et al., 2015)	Census tract (Macro) Roadway segment (Micro)	City, State (New York, Washington)	Pedestrian and Motorized Crash	3 (no injury, possible injury and evident injury)	Multivariate Poisson-lognormal model	√	√	√	√	√	√
(Anastasopoulos, 2016)	Highway segments (Micro)	State (Indiana)	Motorized crash	3 (PDO, injury and fatality)	Random parameter multivariate tobit model, Multivariate zero-inflated negative binomial model	√	√	--	--	--	--

(Barua et al., 2016)	Road segment (Micro)	City (Vancouver)	Any Crash	2 (no injury and injury/fatal crashes)	Bayesian multivariate random parameters spatial model	√	√	√	√	--	--
(Dong et al., 2016)	Intersection (Micro)	State (Tennessee)	Any Crash	2 (disabling injury and non-disabling injury)	Random parameter bivariate zero-inflated negative binomial model	√	--	--	√	--	--
(Bhat et al., 2017)	Census tract (Macro)	Region (Manhattan)	Pedestrian Crash	4 (possible injury, non-incapacitating injury, incapacitating injury and fatal injury)	Random coefficients multivariate count model	√	√	√	--	√	--
(Boulieri et al., 2017)	Ward (macro)	England	Any Crash	2 (slight accidents, fatal accidents	Multivariate Bayesian Model	√	--	--	√	--	--
(Chen et al., 2017)	Highway segment (Micro)	State (Indiana)	Motorized Crash	3 (PDO, possible injury, and injury/fatality)	Multivariate Random Parameters Negative Binomial Approach	√	--	--	√	--	--
(Ma et al., 2017)	Highway segment (Micro)	Interstate I70 (Colorado)	Motorized Crash	2 (injury, no injury)	Multivariate Poisson lognormal (normal, spatial and spatio-temporal)	√	--	--	√	--	√
(Wang et al., 2017)	Road segments, intersections (Micro)	State (Minnesota, Washington)	Any Crash	3 (no injury, possible/non-incapacitating injury and fatal/incapacitating injury crashes)	Multivariate Poisson Lognormal model	√	--	--	√	--	--
(Zeng et al., 2017)	Census tract (Macro) Roadway segment (Micro)	City (Hong Kong)	Any Crash	2 (slight injury crash and killed/seriously injured crashes)	Multivariate Poisson-lognormal model	√	--	--	√	--	√
(Liu and Sharma, 2018)	County (macro)	State (Iowa)	Any Crash	3 (Fatal crashes, major injury crashes, and minor injury crashes)	Multivariate spatio-temporal Bayesian model	√	--	--	√	√	√
(Rahman Shaon et al., 2019)	Highway segment (Micro)	State (Wisconsin)	Any Crash	4 (No injury, minor injury, serious injury and total injury)	Multivariate multiple risk source regression model	√	--	--	√	--	--

*FTC = Follow too close; **FOTS= Failed to observe traffic signal; ***SSV= Stop sign violation, ****ILC= Improper lane change

Table 5.2 Joint Panel Mixed NB-GOPFS Model Results (Count Component)

Variables (np)	Rear-End		Angular		Sideswipe		Head-on		Single vehicle		Non-motorized	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Constant (6)	-0.626	-7.73	-1.684	-15.407	-2.687	-21.582	-3.557	-18.081	-0.744	-9.932	-2.580	-23.487
Roadway Characteristics												
Proportion of arterial roads (2)	0.166	4.034	0.166	4.034	--[1]	--	--	--	-0.284	-5.105	0.166	4.034
Number of intersections (1)	--	--	0.347	11.804	--	--	0.347	11.804	--	--	0.347	11.804
Signal intensity (3)	0.416	2.422	--	--	-0.630	-3.277	--	--	-0.447	-1.746	0.416	2.422
Road length over 55mph (5)	0.468	3.679	-1.573	-7.679	0.468	3.679	-1.022	-2.877	0.892	7.676	-1.172	-4.591
Standard deviation	--	--	0.903	3.288	--	--	--	--	--	--	--	--
Variance of Speed (2)	0.040	3.697	0.040	3.697	0.069	4.451	--	--	--	--	--	--
Roads with separated median (2)	0.172	3.798	0.172	3.798	0.172	3.798	-0.156	-1.411	--	--	--	--
Average outside shoulder width (4)	-0.308	-7.120	-0.439	-8.323	-0.563	-9.800	-0.308	-7.120	-0.115	-2.621	--	--
Average sidewalk width (1)	--	--	--	--	--	--	--	--	--	--	-0.215	-3.693
Traffic Characteristic												
VMT (4)	--	--	0.131	8.496	0.259	16.909	0.185	8.852	--	--	0.021	1.678
Truck VMT (2)	0.179	15.852	--	--	--	--	--	--	0.270	26.819	--	--
Land-use attributes												
Urban area (4)	0.164	15.359	0.164	15.359	0.149	9.530	0.111	4.094	--	--	0.114	6.279
Office area (2)	0.148	10.384	--	--	0.148	10.384			--	--	0.127	7.389
Residential area (1)	--	--	--	--	-0.077	-6.915	-0.077	-6.915	--	--	--	--
Built environment characteristic												

No. of restaurants (3)	0.273	10.432	--	--	0.084	3.394	--	--	--	--	0.175	7.958
No. of shopping centers (1)	0.030	1.712	--	--	0.030	1.712	--	--	--	--	--	--
Socio-demographic characteristics												
Non-motorists (3)	0.052	2.892	0.148	7.166	0.168	7.581	--	--	--	--	0.052	2.892
Transit users (1)	0.222	13.287	--	--	--	--	--	--	--	--	0.222	13.287
Over dispersion (6)	0.671	16.130	0.251	6.515	0.284	8.270	1.002	6.245	0.713	19.270	0.235	4.459
Unobserved Effects												
Correlation 1 (1)	0.585	21.818	--	--	--	--	--	--	--	--	0.585	21.818
Correlation 2 (1)	--	--	0.957	43.658	0.957	43.658	0.957	43.658	--	--	--	--

*np= number of parameters estimated for each variable from a possible set of six (six crash types)

[1] --= attribute insignificant at 90% confidence level

Table 5.3 Joint Panel Mixed NB-GOPFS Model Results (Severity Component)

Variables (np)	Rear End		Angular		Sideswipe		Head-on		Single vehicle		Non-motorized	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Threshold 1	0.564	14.380	0.221	6.466	0.948	9.171	-0.038	-0.331	0.264	5.198	-0.671	-8.045
Threshold 2	-0.395	-12.342	-0.492	-19.890	-0.678	-13.290	-0.688	-9.478	-0.808	-23.022	-0.463	-10.780
Threshold 3	-0.262	-6.067	-0.373	-10.523	-0.440	-6.093	-0.506	-6.480	-0.469	-12.780	-0.062	-1.599
Roadway Characteristics												
Arterial roads (2)	0.085	3.647	0.183	4.797	--[1]	--	--	--	0.085	3.647	--	--
Possible and non-incapacitating injury (1)	-0.082	-1.719	--	--	--	--	--	--	--	--	--	--
Local roads (1)	--	--	--	--	--	--	-0.335	-2.132	--	--	-0.335	-2.132
Number of intersections (1)	--	--	--	--	--	--	-0.051	-3.549	-0.051	-3.549	--	--
Traffic signals (1)	--	--	-0.040	-4.198	-0.040	-4.198	-0.040	-4.198	--	--	--	--
Average inside shoulder width (1)	--	--	--	--	-0.171	-3.469	--	--	--	--	--	--
Average outside shoulder width (1)	-0.046	-1.697	--	--	--	--	--	--	--	--	--	--
Proportion of roads over 55mph speed (2)	0.331	5.112	0.331	5.112	--	--	0.878	3.029	0.331	5.112	0.331	5.112
Non-incapacitating and severe injury (1)	-0.667	-2.959	--	--	--	--	--	--	--	--	-1.335	-3.723
Poor pavement condition (1)	--	--	--	--	0.208	2.822	--	--	--	--	--	--
Traffic Characteristic												
Traffic Intensity (Congested) (1)	-0.074	-3.308	-0.074	-3.308	--	--	--	--	--	--	--	--
Non-incapacitating and severe injury (1)	--	--	0.123	1.980	--	--	--	--	--	--	--	--
Truck VMT (1)	--	--	--	--	0.046	4.591	0.046	4.591	--	--	--	--

Land Use Characteristic												
Urban area (2)	--	--	--	--	-0.402	-5.839	-0.402	-5.839	-0.057	-1.022	--	--
Land use mix (1)	-0.117	-2.241	-0.117	-2.241	--	--	--	--	--	--	--	--
Built environment characteristic												
No. of commercial centers (1)	--	--	--	--	--	--	--	--	--	--	-0.048	-2.101
No. of recreational centers (1)	-0.028	-2.265	--	--	--	--	--	--	--	--	--	--
No. of restaurants (1)	--	--	--	--	--	--	--	--	-0.046	-3.011	--	--
Non-incapacitating and severe injury (1)	--	--	--	--	--	--	--	--	0.049	1.650	--	--
No. of shopping centers (1)	--	--	-0.047	-4.863	-0.047	-4.863	-0.047	-4.863	--	--	--	--
Possible and non-incapacitating injury (1)	--	--	--	--	0.051	1.916	--	--	--	--	--	--
Socio-demographic characteristics												
Employee (1)	--	--	--	--	--	--	--	--	--	--	-0.084	-2.380
Motorcycle users (1)	--	--	0.134	2.354	--	--	--	--	--	--	--	--
Proportion of older people (65+) (1)	--	--	--	--	--	--	--	--	--	--	-0.460	-2.045
Household with no cars (1)	--	--	--	--	--	--	--	--	--	--	0.060	2.368
Unobserved Effects												
Between no injury and possible injury (1)	--	--	--	--	--	--	--	--	--	--	0.072	2.831
Between non-incapacitating injury and severe injury (1)	--	--	--	--	--	--	0.479	4.516	--	--	--	--

*np= number of parameters estimated for each variable from a possible set of six (six crash types)

[1] --= attribute insignificant at 90% confidence level

Table 5.4 Independent Panel NB Model Results (Count Component)

Variables (np)	Rear End		Angular		Sideswipe		Head-on		Single vehicle		Non-motorized	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Constant (6)	-0.611	-8.094	-0.990	-9.461	-1.826	-16.869	-3.080	-16.270	-0.744	-9.933	-2.523	-23.787
Roadway Characteristics												
Proportion of arterial road (2)	0.205	4.418	0.205	4.418	--	--	--	--	-0.284	-5.103	0.205	4.418
Number of intersections (1)	--	--	0.284	9.008	--	--	0.284	9.008	--	--	0.284	9.008
Signal intensity (3)	0.456	2.578	--	--	-0.577	-2.725	--	--	-0.447	-1.746	0.456	2.578
Road length over 55mph (5)	0.568	4.554	-1.451	-7.764	0.568	4.554	-1.346	-3.784	0.892	7.675	-1.298	-5.172
Variance of Speed (2)	0.039	3.499	0.039	3.499	0.067	4.564	--	--	--	--	--	--
Road with separated median (2)	0.164	3.770	0.164	3.770	0.164	3.770	-0.201	-1.741	--	--	--	--
Average outside shoulder width (4)	-0.269	-6.450	-0.381	-7.637	-0.410	-7.666	-0.269	-6.450	-0.115	-2.622	--	--
Average sidewalk width (1)	--	--	--	--	--	--	--	--	--	--	-0.201	-3.583
Traffic Characteristic												
VMT (4)	--	--	0.102	6.738	0.191	13.228	0.197	8.470	--	--	0.048	3.180
Truck VMT (2)	0.174	15.400	--	--	--	--	--	--	0.270	26.825	--	--
Land-use attributes												
Urban area (4)	0.158	14.896	0.158	14.896	0.127	8.396	0.086	3.347	--	--	0.099	5.712
Office area (2)	0.190	11.928	--	--	0.190	11.928			--	--	0.148	7.389
Residential area (1)	--	--	--	--	-0.093	-7.387	-0.093	-7.387	--	--	--	--
Built environment characteristic												
No. of restaurant (3)	0.254	8.912	--	--	0.310	9.759	--	--	--	--	0.198	8.803
No. of shopping center (1)	0.066	2.040	--	--	0.066	2.040	--	--	--	--	--	--

Socio-demographic characteristics												
Non-motorists (3)	0.067	3.996	0.145	7.109	0.144	6.858	--	--	--	--	0.067	3.996
Transit user (1)	0.222	14.898	--	--	--	--	--	--	--	--	0.222	14.898
Over dispersion (6)	0.992	30.183	1.176	25.054	1.024	21.268	1.995	8.123	0.713	19.272	0.455	8.840
Log-Likelihood (np)	-39954.27 (53)											

*np= number of parameters estimated for each variable from a possible set of six (six crash types)

Table 5.5 Independent GOPFS Model Results (Severity Component)

