Exploring Analytical, Simulation-Based, And Hybrid Model Structures For Multivariate Crash Frequency Modeling

Tanmoy Bhowmik
Postdoctoral Scholar
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 1-407-927-6574; Fax: 1-407-823-3315
Email: tanmoy78@knights.ucf.edu
ORCiD number: 0000-0002-0258-1692

Moshiur Rahman
Ph.D.
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 321-276-7580, Fax: 407-823-3315
Email: moshiur@knights.ucf.edu

Shamsunnahar Yasmin
Research Fellow – Road Safety Engineering
Centre for Accident Research & Road Safety – Queensland (CARRS-Q)
Faculty of Health
Queensland University of Technology (QUT)
130 Victoria Park Road, Kelvin Grove, QLD, 4059, Australia
Email: shams.yasmin@qut.edu.au
Telephone: +61731384677
ORCiD number: 0000-0001-7856-5376

Naveen Eluru
Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-823-4815, Fax: 407-823-3315
Email: naveen.eluru@ucf.edu
ORCiD number: 0000-0003-1221-4113

*Corresponding author

ABSTRACT

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 current research effort undertakes a comparison between 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 hybrid (combination of analytical and simulation based system) approach to incorporating unobserved heterogeneity. Within the proposed hybrid 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. Further, to the best of authors' knowledge, this study is the first of its kind to incorporate attribute variability (random parameters) effect within the copula framework. 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 resulting data fit and prediction performance offered by the proposed approach clearly highlights the hybrid model - Random Parameter Copula based approach's superiority over the purely simulation-based multivariate model in our study context. The comparison exercise is further augmented by undertaking an in-depth comparison for different count events across different crash types and a correct classification analysis. The estimated results further reinforce the improved performance of the Random Parameter Copula-based multivariate approach. The applicability of the model for hot spot identification is illustrated by generating plots identifying high-crash and low- crash zones by crash type in the Central Florida region.

Keywords: *Simulation approach; Closed-form approach; Dependency; Copula; Crash types; Comparison exercise; and random parameters within Copula.*

1 BACKGROUND

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; Wang et al., 2019). 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).

A summary of research efforts from the two streams described above are presented in Table 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 (see (Bhowmik et al., 2019a; Yasmin and Eluru, 2018) for details of fractional split approach) 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.

1.1 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 (Bhowmik, 2020). 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 parameters) effect within the copula framework for crash frequency analysis. 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 "(see discussion on copula structures below). 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 high-crash and low-crash zones by crash type for policy implications and prioritizations.

The rest of the paper is organized as follows: The next section presents the methodological framework adopted in the analysis while section 3 provides a detailed description of the dataset. Model findings are offered in the fourth section followed by the comparison results (by evaluating predictive performance) and spatial distribution in section 5 and 6 respectively. Finally, a summary of model findings and conclusions are presented in Section 6.

2 METHODOLOGY

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.

2.1 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 (Bhowmik et al., 2018; Yasmin and Eluru, 2018) 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.

2.2 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 (Bhowmik et al., 2019b, 2018) 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:

$$\eta_{ij} = \boldsymbol{\gamma_j} \boldsymbol{s}_{ij} \tag{6}$$

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 paper, 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 $\boldsymbol{\sigma_j}$.

2.3 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; Yasmin et al., 2018b) for a detailed description on copula framework). 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 $\Lambda_{12\ldots M}(c_{i1}, c_{i2}, \ldots c_{iJ})$ is the M variate joint distribution with corresponding marginal distributions. Subsequently, the M variate distribution $\Lambda_{12\ldots M}(c_{i1}, c_{i2}, \ldots c_{iJ})$ can be generated as a joint cumulative probability distribution of uniform [0, 1] marginal variables U_1, U_2 … U_J as below:

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

The joint distribution (of uniform marginal variable) in equation 2 can be generated by a function $C_{\theta i}(.,.)$ such that:

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

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:

$$\begin{aligned}&\varrho_{i\theta}\left(\Lambda_1(c_{i1}), \Lambda_2(c_{i2}) \dots\ \Lambda_M(c_{iJ})\right) \\ &= \sum_{a_1=1}^{2}\sum_{a_2=1}^{2} \dots \sum_{a_J=1}^{2} (-1)^{a_1+a_2+\dots\ a_J}\left[C_{i\theta}\left(\Lambda_1(c_{i1}+a_1-2), \Lambda_2(c_{i2}+a_2\right.\right. \\ &\qquad \left.\left. -2) \dots \Lambda_M(c_{iJ}+a_J-2)\,;\, \theta_i\right)\right]\end{aligned} \tag{11}$$

The reader would note the probability in Equation 8 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 $\Lambda_1(c_{\mathrm{i1}}), \Lambda_2(c_{\mathrm{i2}})$ … $\Lambda_M(c_{iM})$ 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:

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

Thus, the log-likelihood function (LL) with the joint probability expression in Equation 7 can be written as:

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

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. Figure 1 represents the graphical description of the implied dependency structures for the 4

considered copulas. Archimedean copulas, in their multivariate forms, allow only positive associations and equal dependencies among pairs of random variables. As seen from figure 1a, we can observe that Frank copula has a symmetric dependency structure that ensures higher dependency for unobserved variables around the mean of the distribution. On the other hand, Clayton copula can be applied when there is strong left tail dependence without any significant right tail dependency (as indicated by figure 1b). The Gumbel and Joe copulas (figure 1c and 1d) offer the mirror image to Clayton copula by allowing for stronger dependency toward the right tails of the distribution. Between Joe and Gumbel copula, Joe copula allows for a stronger right tail dependency. In our empirical context, we hypothesize that Clayton copula will be best suited because it is likely that there is a stronger dependency across crash types at lower crash spectrum (compared to the higher end).

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.

2.4 Copula-Based Random Parameter Multivariate NB Model

Building on the model structure in 2.3, 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 2.3 is updated as follows:

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

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 . In this paper, it is assumed that these elements are drawn from independent normal distributions: $\Phi \sim N(0, v_j^2)$. Thus, conditional on Φ, the likelihood function for the joint probability can be expressed as:

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

Finally, the log-likelihood function is:

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

The model estimation routine is coded in GAUSS Matrix Programming software.

3 DATA PREPARATION

Our study area includes the Central Florida region with 4,747 TAZs. The analysis is conducted using the 2016 crash records obtained from Florida Department of Transportation (FDOT) Crash Analysis Reporting System and Signal Four Analytics databases. 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. For the motorists, road segment was found to be most unsafe place (48.5%) followed by intersection (38.9%). Table 2 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.

3.1 Variables Considered

A host of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics are considered for the current research effort (Bhowmik et al., 2019b, 2018). Information about the variables are gathered from FDOT Transportation Statistics Division, US Census Bureau, American Community Survey (ACS) and Florida Geographic Data Library (FGDL) 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. 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. Finally, in our analysis, to accommodate for spatial spillover effects we examined the characteristics of neighboring zones. 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). Across