Variables (np)	Rear End		Angular		Sideswipe		Head-on		Single vehicle		Non-motorized	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Threshold 1	0.564	14.394	0.221	6.458	0.948	9.321	-0.039	-0.344	0.264	5.202	-0.665	-7.978
Threshold 2	-0.395	-12.332	-0.492	-19.891	-0.678	-13.289	-0.688	-9.480	-0.808	-23.019	-0.445	-10.656
Threshold 3	-0.262	-6.063	-0.373	-10.521	-0.439	-6.093	-0.505	-6.480	-0.469	-12.778	-0.062	-1.610
Roadway Characteristics												
Arterial road	0.085	3.644	0.183	4.804	--	--	--	--	0.085	3.644	--	--
Possible and non-incapacitating injury	-0.081	-1.710	--	--	--	--	--	--	--	--	--	--
Local road	--	--	--	--	--	--	-0.334	-2.128	--	--	-0.334	-2.128
Number of intersections	--	--	--	--	--	--	-0.051	-3.564	-0.051	-3.564	--	--
Traffic signal	--	--	-0.040	-4.192	-0.040	-4.192	-0.040	-4.192	--	--	--	--
Average inside shoulder width	--	--	--	--	-0.171	-3.479	--	--	--	--	--	--
Average outside shoulder width	-0.046	-1.702	--	--	--	--	--	--	--	--	--	--
Proportion of road over 55mph speed	0.330	5.101	0.330	5.101	--	--	0.876	3.026	0.330	5.101	0.330	5.101
Non-incapacitating and severe injury	-0.669	-2.965	--	--	--	--	--	--	--	--	-1.338	-3.725
Poor pavement condition	--	--	--	--	0.208	2.821	--	--	--	--	--	--
Traffic Characteristic												
Traffic Intensity (Congested)	-0.074	-3.311	-0.074	-3.311	--	--	--	--	--	--	--	--
Non-incapacitating and severe injury	--	--	0.122	1.965	--	--	--	--	--	--	--	--
Truck VMT	--	--	--	--	0.046	4.640	0.046	4.640	--	--	--	--

Land Use Characteristic												
Urban area	--	--	--	--	-0.402	-5.912	-0.402	-5.912	-0.058	-1.027	--	--
Land use mix	-0.117	-2.245	-0.117	-2.245	--	--	--	--	--	--	--	--
Built environment characteristic												
No. of commercial centers	--	--	--	--	--	--	--	--	--	--	-0.049	-2.140
No. of recreational centers	-0.028	-2.270	--	--	--	--	--	--	--	--	--	--
No. of restaurants	--	--	--	--	--	--	--	--	-0.046	-3.011	--	--
Non-incapacitating and severe injury	--	--	--	--	--	--	--	--	0.049	1.660	--	--
No. of shopping centers	--	--	-0.047	-4.862	-0.047	-4.862	-0.047	-4.862	--	--	--	--
Possible and non-incapacitating injury	--	--	--	--	0.051	1.918	--	--	--	--	--	--
Socio-demographic characteristics												
Employee	--	--	--	--	--	--	--	--	--	--	-0.083	-2.381
Motorcycle user	--	--	0.134	2.351	--	--	--	--	--	--	--	--
Proportion of older people (65+)	--	--	--	--	--	--	--	--	--	--	-0.443	-1.968
Household with no cars	--	--	--	--	--	--	--	--	--	--	0.060	2.384
Sample Size	**2992**		**2585**		**2116**		**806**		**2510**		**1417**	

*np= number of parameters estimated for each variable from a possible set of six (six crash types)

CHAPTER 6: ACCOMMODATING POPULATION HETEROGENEITY WITHIN A PANEL MODEL FRAMEWORK

Road traffic crashes and their consequences (property damage, injuries and fatalities) remain a global health concern given the extent of societal, emotional and economic impacts of these unfortunate events. According to recent report by NHTSA 2018, road traffic crashes, responsible for 36,750 fatalities in the US, ranked as the third deadliest in the decade and a leading cause of death among people aged between 17 and 21 years old. The numbers are declining relative to 2016 and 2017 but still it is 12.2% higher than 2014 (all time low, 32,544) (NHTSA, 2018) which warrants devising appropriate solutions for reducing the number and consequence of such unfortunate events. A major tool employed in the literature to develop counter measures is the application of econometric models for crash frequency analysis. Econometric approaches for developing crash prediction models in safety literature are dominated by traditional count regression frameworks (Poisson and negative binomial (NB)) in a univariate modeling setting. These approaches identify a single count variable (typically the total number of crashes) for a spatial unit and study the impact of exogenous variables. However, studies show that a single total crash model will not be able to parse the distinct crash distribution by different attributes (such as type, injury severity, modes) and such aggregation can result in aggregation bias and loss of information available in the dataset. For example, consider the exogenous variable - presence of left guardrail on the roadway. In the presence of a left guardrail, vehicles are prevented from entering the opposite direction thus reducing head-on crashes. On the other hand, vehicles on hitting the guardrail might collide with other vehicles travelling in the same direction which in turn resulting in an increase in rear-end, sideswipe, angular crashes. So, the overall impact of the

guardrail on total number of crashes would yield a positive sign despite having distinct impact on different crash types.

Hence, it is not surprising that in recent years, safety researchers have focused on disaggregating the data by various attributes such as crash typology (such as head-on or rear-end), injury severity (such as crashes by no injury or crashes by severe injury) and crash location (such as intersection versus non-intersection). In this case, an extension of univariate approach would be to develop multiple univariate models considering counts by different attribute levels as multiple dependent variables (please see (Lord and Mannering, 2010; Yasmin and Eluru, 2018) for a literature review). The separate models allow us to capture distinct and realistic impacts of exogenous variables on different count dimensions. However, these approaches only accommodate for observed factors and inherently neglects the unobserved heterogeneity. In recent years, there is growing recognition that crash counts across different attributes are correlated and hence multivariate in nature. For instance, higher number of blind spots at intersections along a corridor (usually unobserved to analysts) are likely to result in higher number of vehicular conflicts as well as possibly higher number of pedestrian/bicyclists involved crashes. Ignoring such correlation, if present, may lead to biased and inefficient parameter estimates resulting in erroneous policy implications (see Mannering et al., 2016 for an extensive discussion). Recognizing this drawback, several research efforts have developed frameworks that accommodate for the influence of these common unobserved factors employing multivariate modeling approaches. However, there are still several methodological challenges associated with these models in accommodating unobserved heterogeneity. In this context, the current research contributes towards addressing the methodological challenges in crash frequency models by employing an alternative econometric

approach for analyzing multiple crash frequency variables while capturing unobserved heterogeneity.

6.1 Earlier Research

The most common approach employed to address the potential unobserved heterogeneity in safety literature is the development of multivariate crash frequency models (please see (Bhowmik et al., 2019a, 2018; Yasmin and Eluru, 2018) for detailed review on multivariate approach).. Based on the process of capturing correlations, the multivariate approaches can broadly be classified along two major streams: (1) simulation-based approaches and (2) analytically closed-form based approaches. The main difference between these two streams lies in how the dependency across dimensions is captured. In simulation based multivariate approaches, the different propensities are correlated by generating a common error term across dimensions. Further, probability computation requires integrating the probability function over the error term distribution due to the unobserved nature of the error term (see (Bhowmik et al., 2019a)). The exact computation is dependent on the distributional assumption and does not usually have a closed form expression. These approaches rely either on Maximum Simulated Likelihood (MSL) in the classical realm or Markov Chain Monte Carlo (MCMC) approach in the Bayesian realm for model estimation (Anastasopoulos et al., 2012; Aguero-Valverde, 2013; Wang and Kockelman, 2013; Barua et al., 2014; Dong et al., 2014). However, the complexity of the model estimation is dependent on the number of unobserved parameters estimated. Further, applying simulation for such joint processes is likely to be error-prone and the stability of the variance-covariance matrix is often sensitive to model specification and number of simulation draws (see (Bhat, 2011) for a discussion).

On the other hand, the closed form based approach referred to as the copula framework relies on developing multivariate distributions (or approximations of multivariate distributions) with analytical closed form probability expressions that obviate the need for simulation. These model frameworks are estimated employing maximum likelihood or composite maximum likelihood approaches (Wang et al., 2015; Nashad et al., 2016; Yasmin et al., 2018). One of the limitations of the simulation approaches is its accuracy that is strongly tied to the dimensionality of integration (number of unobserved parameters estimated) as well as number of draws considered for the probability function evaluation. On the other hand, in the closed-form regime, the propensity equations for frequency dimensions are tied together by analytical multivariate distributional assumptions. The likelihood function is quite complicated, but once programmed, these closed-form frameworks are less prone to error[7]. Further, the copula framework allows for flexible correlation structures (radial symmetry and asymmetry, and asymptotic tail independence and dependence) across joint dimensions thus enhancing the flexibility of the multivariate approach which results in more stable inference conclusion.

The literature clearly highlights the prevalence of multivariate model frameworks in safety literature. However, there are two major challenges associated with the existing multivariate approach in estimating observed and unobserved effects (see chapter 4 for detail). First, in multivariate approaches, a separate crash propensity equation is adopted for each crash type. Within each propensity equation, we estimate a number of observed parameters in the model

[7]In some cases, where such formulas are of very high dimensions they might not be analytically tractable. In this case, an composite maximum likelihood approach is adopted (Bhat, 2014, 2011; Narayanamoorthy et al., 2013).

structures. For instance, let us consider two crash types: rear-end and sideswipe and we developed a crash prediction model at a zonal level. Now, it is possible that zonal AADT has similar effect on zonal level rear-end and sideswipe crashes. To test this, researchers need to perform a log-likelihood ratio test and with increased number of dependent and independent variables, this process will be burdensome. Due to the additional burden, it is often cumbersome to check whether the variable effects are really different across the propensities. As a result, with higher number of dependent and independent variables, the number of parameters estimated are likely to be very high. Second, the complexity of the model estimation is dependent on the number of unobserved parameters estimated. In traditional multivariate models, the influence of unobserved factors is typically accommodated as random effects and error correlations across dimensions. The random effect referred to the unobserved factors affecting crash propensity within the dimension while the error correlation parameters account for the influence of unobserved factors affecting multiple dependent variables. These effects require simulation[8] for parameter estimation and with higher dimensions (number of dependent variable and unobserved factors to be estimated), the model infrastructure can get computationally demanding.

6.2 Current Study

The work presented in chapter 4 address these challenges by recasting the multivariate crash frequency modeling problem as a pooled univariate crash frequency (with unobserved heterogeneity accommodated) analysis problem. To elaborate, instead of considering the crash

[8] For the closed form approaches, we don't need simulation for correlation but for estimating random effects, we need to rely on simulation.

frequency by crash type as a multivariate distribution, the authors represent it as repeated measures of crash frequency while recognizing that each repetition represents a different crash type. The recasting process allows for the estimation of a parsimonious model system by allowing for an improved specification testing of variable impacts across different crash types (see chapter 4 for detail). Using this consideration, the proposed model system enhances the efficiency of estimation through a single crash frequency model while also allowing for parameter effects to vary across different crash types through crash type specific deviation terms. Further, as only one propensity equation is to be estimated, it allows for reduction in parameters especially for unobserved factors resulting in substantial improvements in model efficiency and computational times. The study presented in chapter 4 accommodate population heterogeneity through unobserved effects. However, it is possible that the influence of observed exogenous factors also might vary across different TAZs and ignoring such heterogeneity might be erroneous.

To illustrate the importance of varying impact of exogenous variables, let us consider the number of non-motorized crashes in two zones (Z_1 and Z_2) with identical attributes except average sidewalk width. Z_1 has lower average sidewalk width while the sidewalk is much wider in Z_2. Now, let us consider the effect of non-motorist activity in these zones. Higher non-motorist activity is associated with higher number of non-motorized crashes. Therefore, Z_1 with narrow sidewalk will experience increased number of non-motorized crashes with higher pedestrian activity. On the other hand, Z_2 has wider sidewalk which will provide additional safety to the non-motorists from colliding with a motorized vehicle and as a result, the impact of non-motorists on non-motorized crashes will be less relative to Z_1 and even in some cases, it could be negative. This is an example of the effect of non-motorist activity exhibiting differential impact on the number of non-motorized crashes based on the width of sidewalk. The illustration provided is a case of one variable (average

sidewalk width) moderating the influence of another variable (number of non-motorists). However, in the context of crash frequency analysis by different crash types, it is possible that multiple variables might serve as a moderating influence on a reasonably large set of exogenous variables. Hence, evaluating crash counts employing a traditional model (population homogeneity assumption) might possibly lead to incorrect coefficient estimates.

There are a number of approaches employed in safety literature to account for such systematic heterogeneity including market segmentation, clustering technique, and random effect models (see (Eluru et al., 2012; Yasmin and Eluru, 2016) for detail). However, market segmentation or clustering approaches allocates data records exclusively to a particular cluster or segment based on the attributes and do not consider the possible effects of unobserved factors that may moderate the impact of observed exogenous variables. Another problem is these techniques might result in reduced sample size in some segments/clusters which in turn result in loss of model estimation efficiency. Random effects model is another alternative approach to capture population heterogeneity. These approaches, though attractive, are focussed on the error component of the model and usually require extensive simulation for model estimation while also not considering observed heterogeneity. To that extent, latent segmentation model offers an approach to incorporate population heterogeneity within the systematic component. In a latent segmentation model, TAZs are allocated probabilistically to different segments and a segment specific model is estimated for each segment. Such an endogenous segmentation scheme is appealing for multiple reasons including: (1) it ensures that the parameters are estimated employing the full sample for each segment while employing all data points for model estimation; (2) provides valuable insights on how the exogenous variables affect segmentation; and (3) the probabilistic assignment explicitly acknowledges the role played by unobserved factors in moderating the impact of

observed exogenous variables. To be sure, using latent segmentation approach in crash count literature is not new (micro level: (Park et al., 2010; Park and Lord, 2009; Zou et al., 2014); macro level: (Yasmin and Eluru, 2016)). However, earlier research on latent class models have been restricted to considering only one count dependent variable. To the best of authors knowledge, this study is the first of its kind to develop a latent class count model considering different crash types while accommodating potential correlations across the count dimensions.

To summarize, our current objective contributes to crash frequency literature both methodologically and empirically by estimating a latent segmentation-based Panel Negative Binomial (LPNB) to study the zonal level crash counts across different crash types. The current research effort extends the previous work presented in chapter 4 by introducing the latent class version of the panel negative binomial (PNB) model to capture the potential variation in the impact of exogenous variables while also explicitly accommodating for unobserved heterogeneity through random parameters and error correlations. The newly formulated model will allow us to partition the TAZs into segments based on their attributes and estimate the influence of exogenous variables on crash counts of different crash types. From *methodological* perspective, the current research makes a threefold contribution to literature on crash frequency analysis: First, the recasting allows us to estimate a parsimonious model system and also reduce the computational time for estimating parameters associated with unobserved factors. Second, by introducing the latent class version of the PNB model, we allow for both observed and unobserved heterogeneity thus relaxing the homogeneity assumption of the traditional count models. Third, we allow for a flexible segment membership function and test for the presence of multiple segments in the model estimation. *Empirically*, the research contributes to our understanding of analyzing zonal level crashes for both motorized and non-motorized road user group while considering different crash types within the

motorized category including rear-end, angular, sideswipe, single vehicle and head-on crashes. The analysis is conducted using the zonal level crash records from Central Florida for the year 2016 considering a comprehensive set of exogenous variables. Further, we undertake a comparison exercise of the proposed LPNB model with its' traditional counterpart proposed in chapter 4 .

The rest of the chapter is organized as follows: the next section presents the methodological framework adopted in the analysis while the section 6.4 provides a detailed description of the model findings. The comparison results are discussed in section 6.5 followed by concluding thoughts in the last section (6.7).

6.3 Econometric Framework

The focus of our current objective is to estimate a latent segmentation based panel mixed NB modeling framework and compare its performance with previously proposed panel mixed NB model. The empirical analysis involves estimation of two different frameworks including: Panel mix NB model with and without the latent segmentation. For the sake of brevity, we will restrict ourselves to the discussion of the latent class model only (please see chapter 4 for the detailed methodology on the Panel Mixed NB model).

As highlighted earlier, we alter the dataset by taking all six types of crashes as repeated measures (same TAZ is repeated 6 times) of crash frequency in a univariate NB formulation while recognizing that each repetition represents a different crash type. In the latent segmentation based approach, crash count records by different crash types for TAZs are probabilistically assigned to s relatively homogenous (but latent to the analyst) segments based on various explanatory variables. Within each segment, the effects of exogenous variables on the number of crashes by different crash types occurring across the TAZ over a given period of time are fixed in the segment.

Hence, the latent segmentation based model consists of two components: (1) assignment component and (2) segment specific count model component. The general structure for all latent segmentation based count models involves specifying these two components. For the ease of presentation, we describe modeling framework by the components.

6.3.1 Assignment Component

Let us assume that s be the index for segments ($s = 1, 2,3, \dots, S$), i be the index for TAZ ($i = 1,2,3, \dots, N = 3{,}815$) and $r(r = 1,2, \dots, R, R = 6)$ be an index for different crash type at TAZ i. y_{ir} be the index for crash counts occurring over a period of time in TAZ *i* and crash type *r*. The assignments of TAZ to different segments are modeled as a function of a column vector of exogenous variable by using the random utility based multinomial logit model (see (Dey et al., 2018; Eluru et al., 2012; Wedel et al., 1993; Yasmin and Eluru, 2016) for similar formulation) as:

$$P_{is} = \frac{exp[\boldsymbol{\alpha}_s \boldsymbol{z}_s]}{\sum_{s=1}^{S} exp[\boldsymbol{\alpha}_s \boldsymbol{z}_s]} \tag{43}$$

where, P_{is} is the probability of TAZ i to be assigned to segment s, $\boldsymbol{z}_s$ is a vector of attributes and $\boldsymbol{\alpha}_s$ is a conformable parameter vector to be estimated. The assignment process is the same for all latent class models.

Within any latent segmentation approach, the unconditional probability of y_{ir} can be given as:

$$P_i(y_{ir}) = \sum_{s=1}^{S} \big(P_i(y_{ir}|s)\big) \times (P_{is}) \tag{44}$$

where $P_i(y_{ir}|s)$ corresponds to the probability of count y_{ir} in segment s.

6.3.2 Segment Specific Count Component

As mentioned earlier, we estimated a panel mixed univariate NB modeling framework within each segment. Then the probability equation of the NB formulation can be rewritten as follow:

$$P_{is}(y_{ir}|s) = \frac{\Gamma\left(y_{ir} + \frac{1}{\lambda'}\right)}{\Gamma(y_{ir}+1)\Gamma\left(\frac{1}{\lambda'}\right)}\left(\frac{1}{1+\lambda' v_{ir}}\right)^{\frac{1}{\lambda'}}\left(1-\frac{1}{1+\lambda' v_{ir}}\right)^{y_{ir}} \quad (45)$$

where, $P(y_{ir})$ is the probability that TAZ i has y_{ir} number of crashes for crash type r. λ' is NB over dispersion parameter and v_{ir} is the expected number of crashes occurring in i over a given time period for crash type r. v_{ir} an be expressed as a function of explanatory variables using a log-link function as follows:

$$v_{ir} = E(y_{ir}|\boldsymbol{x}_{ir}) = exp((\boldsymbol{\beta} + \boldsymbol{\theta}_i + \boldsymbol{\varrho}_{ir})\boldsymbol{x}_{ir} + \varepsilon_{ir}) \quad (46)$$

where, $\boldsymbol{x}_{ir}$ is a vector of explanatory variables associated with observations i for crash type r. $\boldsymbol{\beta}$ is a vector of coefficients to be estimated. $\boldsymbol{\theta}_i$ is a vector of unobserved factors moderating the influence of attributes in $\boldsymbol{x}_{ir}$ on the crash count propensity for TAZ i, $\boldsymbol{\varrho}_{ir}$ is a vector of unobserved effects specific to crash type r. ε_{ir} is a gamma distributed error term with mean 1 and variance λ'. In estimating the model, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\theta}, \boldsymbol{\varrho}$ represented by Ψ. In this framework, it is assumed that these elements are drawn from independent normal distribution: $\Psi \sim N(0, (\boldsymbol{\pi}'^{\mathbf{2}}, \boldsymbol{\Phi}^{\mathbf{2}}))$. This $\boldsymbol{\varrho}_{ir}$ will be same across crash types in our case and thus the unobserved heterogeneity across crash types will be captured. Moreover, $\boldsymbol{\theta}_i$ term will capture the random effect across observations.

6.3.3 Model Estimation

Thus, conditional on Ψ, the likelihood function for the latent segmentation based count model across TAZ can be expressed as

$$L_i = (\int_{\Psi} \prod_{r=1}^{R} \left(\sum_{s=1}^{S} \left(P_i(y_{ir}|s) \right) \times (P_{is}) \right) f(\Psi) d\Psi \quad (47)$$

Finally, the log-likelihood function is:

$$LL = \sum_{i} Ln(L_i) \quad (48)$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 48. The parameters to be estimated in the model are: $\boldsymbol{\theta}$, ϱ, Φ and π. To estimate the proposed model, we apply Quasi-Monte Carlo simulation techniques based on the scrambled Halton sequence to approximate this integral in the likelihood function and maximize the logarithm of the resulting simulated likelihood function across individuals (see (Bhat, 2001; Eluru et al., 2008) for examples of Quasi-Monte Carlo approaches in literature).The model estimation routine is coded in GAUSS Matrix Programming software (Aptech).

6.4 Model Specification and Overall Measure of Fit

As discussed earlier, out of 4,747 TAZs, 3,815 TAZs are randomly selected for model estimation and the remaining TAZs (932) are set aside for validation purpose. The number of count dependent variables (crash types) to be analyzed in the current study is six and so every TAZ is repeated six times recognizing that each repetition represents a different crash type (see chapter 4 for detail). Thus, the estimation sample has 22,890 (3,815*6) records and the validation sample has 5,592 (932*6) data records. The empirical analysis involved a series of model estimations. First, we

estimated six separate independent NB model for six crash types to establish a benchmark for comparison. Second, we estimated a parsimonious model structure (Panel independent NB model) using the same independent model system while restricting the parameters across different crash types considered. To elaborate, we estimate a base effect for each exogenous variable that is common across the crash types and estimate deviations for each crash types relative to the base effect. If a deviation is insignificant, it concludes that there is no significant difference in effect for that particular variable between the base crash type and crash type for which the deviation was computed ((see chapter 4 for more details). Thus, the model estimated in such panel formulation results in fewer parameters. Third, we estimated a latent class version of the panel negative binomial (LPNB) model to capture the potential variation in the impact of exogenous variables. Fourth, within the Panel NB model and Latent panel NB model, we consider unobserved heterogeneity in terms of correlation (across the crash count dimensions) and random effect (within the crash count propensity).

6.4.1 Determining Appropriate Number of Segments for Latent Models

In case of the latent models, determining the appropriate number of segments is a critical issue with respect to interpretation and inferences. The estimation process for such latent class model begins with the independent model considering two segments. Then we continued adding additional segments until further addition does not enhance intuitive interpretation and data fit ((Eluru et al., 2012). The decision regarding the optimal number of segments is taken considering criteria like Bayesian Information Criterion (BIC) as well as the interpretability and model parsimony. Specifically, we estimated independent latent NB model with different number of segments (2, 3…) and select the model with the lowest BIC value. Once, the independent latent

model is finalized with appropriate number of segments, we estimated the mixed version of the corresponding independent model.

Within the latent independent Panel NB frameworks, we estimated two models including i) LPNB model with two segments and ii) LPNB model with three segments. The BIC values for these estimated models are: i) LPNB model with two segments is 80, 250.87 (log-likelihood is -39,890.40 with 57 parameters) and ii) LPNB model with three segments is 80, 157.94 (log-likelihood is -39,197.98 with 58 parameters). Based on the BIC value, we can observe that the three segments model provide improved data fit. However, the sample share of one of the segments for the three segments model represents only 5% of the TAZs and thus does not yield any interpretable segment characteristics. As a result, we did not proceed further in adding segments and select the model with two segments as the preferred model for the current analysis. From here on, we restrict ourselves to the discussion of only the LPNB model with two segments.

6.4.2 Comparison Between Models

In summary, we estimated total five models in two regimes: a) unsegmented models including: 1) Independent NB model; 2) Panel independent NB model (PNB); 3) Panel Mixed NB model (PMNB); and b) segmented model including: 4) Latent Panel Independent NB model with two segments (LPNB II) and 5) Latent Panel Mixed NB model with two segments (LPMNB II). Finally, we compare the unsegmented models with the latent segmentation based count models in order to assess the importance of accounting for population heterogeneity in estimating zonal level crash frequency models. The reader would note that all the models mentioned above are non-nested in nature and so, we employ the BIC measure for the comparison exercise.

The log-likelihood values at convergence for these estimated models are: 1) Independent NB model (68 parameters) is -39,954.90; 2) PNB (52 parameters) is -39,961.82; 3) PMNB (53 parameters) is -39,235.75; and b) segmented model including: 4) LPNB II (57 parameters) is -39,890.40 and 5) LPMNB II (57 parameters) is -39,352.26. The corresponding BIC values for these models are: 1) Independent NB model is 80,592.41; 2) PNB is 80,352.47; 3) PMNB is 78,908.57; and b) segmented model including: 4) LPNB II is 80,250.87 and 5) LPMNB II is 79,174.58. Based on the BIC values, several observations can be made. First, the PNB model that accounts for penalty for additional parameters provide improved data fit compared to the independent NB model. This supports our hypothesis that the impact of some variables may not differ across the crash types and through the recasting, we can have a parsimonious model system with improved parameter efficiency. Second, the segmented independent LPNB II model performs better relative to the PNB model. This result provides strong evidence in favour of our hypothesis that crash counts by different crash types can be investigated in a more efficient way through the segmentation of the TAZs. Third, models accommodating unobserved effects perform better than their corresponding independent models in both unsegmented (PMNB vs PNB) and segmented regimes (LPMNB II vs LPNB II) highlighting the importance of accommodating for unobserved heterogeneity in examining crash count by different crash types. Fourth, within the mixed models, the unsegmented model (PMNB) provides improved data fit relative to the segmented model (LPMNB II). Based on the results provide above, we can conclude that the segmented model is a preferred choice as long as the framework is estimated in a closed form structure (independent models that do not account for unobserved heterogeneity; no need for simulation). However, when we rely on simulation for capturing the unobserved effects, the unsegmented model outperforms its segmented counterparts.

6.5 Estimation Results

This section offers a detailed discussion of exogenous variable effects on the crash count outcome for different crash types. Table 6.2 presents the model estimation results for the proposed Latent Panel Mixed NB model (LPMNB II). The estimation results of the PMNB model are presented in Table 6.3 for comparison. In discussing the model results, for the sake of brevity, we will restrict ourselves to the discussion of the LPMNB II model only. As discussed earlier, the latent models are comprised of two parts including segmentation component and segment specific count component. For the ease of presentation, we first present an intuitive discussion of the segmentation component followed by the segment specific count component by different variable groups.

6.5.1 Segmentation Component

6.5.1.1 Descriptive Characteristics of the Segments

To delve into the segmentation characteristics, the model estimates are used to generate information on two criterion including: 1) percentage TAZ share across the two segments, and 2) expected mean of crash count events of different crash types within each segment (see (Eluru et al., 2012) for detail). Table 6.1 provides these estimates. From the estimates, it is clear that the likelihood of a TAZ being assigned to segment 1 is substantially higher than the likelihood of being assigned to segment 2 (0.74 vs 0.26). Further, the expected number of crash counts by different crash types conditional on their belonging to a particular segment offer contrasting results indicating that the two segments exhibit distinct crash risk profiles for different crash types in the current study. As evident form table 6.1, we can observe that relative to observed sample mean, the expected mean crash counts by different crash types is higher in segment 1 (except head -on)

while in segment 2, the expected mean is lower for every crash types except head-on crashes. Interestingly, we find that, segment 2 has higher risk for head-on crashes relative to segment 1. Based on overall results, it is clear that a TAZ, if allocated to segment 1 is likely to experience higher number of crashes by most of the crash types than if allocated to segment 2. Thus, we may label segment 1 as the "high risk segment" and segment 2 as the "low risk segment".

6.5.1.2 Segment Membership Component

The latent segmentation component determines the relative prevalence of each segments, as well as the likelihood of a TAZ being allocated to one of the two segments based on some zonal level exogenous variables. In our analysis, we find that segment share is influenced by zonal level roadway and land use attributes. In particular, number of intersections, average outside should width, urban area and residential area in a zone affect the assignment of a TAZ to a segment. The first row panel of Table 6.2 represent the effect of these control variables. In the segmentation component, one of the segment must be the base for every variable for the sake of identification. In our current analysis, the high risk segment (segment 1) chosen to be the base and the coefficients presented in the table correspond to the propensity for being a part of the low risk segment (Segment 2). Thus, a positive (negative) sign for a variable in the segmentation component indicates that TAZs with the variable characteristics are more (less) likely to be assigned to the low risk segment relative to the high risk segment.

The positive sign on the constant does not have any substantive interpretation and simple indicates the larger size of the low risk segment relative to the high risk segments. From the estimated results, we can observe that higher number of intersections in a zone increase the likelihood of assigning the TAZ to the high risk segment while TAZ with wider shoulder width

have a higher probability to be allocated to the low-risk segment. TAZ with more urbanized area are more likely to be assigned to the high-risk segment. On the other hand, with increase in residential area, the likelihood of assigning the TAZ to low risk segment increases. Based on these results, we can argue that high risk segment consists of urbanized zone having higher number of intersection with narrow average outside shoulder and less residential area. On the other hand, zones within segment 2 are more likely to be characterized by rural area with less number of intersections, wider average outside should width and more residential area.

6.5.2 Segment Specific Count Component

The coefficients in Table 6.2 represent the effect of exogenous variables on the frequency component of each crash type within each segment. The reader would note that, within each segment, the variables in the crash count component of Table 6.2 with positive (negative) sign indicates that an increase in the variable is likely to result in more (less) crashes. In the subsequent sections, we provide a discussion of model results for different crash types by segment groups.

6.5.2.1 High Risk Segments (Segment 1)

The crash risk component for different crash types within the high risk segment (segment 1) is discussed in this section by variable groups. Within the high-risk segment, the impact of explanatory attributes within different groups are along expected lines.

6.5.2.1.1 Crash Specific Constants:

The crash specific constants represent the intercept of crash propensity after adding the various exogenous variables and do not have any substantive interpretation.

6.5.2.1.2 Roadway Characteristics:

The results regarding the impact of proportion of arterial roads reveal that a TAZ with higher proportion of arterial roads is more likely to experience increased incidence of rear-end, angular and non-motorized crashes while the number of single vehicle crashes reduces. This is expected as single vehicle crashes usually occur on high speed roads while on arterial roads, drivers are restricted to operate at lower operating speed due to higher vehicular interactions. At the same time, the increased traffic interactions result in higher number of rear-end, angular crashes and non motorized crash ((Bhowmik et al., 2019b). Further, the estimated results show that TAZ with higher variance in speed limit results in higher number of rear-end, sideswipe and non-motorized crashes within the high risk segment. Interesting thing to note is that the influence of variance of speed limit is not different for the three crash types which support our hypothesis that the impact of some variables may not differ across the crash types. Traditional approaches in frequency modeling would have estimated three separate parameters for the three crash types while in our approach, a single parameter is adequate to accommodate for the impact of the variable (variance of speed limit).

In terms of proportion of roads over or equal 55mph speed limits, we find contrasting results across different crash types within the high risk segment. For instance, the positive coefficient offered by the variable on rear-end , sideswipe and single vehicle crashes (same effect) indicates an increased likelihood of these crash types in a TAZ having higher percentage of roads over 55mph speed limit. On the other hand, the estimated results show that TAZ with more high-speed roads (≥55mph) results in reduced incidence of angular, head-on and non-motorized crashes. The result is expected since high speed roads are usually straight (less curvature) with a divider or

median which eventually reduce the risk of angular and head-on crashes. Further, we found that the impact of the proportion of road over 55mph has significant variability on angular crashes (indicated by the standard deviation parameter) which implies that the overall impact is most likely to be negative (98%).

6.5.2.1.3 Land-use Characteristics:

Within the high risk segment, the only land use characteristic influencing crash risk by different crash types is the amount of office area in a zone. As evident from Table 6.2, we can see that office area is positively associated with rear-end, sideswipe and non-motorized crashes indicating a higher likelihood of these crash types in a TAZ with increased office areas. This variable basically reflects the presence of higher vehicular and non-motorist interactions and in turn, higher exposure for both road user groups.

6.5.2.1.4 Built Environment Characteristics:

In terms of built environment attributes, we considered a number of variables, among which only number of restaurants and shopping centers have significant impact on zonal level crash risks within the high risk segment. In particular, higher number of restaurant and shopping centers in a TAZ results in higher incidence of rear-end and sideswipe crashes perhaps due to the higher density of traffic volume for these zones. With respect to non-motorized crashes, number of restaurants is found to be a significant determinant with a positive impact (see (Yasmin et al., 2018a) for similar result).

6.5.2.1.5 Traffic Characteristics:

The parameters associated with traffic characteristics offer expected results. The parameter associated with VMT proxies for traffic volume reveals a positive impact on angular, sideswipe, head-on and non-motorized crashes indicating a higher risk of such crashes in a TAZ with increased VMT. Interestingly, the study found no significant impact of the truck volume on any of the crash types within the high risk segment.

6.5.2.1.6 Socio-demographic Characteristics:

For socio-demographic attributes, we consider the number of non-motorists (walk/bike) and transit commuters in a zone serving as additional exposure measures for the crash risk model. As evident from table 6.2, our analysis shows that TAZ with increased number of non-motorist commuters is likely to experience increased number of rear-end, sideswipe, non-motorized and angular crashes. In fact, the reader would note that the magnitude of these impacts is same across the three crash types (rear-end, sideswipe and non-motorized) while a more profound impact is observed for the angular crashes. On the other hand, the likelihood of being involved in a rear-end and non-motorized crashes increases with increasing share of transit commuters in a zone.

6.5.2.1.7 Unobserved Common Factors:

The final set of variables in Table 6.2 correspond to the potential correlation affecting zonal level crash counts by different crash types simultaneously. The reader would note that, in estimating the model, we found significant impact of two common unobserved components including (1) common unobserved factors affecting rear-end and non-motorized crashes and (2) common unobserved factors affecting angular, sideswipe and all single vehicle crashes. Overall, the results

clearly indicates the presences of common unobserved heterogeneity across different crash types within the high risk segment.

6.5.2.2 Low Risk Segments (Segment 2)

The crash risk component for different crash types within the low risk segment (segment 2) is discussed in this section by variable groups. Similar to the high-risk segment, the effect observed for different attributes on different crash types are also intuitive in the low risk segment. As evident form table 6.2, we can see that the crash count propensity for different crash types for the "low risk" segment provides variable impacts that are significantly different, in magnitude (for a few variables), from the impacts offered by the exogenous variables in "high risk" segment. Additionally, the number of variables influencing the zonal level crash frequency by different crash types are significantly lower in the low risk segment relative to the high risk segment which further highlights the difference between the two segments.

6.5.2.2.1 Crash Specific Constants:

Similar to the high risk segment, the crash specific constants in the low risk segments also represent the intercept of crash propensity after adding the various exogenous variables and do not have any substantive interpretation. As expected, we can observe

6.5.2.2.2 Roadway Characteristics:

As in the high risk segment, proportion of arterial roads offers a negative influence on single vehicle crashes in the low risk segment also (same reasoning as segment 1) though the magnitude is much higher in the low risk segment. One possible explanation can be attributed to the fact that

segment 2 consists of zone with wider outside average shoulder width. Outside shoulder width in a road reflects the extra margin of safety for vehicular maneuvers and thus reduce the potential for single vehicle crashes. Further, the parameter associated with signal intensity offers contrasting effects on different crash types. While an increase in the variable positively influence the rear-end, sideswipe and non-motorized crashes, a negative associated is observed for single vehicle crashes. This is intuitive as with more signals on the road, the traffic density increases thus results in increased conflicts between vehicles to vehicles and vehicles to non-motorists. At the same time, these conflicts result in lower operating speed which in turn reduce the potential for single vehicle crashes. Interesting thing to note is that the influence of signal intensity is not different for the three crash types (rear-end, sideswipe and non-motorized) which again lends support to our hypothesis that the impact of some variables may not differ across the crash types.

Similar to the segment 1, variance of speed limit reflects a same positive impact on rear-end, sideswipe and non motorized crashes in segment 2, but the impact is more profound in the second segment. Further our analysis shows that TAZ with higher proportion of high-speed roads (≥55mph) is more likely to experience increased number of single vehicle crashes relative to other zones in the low risk segment. Relative to segment 1, the effect (magnitude) is less in the low risk segment. In addition, we found that proportion of road over 55mph has significant variability specific to single vehicle crashes as indicated by the standard deviation parameter. The reader would note that the distributional parameter indicates that the overall impact of the variable on single vehicle crashes is likely to be positive (84%). In terms of proportion of road with separate median, the variable is found to have the same positive effect on rear-end, angular and sideswipe crashes while a negative coefficient is observed for head-on crashes. Separated median such as guardrail on a road provide additional safety margin to a vehicle from colliding with the opposite

direction traffic thus reduce the risk for head-on crashes. At the same time, vehicle hitting the guardrail have a higher likelihood of colliding with same direction traffic and hence the positive impact is also intuitive.

6.5.2.2.3 Land-use Characteristics:

For low risk segment, none of the variables within land use characteristics are found to significantly influence zonal level crash counts of any crash types in the current study context.

6.5.2.2.4 Built Environment Characteristics:

Similar to the land use attributes, we did not find any variable specific to build environment characteristics to significantly affect the zonal level crash counts of different crash types in the low risk segment.

6.5.2.2.5 Traffic Characteristics:

Unlike the high risk segment, we did not find any significant impact of VMT on any crash types. In terms of traffic characteristics, the only variable influencing the crash counts of different crash types in the low risk segment is the truck VMT. Truck VMT serves as a surrogate for exposure for truck volume. As expected, truck VMT is found to positively influence the rear-end and all single vehicle crash propensity indicating a higher risk of getting involved in rear-end and all single vehicle specific crashes with increased exposure to truck volume.

6.5.2.2.6 Socio-demographic Characteristics:

With respect to socio-demographic characteristics, we find that increased presence of transit commuters are associated with higher risk of rear-end and non-motorized crashes in the low risk segment (same as high risk segment). However, the magnitude of the impact of the variable is more pronounced in the low risk segment.

6.5.2.2.7 Unobserved Common Factors:

Within the low risk segments, we found the presence of common unobserved factors affecting angular, sideswipe and all single vehicle crashes simultaneously. Unlike high risk segments, we did not find any other common factors between rear-end and non motorized crashes.

6.6 COMPARISON EXERCISE

6.6.1 Predictive Performance

In an effort to assess the predictive performance of the estimated models, we compute several goodness fit of measures at disaggregate level including MPB (Mean prediction bias), MAD (mean absolute deviation), MAPE (mean absolute percentage error), RMSE (Root mean square error) and predictive log-likelihood (see chapter 4 for a discussion on estimating these measures). Specifically, we employ these measure on two datasets: 1) in-sample dataset: for the records used in the model estimation (sample size = 3,815 TAZs) and 2) holdout sample: records that are set aside for validation analysis (sample size = 932 TAZs). The reader would note that model with lower value of predictive measures and higher value of predictive log-likelihood will reflect better performance in terms of prediction and statistical fit relative to the observed data. Table 6.4

presents the values of these measures for PMNB and LPMNB models for both in-sample and holdout-sample measures.

Several observations can be made based on the measures presented in Table 6.4. First, a total of 70 prediction measures are estimated considering six crash types and total crash counts in both estimation and validation sample. Out of these 70 measures, LPMNB model provide improved predictive performance for most of the measures (52). Second, whenever PMNB model performs better, the differences are very marginal. For example, the RMSE value estimated for sideswipe crashes form PMNB model is 4.171 (for estimation sample) while LPMNB model provides a RMSE value of 4.216. On the other hand, for rear-end, the RMSE value found from PMNB is 38.098 (for estimation sample) whereas for LPMNB, it is only 18.682. This clearly highlights the improved predictive power of the segmented model over its' unsegmented counterpart. Third, with respect to predictive log-likelihood, again LPMNB model performs better in most of the crash types (10 out of 14). The reader would note that, there is a difference between estimated and predicted log-likelihood. When we estimate our model considering correlation and unobserved effects, for every observation unit (TAZ), we get a joint probability and hence estimate the log-likelihood. However, in terms of prediction, our objective is to compare performance across crash types and not across the overall joint likelihood. Though PMNB model provides improved data fit in terms of model estimation (estimated log-likelihood, discussed in section 4.1.2), it falls short in prediction (based on predictive log-likelihood). In summary, the resulting goodness of fit measures and predictive log-likelihood offer by the LPMNB model clearly highlight its improved performance over the PMNB model.

6.6.2 Elasticity Effects

The parameters of the exogenous variables in Table 6.2 do not directly provide the exact magnitude of the effects of variables on the zonal level crash counts across different crash types. However, it might be possible that the effects (exact magnitude) of some attributes could differ considerably across the two frameworks. To evaluate this, we compute aggregate level elasticity effects for both PMNB and LPMNB models. In particular, we estimate the percentage change in the expected zonal level crash counts for every crash types in response to the increase of the explanatory variable by 10% (see Eluru and Bhat, 2007 for a discussion on the methodology for computing elasticities). For this purpose, we identify a subset of exogenous variables including proportion of arterial roads, variance of speed limit, proportion of roads over 55mph, proportion of roads with separated median and number of transit commuters in a zone. Further, for the LPMNB model, we estimate the aggregate level elasticities for the overall sample as well as for each segment separately to emphasize policy repercussions based on most critical contributory factors. For the overall sample, we took the segmentation probabilities into consideration. Table 6.5 provides a detailed documentation of the elasticities effect across the crash types for both PMNB and LPMNB models.

Several observations can be made based on the elasticity effects presented in Table 6.5. First, from the elasticity effects presented in table 6.5, we can clearly see some significant differences across two segments for some variables which highlights the importance of allowing for population heterogeneity in examining aggregate level crash counts across different crash types. For instance, due to the 10% increase in proportion of arterial roads, the expected mean of single vehicle crashes will reduce by 0.97% in the high risk segment whereas the effect is more significant in low risk segment with a reduction rate of 1.66%. Such differences can also be observed for other variables including variance of speed limit on rear-end, angular and sideswipe

crash counts; proportion of roads over 55mph on single vehicle crashes; and number of transit commuters on rear-end and non-motorized crashes. Second, interestingly, with respect to the variables present in both segments, TAZs assigned to low risk segment have higher elasticities relative to the high risk segment. Third, in terms of comparison across the two models adopted in the study, we found substantial differences in elasticities. For example, for the transit commuter variable, the PMNB model predicts an increase of 6.45% in expected mean for rear end crashes while LPMNB model predicts 4.39%. Similarly, with 10% increase in the proportion of road over 55mph speed, PMNB model predicts an 0.88% increase in expected mean for single vehicle crashes whereas we found an increase of about 1.16% from LPMNB model. This, it is evident that allowing for the population heterogeneity in both observed and unobserved factors provides more accurate representation of the variable impacts.

6.7 Summary

The current chapter extends the previous work of presented in chapter 4 by introducing the latent class version of the panel negative binomial (PNB) model to capture the potential variation in the impact of exogenous variables while also explicitly accommodating for unobserved heterogeneity through random parameters and error correlations. Further, we undertake a comparison exercise of the proposed LPNB model with its' traditional counterpart PMNB model proposed in chapter 4 in order to assess the importance of accounting for population heterogeneity in estimating zonal level crash frequency models.

Based on the statistical data fit, we can conclude that the segmented model is a preferred choice as long as the framework is estimated in a closed form structure (independent models that do not account for unobserved heterogeneity; no need for simulation). However, when we rely on

simulation for capturing the unobserved effects, the unsegmented model outperforms its' segmented counterparts. In an effort to assess the predictive performance of the estimated models, we compute several goodness fit of measures at disaggregate level including MPB, MAPE, RMSE and predictive log-likelihood for a discussion on estimating these measures). The resulting goodness of fit measures and predictive log-likelihood offer by the LPMNB model clearly highlight its improved performance over the PMNB model. Further, we compute aggregate level elasticity effects for both PMNB and LPMNB models to quantify whether the effect of variables significantly differs across the two frameworks. From the elasticity effects, we can clearly see substantial differences in elasticities which proves our hypothesis that allowing for the population heterogeneity in both observed and unobserved factors provides more accurate representation of the variable impacts.

Table 6.1 Segment Characteristics for LPMNB model

Crash Type	*Observed*	*Segment 1 (0.74)*	*Segment 2 (0.26)*	*Overall*
Rear-end	10.934	13.183	5.899	11.757
Angular	4.176	4.820	1.770	4.216
Sideswipe	2.687	2.791	1.799	2.616
Single Vehicle	2.390	2.489	1.986	2.361
Head-on	0.334	0.301	0.466	0.338
Non-motorized	0.712	0.869	0.239	0.755
Overall	3.539	4.075	2.027	3.674

Table 6.2 LPMNB Model Results

Segment Component				
Variables	*Segment 1*		*Segment 2*	
Constant	--	--	1.532	11.898
Number of intersections	--	--	-0.660	-14.394
Average outside shoulder width	--	--	-0.534	-21.139
Urban Area (acre)	--	--	0.897	13.163
Residential area	--	--	0.056	2.932
Crash Count Component				
Crash Specific Characteristic				
Rear-end	-0.171	-3.372	-3.298	-13.797
Angular	-1.654	-27.320	-4.363	-13.680
Sideswipe	-0.325	-6.400	-4.225	-11.594
Single Vehicle	-0.345	-8.048	-3.185	-14.410
Head-on	-2.882	-18.544	-4.227	-12.654
Non-motorized	-2.040	-15.908	-5.338	-14.331
Roadway Characteristic				
Proportion of arterial roads				
Rear-end+angular+NMT	0.166	4.933	--	--
All single vehicle	-0.260	-4.087	-0.472	-3.312
Signal Intensity				
Rear-end+sideswipe+NMT	--	--	2.350	3.479
Single vehicle	--	--	-1.760	-1.686
Variance of speed limit				
Rear-end+sideswipe+NMT	0.036	5.167	0.133	5.244
Road length over 55mph				
Rear-end+sideswipe	0.846	12.212	--	--
Angular	-2.058	-11.470	--	--
Standard Deviation	0.452	1.904	--	--
Single vehicle	0.846	12.212	0.753	2.921
Standard Deviation	--	--	0.930	3.149
Head-on	-2.103	-4.559	--	--
Non-motorized	-1.900	-6.312		
Roads with separated median				
Rear-end+angular+sideswipe	--	--	0.925	6.286
Head-on	--	--	-0.276	-1.138
Land Use Characteristic				
Office area (acre)				
Rear-end+sideswipe	0.195	20.947	--	--
Non-motorized	0.169	6.687	--	--
Built Environment Characteristic				

Number of restaurants				
Rear-end+sideswipe	0.192	13.919	--	--
Non-motorized	0.190	6.635		
Number of shopping centers				
Rear-end+sideswipe	0.034	2.676	--	--
Traffic Characteristic				
VMT				
Angular+sideswipe	0.147	45.205	--	--
Head-on	0.171	10.550	--	--
Non-motorized	0.102	8.365		
Truck VMT				
Rear-end	--	--	0.418	14.386
Single vehicle	--	--	0.554	21.272
Socio-economic Characteristic				
Non-motorist commuter				
Rear-end+sideswipe+NMT	0.076	3.924	--	--
Angular	0.170	8.790		
Transit commuter	--	--	--	--
Rear-end+ Non-motorized	0.217	11.883	0.576	8.584
Over Dispersion Parameter				
Rear-end	0.279	9.521	0.965	11.865
Angular	0.190	7.825	1.512	3.064
Sideswipe	0.284	8.294	0.965	11.865
Single Vehicle	0.726	17.746	0.115	1.554
Head-on	0.190	7.825	1.512	3.064
Non-motorized	0.279	9.521	0.965	11.865
Correlations				
Rear-end+NMT	0.679	23.659	--	--
Angular+sideswipe+single vehicle	0.840	34.284	1.245	7.913

Table 6.3 PMNB Model Results

Variables (np)	Rear-End		Angular		Sideswipe		Head-on		Single vehicle		Non-motorized	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Constant (6)	-0.930	-13.685	-1.623	-20.072	-2.590	-22.568	-3.499	-23.345	-0.747	-15.927	-3.016	-19.626
Roadway Characteristics												
Proportion of arterial roads (2)	0.158	4.732	0.158	4.732	--	--	--	--	-0.287	-5.422	0.158	4.732
Number of intersections (1)	--	--	0.359	14.033	--	--	0.359	14.033	--	--	0.359	14.033
Signal intensity (3)	0.716	3.347	--	--	-0.494	-1.828	--	--	-0.443	-2.693	0.716	3.347
Road length over 55mph (5)	0.422	5.047	-1.599	-8.872	0.422	5.047	0.866	6.410	-1.098	-4.575	-1.135	-4.580
Standard deviation	--	--	0.703	2.171	--	--	--	--	-0.509	-2.253	--	--
Variance of Speed (2)	0.038	5.079	0.038	5.079	0.070	5.021	--	--	--	--	--	--
Roads with separated median (2)	0.204	7.758	0.204	7.758	0.204	7.758	-0.108	-1.516	--	--	--	--
Average outside shoulder width (4)	-0.252	-7.489	-0.428	-9.693	-0.530	-10.186	-0.252	-7.489	-0.118	-3.221	--	--
Traffic Characteristic												
VMT (4)	--	--	0.1219	11.19	0.2392	18.689	0.1546	9.292	--	--	0.0182	1.800
Truck VMT (2)	0.1909	19.089	--	--	--	--	--	--	0.2708	34.334	--	--
Land-use attributes												
Urban area (4)	0.156	20.876	0.156	20.876	0.142	9.762	0.106	4.882	--	--	0.115	5.284
Office area (2)	0.163	18.620	--	--	0.163	18.620			--	--	0.164	6.635
Residential area (1)	--	--	--	--	-0.077	-7.218	-0.077	-7.218	--	--	--	--
Built environment characteristic												
No. of restaurants (3)	0.3082	13.34	--	--	0.1091	4.297	--	--	--	--	0.2568	9.068
No. of shopping centers (1)	0.029	2.029	--	--	0.029	2.029	--	--	--	--	--	--

Socio-demographic characteristics												
Non-motorists (3)	0.070	3.408	0.148	6.956	0.164	6.505	--	--	--	--	0.070	3.408
Transit users (1)	0.239	13.596	--	--	--	--	--	--	--	--	0.239	13.596
Over dispersion (6)	0.396	31.904	0.384	14.952	0.396	31.904	0.384	14.952	0.700	22.059	0.396	31.904
Unobserved Effects												
Correlation 1 (1)	0.741	33.753	--	--	--	--	--	--	--	--	0.741	33.753
Correlation 2 (1)	--	--	0.936	40.216	0.936	40.216	0.936	40.216	--	--	--	--

*np= number of parameters estimated for each variable from a possible set of six (six crash types)
--= attribute insignificant at 90% confidence level

Table 6.4 Predictive Performance Measure of Two Models (PMNB and LPMNB)

Dataset	Crash Type	MPB		MAD		MAPE		RMSE		Predicted Log-likelihood	
		PMNB*	LPMNB	PMNB	LPMNB	PMNB	LPMNB	PMNB	LPMNB	PMNB	LPMNB
In-Sample Measures (3,815 TAZs)	*Rear-end*	-0.312	-0.823	8.519	7.741	3.077	2.980	38.098	18.682	-11113.5	-11087.6
	Angular	1.148	-0.040	3.126	3.445	1.892	1.010	5.834	5.769	-8645.03	-8635.75
	Sideswipe	0.868	0.071	2.028	2.210	0.861	0.697	4.171	4.216	-6744.93	-6747.99
	Single Vehicle	0.062	0.029	1.809	1.866	1.547	0.333	2.903	3.070	-7098.68	-7074.95
	Head-on	0.107	-0.004	0.429	0.494	0.089	0.153	0.990	1.001	-2584.61	-2596.1
	Non-motorized	0.077	-0.043	0.680	0.699	0.067	0.133	1.360	1.203	-3761.8	-3756.02
	Overall	1.950	-0.809	16.590	16.454	7.533	5.306	38.912	20.296	-39948.5	-39898.4
Hold-out sample Measures (932 TAZs)	*Rear-end*	-0.615	1.833	19.694	14.999	2.144	4.161	74.047	34.174	-3783.87	-3758.93
	Angular	4.660	3.311	6.046	5.856	3.274	0.925	10.048	9.627	-3086.49	-3072.68
	Sideswipe	3.287	2.167	4.173	4.079	2.241	0.616	7.292	7.214	-2628.16	-2662.63
	Single Vehicle	1.195	1.261	2.513	2.594	1.661	0.747	3.979	4.156	-2271.89	-2259.24
	Head-on	0.151	0.053	0.515	0.555	0.038	0.101	0.769	0.768	-828.111	-833.142
	Non-motorized	0.177	-0.010	1.186	1.172	0.402	0.085	2.308	1.949	-1405.31	-1402.91
	Overall	8.855	8.615	34.129	29.254	9.760	6.635	75.225	36.527	-14003.8	-13989.5

Table 6.5 Elasticity Effects Across Two Models (PMNB and LPMNB)

Variables	Models		Crash Types					
			Rear-end	Angular	Sideswipe	Single Vehicle	Head-on	Non-motorized
Arterial Roads	LPMNB	Segment 1	0.800	0.735	0.000	-0.974	0.000	0.787
		Segment 2	0.000	0.000	0.000	-1.655	0.000	0.000
		Overall	0.736	0.704	0.000	-1.110	0.000	0.752
	PMNB		0.859	0.753	0.000	-1.093	0.000	0.816
Variance	LPMNB	Segment 1	1.178	1.061	1.171	0.000	0.000	0.000
		Segment 2	5.036	5.056	5.032	0.000	0.000	0.000
		Overall	1.556	1.315	1.539	0.000	0.000	0.000
	PMNB		1.343	1.331	1.409	0.000	0.000	0.000
Speed ≥55mph	LPMNB	Segment 1	0.824	-1.137	0.886	1.115	-1.184	-0.906
		Segment 2	0.000	0.000	0.000	1.349	0.000	0.000
		Overall	0.640	-0.954	0.684	1.163	-0.826	-0.782
	PMNB		0.246	-0.769	0.322	0.887	-0.615	-0.465
Road with Median	LPMNB	Segment 1	0.000	0.000	0.000	0.000	0.000	0.000
		Segment 2	7.754	7.776	7.793	0.000	-1.703	0.000
		Overall	0.741	0.443	0.716	0.000	-0.342	0.000
	PMNB		1.623	1.469	1.590	0.000	-0.723	0.000
Transit Commuter	LPMNB	Segment 1	3.404	0.000	0.000	0.000	0.000	3.352
		Segment 2	14.668	0.000	0.000	0.000	0.000	14.586
		Overall	4.387	0.000	0.000	0.000	0.000	4.021
	PMNB		6.450	0.000	0.000	0.000	0.000	6.310

CHAPTER 7: CONCLUSIONS

Road traffic crash related morbidity and mortality is acknowledged to be a global challenge. In reducing the burden of such unavoidable incidents, safety researchers are investigating approaches for crash occurrence reduction and crash consequence mitigation. A major analytical tool employed for examining the critical factors influencing crash occurrence include the econometric crash frequency models. The traditional modeling framework for crash frequency analysis is the univariate frequency model such as Poisson and Negative binomial model. However, these approaches do not account for the common unobserved factors affecting the multiple dependent variables for the same observational unit. Recognizing this drawback, several research efforts have developed frameworks that accommodate for the influence of these common unobserved factors referred to as multivariate modeling approaches. However, there are still several methodological challenges associated with such existing models suggesting continual needs to develop advanced econometric framework to address these gaps.

In this context, the current dissertation contributes towards addressing the methodological challenges in crash frequency analysis for analyzing multiple crash frequency variables for the same study unit by proposing advanced econometric approaches. The first objective of the dissertation contributes to safety literature by conducting a comparison exercise between the two major streams of multivariate approaches - (1) simulation-based approach and (2) analytical closed form approach - for analyzing the crash counts considering different crash types. In the second objective of the dissertation, we propose an alternative and mathematically simpler approach for analyzing multiple crash frequency variables for the same study unit by recasting a multivariate distributional problem as a repeated measures univariate problem. The recasting allows us to

estimate parsimonious model systems thus improving parameter estimation efficiency. The third objective of the dissertation contributes to burgeoning econometric and safety literature by developing a joint modeling approach that can accommodate for several dependent variables within a parsimonious structure. By recasting the analysis levels for dependent variables, the proposed approach allows for flexible consideration of crashes by type and severity within a single framework. The final objective of the dissertation contributes to literature on crash frequency analysis by accommodating population heterogeneity in the impact of exogenous variables. The empirical analysis is based on traffic analysis zone (TAZ) level crash count data for both motorized and non-motorized crashes from Central Florida for the year 2016. A comprehensive set of exogenous variables including roadway, built environment, land-use, traffic, socio-demographic and spatial spillover characteristics are considered for the analysis.

The proposed contributions are organized along four parts. The rest of the chapter is organized as follows. Section 7.1 through 7.4 discusses the substantive and methodological contributions of the dissertation for each objective examined in the dissertation. Section 7.6 concludes the dissertation by discussing the limitations of the dissertation and offering directions for future research.

7.1 Exploration of Analytical, Simulation and Combined Model Structures

The most common approach employed to address the correlation across multiple frequency dependent variables in existing safety literature is the development of multivariate frameworks. These multivariate approaches can broadly be classified along two major streams: (1) simulation-based approaches and (2) analytically closed-form based approaches. The main difference between these two streams lies in how the dependency across dimensions is captured. In the simulation-

based models, probability computation requires integrating the probability function over the error term distribution and the exact computation is dependent on the distributional assumption due to the inherently unobserved nature of the error term. Thus, the accuracy of the simulation-based approach is affected by number of dimensions as well as number of draws considered for the function evaluation. On the other hand, in the closed-form regime, the propensity equations for frequency dimensions are tied together by analytical multivariate distributional assumptions. Though the likelihood function is complicated in the closed-form approach, but once programmed, these frameworks are less prone to error.

In our research, we compare the performance of the simulation-based framework with closed-form copula-based frameworks. In addition, we build on the closed-form copula based frameworks to incorporate unobserved heterogeneity associated with variable impacts on crash types (random parameters). The proposed model system is compared with the simulation based and analytical multivariate models. The comparison exercise is undertaken with the univariate models following negative binomial model structure. Within the copula framework, we estimate models for four copula structures: (1) Frank, (2) Gumbel, (3) Clayton and (4) Joe which cover a wide range of dependency structures, including radial symmetry and asymmetry, and asymptotic tail independence and dependence. The empirical analysis is based on the traffic analysis zone (TAZ) level crash count data for both motorized and non-motorized crashes from Central Florida for the year 2016. The models were estimated employing a comprehensive set exogenous variable including roadway, built environment, land-use, traffic, socio-demographic characteristics and spatial spillover effects. The model fit measures clearly highlight that the RPCC (random parameter Clayton copula) model outperforms simulation-based RPMNB model. The comparison exercise was further augmented by generating a host of comparison metrics for both estimation

sample and hold-out sample. In an effort to further assess the predictive performance of the estimated models, an in-depth comparison for different count events across different crash types and correct classification analysis are carried out. The estimated results further reinforce the superiority of the RPCC-based multivariate approach. The RPCC based copula model is also employed to generate hot and cold zone categorization of TAZs in the Central Florida region to identify potential vulnerable zones by crash type.

The proposed model results offer insights on important variables affecting crash frequency by crash types (road user and location for the current study context). The macro-level model outcomes can be used to devise safety-conscious decision support tools to facilitate a proactive approach in assessing medium and long-term policy-based countermeasures. Moreover, with the spatial illustration, high risk zones for every crash type can be easily identified and thus help the planners in enhancing safety for these high crash risk zones.

7.2 Panel Mixed Approach to Modeling Crash Frequency by Crash Types

The most common approach employed to address correlation across multiple crash frequency dependent variables in safety literature is the development of simulation-based multivariate frameworks. However, with higher dimensions, the multivariate model estimation infrastructure can get computationally demanding in terms of the number of observed and unobserved parameters to estimate. In this context, our proposed research attempts to contribute to simulation-based multivariate approaches by altering how the multiple dependent variables are analyzed. Specifically, instead of considering the crash frequency by crash type as a multivariate distribution, we represent it as a repeated measures of crash frequency while recognizing that each repetition represents a crash type specific to a zone. Thus, in this process we cast a multivariate distribution

as a univariate distribution with repeated measures. The recasting allows us to estimate parsimonious model systems as well as simplify the specification process. This simplification leading to parsimonious specification can reduce the computational time for estimating parameters associated with unobserved factors. To the best of authors' knowledge, this study is the first of its kind to simplify current modeling infrastructure for multivariate analysis in safety literature.

In our current research effort, a simple random parameter based univariate model code was employed to analyze zonal level crash counts for different crash types including rear-end, angular, sideswipe, all single vehicle, other multiple vehicle and non-motorized crashes. The empirical analysis was based on the traffic analysis zone (TAZ) level crash count data from Central Florida for the year 2016. A host of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics were considered in the current research effort. A comprehensive comparison of the proposed model with the most commonly used multivariate negative binomial (NB) model was conducted. The comparison exercise based on the BIC value clearly highlighted the superiority of the proposed approach over the traditional multivariate formulation in terms of data fit. The comparison exercise was further augmented by generating several predictive measures for both estimation and holdout samples. Based on the resulting fit measures, the study concludes that the proposed formulation has offered equivalent predictions relative to the most traditional multivariate NB model even though there is a significant difference in the number of parameters within these two frameworks (61 vs 92). Further, we compute aggregate level elasticity effects for both PMNB and RPMNB models to quantify whether the effect of variables significantly differs across the two frameworks. For this purpose, we identify a subset of exogenous variable including proportion of arterial roads, length of divided roads, proportion of roads over 55mph, institutional areas and number of non-motorist commuters. The

elasticity results clearly indicate that for most of the variables, the effects are quite similar for both models across different crash types. However, for some variables, we found some significant and substantial differences in the elasticity effects across the two frameworks for some crash types. Such differences could be attributed to the non-linearity embedded within the two model structures estimated with similar data fit.

The current research effort contributes to literature on crash frequency analysis by suggesting an alternative and mathematically simpler approach for analyzing multiple crash frequency variables for the same study unit. Specifically, the proposed framework while simplifying the model estimation process, allows for parsimonious specification without compromising the model explanatory power and provides similar performance (predictions) as the currently employed multivariate NB model. In conclusion, the aim of the proposed scheme is to augment the inventory of crash frequency models with an alternative formulation and serves as a viable approach to reduce the parameter explosion that is common within a multivariate NB model with large number of dependent variable dimensions.

7.3 Econometric Approach for Modeling Crash Counts by Crash Type and Severity

Despite the distinct injury severity profile, there is limited adoption of research modeling severity frequency or proportion considering different crash types. The main challenge is with the number of dependent variables as accommodating unobserved heterogeneity for such large number of dimensions is substantially burdensome. The probability evaluation with high dimensional integrals is potentially affected by several challenges including - requirements of generating high dimensionality of random numbers, empirical identification issues due to relatively flat objective functions in larger dimensions and longer computational run times. In this context, the proposed

research contributes to burgeoning econometric and safety literature by developing a joint modeling approach that can accommodate for a large number of dependent variables (considering crash types and severities) within a parsimonious structure. With respect to crash type specific component, instead of considering the crash frequency by crash type as a traditional multivariate distribution, we recasted it as a repeated measures of crash frequency while recognizing that each repetition represents a crash type specific to a zone. At the same time, for the severity component, as opposed to modeling the count events, count proportions by different severity level for a study unit were examined. Finally, we developed a joint model to tie the two components in a single integrated framework while accommodating unobserved heterogeneity across and within the two components (crash frequency and crash severity proportions by types).

In our current research effort, we employed a Panel mixed Negative Binomial- Generalized Ordered Probit Fractional Spilt (PM-NB-GOPFS) model where the first component (NB) accommodated for crash frequency by crash type and the later component (GOPFS) studied the fraction of severity outcome for different crash types. The empirical analysis was conducted using the zonal level crash count data for the year 2016 from Central Florida while considering a comprehensive set of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics. The empirical analysis involved a series of model estimations including: 1) Independent NB-GOPFS model; 2) Panel NB-GOPFS model without unobserved component parameters; and 3) Joint Panel NB-GOPFS model with unobserved heterogeneity. The comparison exercise, based on the Bayesian Information Criterion (BIC)value highlighted the superiority of the proposed framework that accounts for penalty for additional parameters (model 2 and 3) and within the proposed approach, the model considering unobserved heterogeneity (model 3) outperformed its' counterpart (model 2).

The analysis was further augmented by undertaking a prediction exercise using the final model parameter estimates. One of the major advantage of the proposed framework is that in a single econometric framework, we can predict severaldimensions including total crash counts, total crash counts by crash types, crash proportions for each severity level, crash counts for each severity level and finally, proportions and counts of crashes for each crash type by severity. In evaluating the predictive performance, we compute the errors (MAD and MAPE) across all the aforementioned dimensions. Specifically, we compute MAD at a disaggregate level by generating measures at the study unit level (TAZ). On the other hand, MAPE measures are generated at an aggregate level where we estimate the number and proportion of crashes for corresponding dimension (crash types, severities) and predict the TAZ shares for different count and proportion alternatives and compared it with the observed shares. The prediction results clearly indicated that the joint model for crash counts and severity proportions by crash type performed adequately (for both in-sample and validation samples) under consideration.

In summary, the current study contributes to safety literature both methodologically and empirically. Methodologically, we developed a joint framework analysing 24 dependent variables (6*4 from 6 crash types and 4 severities). Empirically, by increasing the dimensionality of the dependent variable, the proposed approach allows for flexible consideration of crashes by type and severity within a single framework. Further, the proposed model results offer insights on important variables affecting crash frequency and severity for different crash types. Such macro-level model outcomes can be used to devise safety-conscious decision support tools to facilitate proactive approach in assessing medium and long-term policy-based countermeasures.

7.4 Accommodating Population Heterogeneity Within A Panel Model Framework

The literature clearly highlights the prevalence of multivariate model frameworks in safety literature. However, there are some major challenges associated with the existing multivariate approach in estimating observed and unobserved effects. To that extent, our current study contributes to crash frequency literature both methodologically and empirically by estimating a latent segmentation-based Panel Negative Binomial (LPNB) to study the zonal level crash counts across different crash types. The fourth objective of the dissertation extends the previous work presented in objective two by introducing the latent class version of the panel negative binomial (PNB) model to capture the potential variation in the impact of exogenous variables while also explicitly accommodating for unobserved heterogeneity through random parameters and error correlations. The latent segmentation scheme is appealing for multiple reasons including: (1) it ensures that the parameters are estimated employing the full sample for each segment while employing all data points for model estimation; (2) provides valuable insights on how the exogenous variables affect segmentation; and (3) the probabilistic assignment explicitly acknowledges the role played by unobserved factors in moderating the impact of observed exogenous variables. Further, we undertake a comparison exercise of the proposed LPNB model with its' traditional counterpart PMNB model proposed in chapter 4 (objective two) in order to assess the importance of accounting for population heterogeneity in estimating zonal level crash frequency models.

Based on the statistical data fit, we can conclude that the segmented model is a preferred choice as long as the framework is estimated in a closed form structure (independent models that do not account for unobserved heterogeneity; no need for simulation). However, when we rely on simulation for capturing the unobserved effects, the unsegmented model outperforms its'

segmented counterparts. In an effort to assess the predictive performance of the estimated models, we compute several goodness fit of measures at disaggregate level including MPB (Mean prediction bias), MAD (mean absolute deviation), MAPE (mean absolute percentage error), RMSE (Root mean square error) and predictive log-likelihood for a discussion on estimating these measures). Specifically, we employ these measure on two datasets: 1) in-sample dataset: for the records used in the model estimation (sample size = 3,815 TAZs) and 2) holdout sample: records that are set aside for validation analysis (sample size = 932 TAZs). The resulting goodness of fit measures and predictive log-likelihood offer by the LPMNB model clearly highlight its improved performance over the PMNB model. Further, we compute aggregate level elasticity effects for both PMNB and LPMNB models to quantify whether the effect of variables significantly differs across the two frameworks. For this purpose, we identify a subset of exogenous variables including proportion of arterial roads, variance of speed limit, proportion of roads over 55mph, proportion of roads with separated median and number of transit commuters in a zone. Further, for the LPMNB model, we estimate the aggregate level elasticities for the overall sample as well as for each segment separately to emphasize policy repercussions based on most critical contributory factors. From the elasticity effects, we can clearly see some significant differences across two segments for some variables which highlights the importance of allowing for population heterogeneity in examining aggregate level crash counts across different crash types. In terms of comparison across the two models adopted in the study, we found substantial differences in elasticities which proves our hypothesis that allowing for the population heterogeneity in both observed and unobserved factors provides more accurate representation of the variable impacts.

In summary, the newly formulated model will allow us to partition the TAZs into segments based on their attributes and estimate the influence of exogenous variables on crash counts of

different crash types. From *methodological* perspective, the current research makes a threefold contribution to literature on crash frequency analysis: First, the recasting allows us to estimate a parsimonious model system and also reduce the computational time for estimating parameters associated with unobserved factors. Second, by introducing the latent class version of the PNB model, we allow for both observed and unobserved heterogeneity thus relaxing the homogeneity assumption of the traditional count models. Third, we allow for a flexible segment membership function and test for the presence of multiple segments in the model estimation. *Empirically*, the research contributes to our understanding of analyzing zonal level crashes for both motorized and non-motorized road user group while considering different crash types within the motorized category including rear-end, angular, sideswipe, single vehicle and head-on crashes.

7.5 Contribution of The Dissertation

The current dissertation contributes substantially towards methodological gaps in the state of art for analyzing multiple crash frequency variables along six directions: (1) considering crashes from both road user groups (motorists and non-motorists) (2) consider different crash types within the motorized crashes; (3) undertake a comparison exercise between the analytical and simulation based multivariate model for capturing unobserved heterogeneity; (4) propose a new alternative simpler model for analyzing multiple dependent variables; (5) propose a new econometric approach for analyzing a large number of dependent count variables and by increasing the dimensionality of the dependent variable, the proposed approach allows for flexible consideration of crashes by type and severity within a single framework; and (6) a latent segmentation framework to capture the potential variation in the impact of explanatory variables at the zonal level crash counts by crash type. In addition to making the aforementioned methodological contributions, the

dissertation also makes a substantial empirical contribution to the existing safety literature. All the models developed at zonal level can be used to devise safety-conscious decision support tools to facilitate proactive approach in assessing medium and long-term policy-based countermeasures.

7.6 Limitations and Future Research

To be sure, the dissertation is not without limitations. In our study, left-turn and right-turn crashes were considered in the same category due to sample size restrictions despite differences in crash mechanisms of these two categories. In future research efforts, it might be useful to consider them separately given that the crash mechanisms for these crash types could be potentially different. Moreover, given the inherent aggregation of the dataset, it would be beneficial to accommodate for the presence of spatial unobserved effects as well. Further, it might be interesting to explore the transferability of models developed for crash count by estimating similar models for multiple spatial units and several years. Finally, it would be an interesting research exercise to evaluate if the findings are confirmed for other count model kernels (such a log-normal frameworks).

REFERENCES

Abdel-Aty, M., Keller, J., Brady, P.A., 2005. Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression. Transp. Res. Rec. 1908 1908 , 37–45. doi:10.3141/1908-05

Abdelwahab, H.T., Abdel-Aty, M.A., 2002. Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transp. Res. Rec. 1784 1784 , 115–125. doi:10.3141/1784-15

Aguero-Valverde, J., 2013. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. Accid. Anal. Prev. 59, 365–373. doi:10.1016/j.aap.2013.06.014

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accid. Anal. Prev. 38 3 , 618–625. doi:10.1016/j.aap.2005.12.006

Alarifi, S.A., Abdel-Aty, M., Lee, J., 2018. A Bayesian multivariate hierarchical spatial joint model for predicting crash counts by crash type at intersections and segments along corridors. Accid. Anal. Prev. 119, 263–273. doi:10.1016/j.aap.2018.07.026

Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Anal. Methods Accid. Res. 11, 17–32. doi:10.1016/j.amar.2016.06.001

Anastasopoulos, P.C., Shankar, V.N., Haddock, J.E., Mannering, F.L., 2012. A multivariate tobit analysis of highway accident-injury-severity rates. Accid. Anal. Prev. 45, 110–119. doi:10.1016/j.aap.2011.11.006

Aptech, 2015. Aptech [WWW Document]. Aptech 2015, Aptech Syst. Inc, accessed from http//www.aptech.com/ Sept. 19th 2015. URL http://www.aptech.com/ (accessed 9.19.15).

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Anal. Methods Accid. Res. 9, 1–15. doi:10.1016/j.amar.2015.11.002

Barua, S., El-Basyouny, K., Islam, M.T., 2014. A Full Bayesian multivariate count data model of collision severity with spatial correlation. Anal. Methods Accid. Res. 3–4, 28–43. doi:10.1016/j.amar.2014.09.001

Bhat, C.R., 2014. The composite marginal likelihood (CML) inference approach with applications to discrete and mixed dependent variable models. Found. Trends Econom. 7 1 , 1–117. doi:10.1561/0800000022

Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. Transp. Res. Part B Methodol. 45 7 , 923–939. doi:10.1016/j.trb.2011.04.005

Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transp. Res. Part B Methodol. 35 7 , 677–693. doi:10.1016/S0191-2615(00)00014-X

Bhat, C.R., Astroza, S., Lavieri, P.S., 2017. A new spatial and flexible multivariate random-coefficients model for the analysis of pedestrian injury counts by severity level. Anal. Methods Accid. Res. 16, 1–22. doi:10.1016/j.amar.2017.05.001

Bhat, C.R., Born, K., Sidharthan, R., Bhat, P.C., 2014. A count data model with endogenous covariates: Formulation and application to roadway crash frequency at intersections. Anal. Methods Accid. Res. 1, 53–71. doi:10.1016/j.amar.2013.10.001

Bhat, C.R., Eluru, N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. Transp. Res. Part B Methodol. 43 7 , 749–765.

doi:10.1016/j.trb.2009.02.001

Bhowmik, T., Rahman, M., Yasmin, S., Eluru, N., 2019a. Alternative Model Structures for Multivariate Crash Frequency Analysis: Comparing Simulation-based Multivariate Model with Copula-based Multivariate Model.

Bhowmik, T., Yasmin, S., Eluru, N., 2019b. Do we need multivariate modeling approaches to model crash frequency by crash types? A panel mixed approach to modeling crash frequency by crash types. Anal. Methods Accid. Res. 24. doi:10.1016/j.amar.2019.100107

Bhowmik, T., Yasmin, S., Eluru, N., 2019c. A multilevel generalized ordered probit fractional split model for analyzing vehicle speed. Anal. Methods Accid. Res. 21, 13–31. doi:10.1016/j.amar.2018.12.001

Bhowmik, T., Yasmin, S., Eluru, N., 2018. A joint econometric approach for modeling crash counts by collision type. Anal. Methods Accid. Res. 19 3 , 16–32. doi:10.1016/j.amar.2018.06.001

Boulieri, A., Liverani, S., de Hoogh, K., Blangiardo, M., 2017. A space–time multivariate Bayesian model to analyse road traffic accidents by severity. J. R. Stat. Soc. Ser. A Stat. Soc. 180 1 , 119–139. doi:10.1111/rssa.12178

Cai, Q., Lee, J., Eluru, N., Abdel-Aty, M., 2016. Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. Accid. Anal. Prev. 93, 14–22. doi:10.1016/j.aap.2016.04.018

Cameron, A.C., Li, T., Trivedi, P.K., Zimmer, D.M., 2004. Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts*. Econom. J. 7 2 , 566–584. doi:10.1111/j.1368-423x.2004.00144.x

Chen, S., Saeed, T.U., Labi, S., 2017. Impact of road-surface condition on rural highway safety:

A multivariate random parameters negative binomial approach. Anal. Methods Accid. Res. 16, 75–89. doi:10.1016/j.amar.2017.09.001

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X., Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accid. Anal. Prev. 99, 330–341. doi:10.1016/j.aap.2016.11.022

Cheng, W., Gill, G.S., Ensch, J.L., Kwong, J., Jia, X., 2018. Multimodal crash frequency modeling: Multivariate space-time models with alternate spatiotemporal interactions. Accid. Anal. Prev. 113, 159–170. doi:10.1016/j.aap.2018.01.034

Chiou, Y.C., Fu, C., 2015. Modeling crash frequency and severity with spatiotemporal dependence. Anal. Methods Accid. Res. 5–6, 43–58. doi:10.1016/j.amar.2015.03.002

Chiou, Y.C., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accid. Anal. Prev. 50, 73–82. doi:10.1016/j.aap.2012.03.030

Chiou, Y.C., Fu, C., Hsieh, C.W., 2014. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. Anal. Methods Accid. Res. 2, 1–11. doi:10.1016/j.amar.2013.12.001

Dey, B.K., Anowar, S., Eluru, N., Hatzopoulou, M., 2018. Accommodating exogenous variable and decision rule heterogeneity in discrete choice models: Application to bicyclist route choice. PLoS One 13 11 . doi:10.1371/journal.pone.0208309

Dong, C., Clarke, D.B., Nambisan, S.S., Huang, B., 2016. Analyzing injury crashes using random-parameter bivariate regression models. Transp. A Transp. Sci. 12 9 , 794–810. doi:10.1080/23249935.2016.1177134

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters

zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accid. Anal. Prev. 70, 320–329. doi:10.1016/j.aap.2014.04.018

El-Basyouny, K., Barua, S., Islam, M.T., Li, R., 2014. Assessing the Effect of Weather States on Crash Severity and Type by Use of Full Bayesian Multivariate Safety Models. Transp. Res. Rec. J. Transp. Res. Board 2432 1 , 65–73. doi:10.3141/2432-08

Eluru, N., Bagheri, M., Miranda-Moreno, L.F., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. Accid. Anal. Prev. 47, 119–127. doi:10.1016/j.aap.2012.01.027

Eluru, N., Bhat, C.R., 2007. A joint econometric analysis of seat belt use and crash-related injury severity. Accid. Anal. Prev. 39 5 , 1037–1049. doi:10.1016/j.aap.2007.02.001

Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accid. Anal. Prev. 40 3 , 1033–1054. doi:10.1016/j.aap.2007.11.010

Eluru, N., Chakour, V., Chamberlain, M., Miranda-Moreno, L.F., 2013. Modeling vehicle operating speed on urban roads in Montreal: A panel mixed ordered probit fractional split model. Accid. Anal. Prev. 59, 125–134. doi:10.1016/j.aap.2013.05.016

Eluru, N., Paleti, R., Pendyala, R.M., Bhat, C.R., 2010. Modeling Injury Severity of Multiple Occupants of Vehicles. Transp. Res. Rec. J. Transp. Res. Board 2165 1 , 1–11. doi:10.3141/2165-01

Eluru, N., Yasmin, S., 2015. A note on generalized ordered outcome models. Anal. Methods Accid. Res. 8, 1–6. doi:10.1016/j.amar.2015.04.002

Fountas, G., Anastasopoulos, P.C., 2017. A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. Anal. Methods Accid. Res. 15,

1–16. doi:10.1016/j.amar.2017.03.002

Geedipally, S.R., Patil, S., Lord, D., 2010. Examination of methods to estimate crash counts by collision type. Transp. Res. Rec. 2165 2165 , 12–20. doi:10.3141/2165-02

GHSA, 2009. Governors Highway Safety Association (GHSA), 2009. Toward Zero Deaths: Every Life Counts, GHSA, Washington, D.C.

Heydari, S., Fu, L., Miranda-Moreno, L.F., Joseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. Anal. Methods Accid. Res. 13, 16–27. doi:10.1016/j.amar.2016.12.002

Hosseinpour, M., Yahaya, A.S., Sadullah, A.F., 2014. Exploring the effects of roadway characteristics on the frequency and severity of head-on crashes: Case studies from Malaysian Federal Roads. Accid. Anal. Prev. 62, 209–222. doi:10.1016/j.aap.2013.10.001

Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. Anal. Methods Accid. Res. 14, 10–21. doi:10.1016/j.amar.2017.01.001

Jonathan, A.V., Wu, K.F., Donnell, E.T., 2016. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accid. Anal. Prev. 87, 8–16. doi:10.1016/j.aap.2015.11.006

Lee, J., Abdel-Aty, M., Jiang, X., 2015. Multivariate crash modeling for motor vehicle and non-motorized modesat the macroscopic level. Accid. Anal. Prev. 78, 146–154. doi:10.1016/j.aap.2015.03.003

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., Cai, Q., 2018. Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects. Accid. Anal. Prev. 111, 12–22. doi:10.1016/j.aap.2017.11.017

Li, Z., Wang, W., Liu, P., Bigham, J.M., Ragland, D.R., 2013. Using Geographically Weighted Poisson Regression for county-level crash modeling in California. Saf. Sci. 58, 89–97. doi:10.1016/j.ssci.2013.04.005

Liu, C., Sharma, A., 2018. Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. Anal. Methods Accid. Res. 17, 14–31. doi:10.1016/j.amar.2018.02.001

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44 5 , 291–305. doi:10.1016/j.tra.2010.02.001

Ma, X., Chen, S., Chen, F., 2017. Multivariate space-time modeling of crash frequencies by injury severity levels. Anal. Methods Accid. Res. 15, 29–40. doi:10.1016/j.amar.2017.06.001

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16. doi:10.1016/j.amar.2016.04.001

Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accid. Anal. Prev. 40 1 , 260–266. doi:10.1016/j.aap.2007.06.006

Mothafer, G.I.M.A., Yamamoto, T., Shankar, V.N., 2016. Evaluating crash type covariances and roadway geometric marginal effects using the multivariate Poisson gamma mixture model. Anal. Methods Accid. Res. 9, 16–26. doi:10.1016/j.amar.2015.11.001

Narayanamoorthy, S., Paleti, R., Bhat, C.R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transp. Res. Part B Methodol. 55, 245–264. doi:10.1016/j.trb.2013.07.004

Nashad, T., Yasmin, S., Eluru, N., Lee, J., Abdel-Aty, M.A., 2016. Joint modeling of pedestrian and bicycle crashes: Copula-based approach. Transp. Res. Rec. 2601 2601 , 119–127. doi:10.3141/2601-14

National Highway Traffic Safety Administration (NHTSA), 2013. Traffic safety facts 2011 data - Pedestrians. Ann. Emerg. Med. 62 6 , 612. doi:10.1016/j.annemergmed.2013.09.018

NHTSA, 2018 [WWW Document], n.d. URL https://www.usatoday.com/story/money/cars/2019/06/17/car-crashes-36-750-people-were-killed-us-2018-nhtsa-estimates/1478103001/ (accessed 2.17.20).

Pai, C.W., Saleh, W., 2008. Modelling motorcyclist injury severity by various crash types at T-junctions in the UK. Saf. Sci. 46 8 , 1234–1247. doi:10.1016/j.ssci.2007.07.005

Papke, L.E., 1996. Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. J. Appl. Econom. 11 6 , 619–632. doi:10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1

Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41 4 , 683–691. doi:10.1016/j.aap.2009.03.007

Park, B.J., Lord, D., Hart, J.D., 2010. Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis. Accid. Anal. Prev. 42 2 , 741–749. doi:10.1016/j.aap.2009.11.002

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. Accid. Anal. Prev. 40 4 , 1486–1497. doi:10.1016/j.aap.2008.03.009

Rahman Shaon, M.R., Qin, X., Afghari, A.P., Washington, S., Haque, M.M., 2019. Incorporating behavioral variables into crash count prediction by severity: A multivariate multiple risk

source approach. Accid. Anal. Prev. 129, 277–288. doi:10.1016/j.aap.2019.05.010

Reynolds, C.C.O., Harris, M.A., Teschke, K., Cripton, P.A., Winters, M., 2009. The impact of transportation infrastructure on bicycling injuries and crashes: A review of the literature. Environ. Heal. A Glob. Access Sci. Source 8 1 , 2–4. doi:10.1186/1476-069X-8-47

Sener, I.N., Eluru, N., Bhat, C.R., 2010. On jointly analyzing the physical activity participation levels of individuals in a family unit using a multivariate copula framework. J. Choice Model. 3 3 , 1–38. doi:10.1016/S1755-5345(13)70012-5

Serhiyenko, V., Mamun, S.A., Ivan, J.N., Ravishanker, N., 2016. Fast Bayesian inference for modeling multivariate crash counts. Anal. Methods Accid. Res. 9, 44–53. doi:10.1016/j.amar.2016.02.002

Soulez, F., Denis, L., Fournier, C., Thiébaut, É., Goepfert, C., 2007. Inverse-problem approach for particle digital holography: accurate location based on local optimization. J. Opt. Soc. Am. A 24 4 , 1164. doi:10.1364/josaa.24.001164

U.S. NHTSA, 2017. 2016 Fatal Motor Vehicle Crashes: Overview. Traffic Saf. Facts Res. Note (DOT HS 812456) October , 1–9. doi:DOT HS 812 456

Wang, K., Ivan, J.N., Ravishanker, N., Jackson, E., 2017. Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways. Accid. Anal. Prev. 99, 6–19. doi:10.1016/j.aap.2016.11.006

Wang, K., Yasmin, S., Konduri, K.C., Eluru, N., Ivan, J.N., 2015. Copula-based joint model of injury severity and vehicle damage in two-vehicle crashes. Transp. Res. Rec. 2514, 158–166. doi:10.3141/2514-17

Wang, X., Abdel-Aty, M., 2008. Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. Accid. Anal. Prev. 40 5 , 1674–1682.

doi:10.1016/j.aap.2008.06.001

Wang, Y., Kockelman, K.M., 2013. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accid. Anal. Prev. 60, 71–84. doi:10.1016/j.aap.2013.07.030

Wedel, M., Desarbo, W.S., Bult, J.R., Ramaswamy, V., 1993. A latent class poisson regression model for heterogeneous count data. J. Appl. Econom. 8 4 , 397–411. doi:10.1002/jae.3950080407

World Health Organization, 2015. Global status report on road safety. Inj. Prev. 318. doi:http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

Xin, C., Guo, R., Wang, Z., Lu, Q., Lin, P.S., 2017. The effects of neighborhood characteristics and the built environment on pedestrian injury severity: A random parameters generalized ordered probability model with heterogeneity in means and variances. Anal. Methods Accid. Res. 16, 117–132. doi:10.1016/j.amar.2017.10.001

Xuesong, W., Abdel-Aty, M., Brady, P.A., 2006. Crash estimation at signalized intersections significant factors and temporal effect. Transp. Res. Rec. 1953 1953 , 10–20. doi:10.3141/1953-02

Yan, X., Radwan, E., Mannila, K.K., 2009. Analysis of truck-involved rear-end crashes using multinomial logistic regression. Adv. Transp. Stud. 17 17 , 39–52.

Yasmin, S., Bhowmik, T., Rahman, M., Eluru, N., 2018a. Enhancing Non-Motorized Safety by Simulating Non-Motorized Exposure using a Transportation Planning Approach. Present. Transp. Res. Board Annu. Meet. Washingt. D.C., 2018 1–9.

Yasmin, S., Eluru, N., 2016. Latent segmentation based count models: Analysis of bicycle safety in Montreal and Toronto. Accid. Anal. Prev. 95, 157–171. doi:10.1016/j.aap.2016.07.015

Yasmin, S., Eluru, N., Lee, J., Abdel-Aty, M., 2016. Ordered fractional split approach for aggregate injury severity modeling. Transp. Res. Rec. 2583 1 , 119–126. doi:10.3141/2583-15

Yasmin, S., Eluru, N., Pinjari, A.R., Tay, R., 2014. Examining driver injury severity in two vehicle crashes - A copula based approach. Accid. Anal. Prev. 66, 120–135. doi:10.1016/j.aap.2014.01.018

Yasmin, S., Momtaz, S.U., Nashad, T., Eluru, N., 2018b. A multivariate copula-based macro-level crash count model. Transp. Res. Rec. 2672 30 , 64–75. doi:10.1177/0361198118801348

Ye, X., Pendyala, R.M., Shankar, V., Konduri, K.C., 2013. A simultaneous equations model of crash frequency by severity level for freeway sections. Accid. Anal. Prev. 57, 140–149. doi:10.1016/j.aap.2013.03.025

Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Saf. Sci. 47 3 , 443–452. doi:10.1016/j.ssci.2008.06.007

Yu, R., Abdel-Aty, M., 2013. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. Accid. Anal. Prev. 58, 97–105. doi:10.1016/j.aap.2013.04.025

Zeng, Q., Huang, H., Pei, X., Wong, S.C., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Anal. Methods Accid. Res. 10, 12–25. doi:10.1016/j.amar.2016.03.002

Zeng, Q., Wen, H., Huang, H., Pei, X., Wong, S.C., 2018. Incorporating temporal correlation into a multivariate random parameters Tobit model for modeling crash rate by injury severity. Transp. A Transp. Sci. 14 3 , 177–191. doi:10.1080/23249935.2017.1353556

Zeng, Q., Wen, H., Huang, H., Pei, X., Wong, S.C., 2017. A multivariate random-parameters Tobit

model for analyzing highway crash rates by injury severity. Accid. Anal. Prev. 99, 184–191. doi:10.1016/j.aap.2016.11.018

Zhan, X., Abdul Aziz, H.M., Ukkusuri, S. V., 2015. An efficient parallel sampling technique for Multivariate Poisson-Lognormal model: Analysis with two crash count datasets. Anal. Methods Accid. Res. 8, 45–60. doi:10.1016/j.amar.2015.10.002

Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Anal. Methods Accid. Res. 1, 39–52. doi:10.1016/j.amar.2013.11.001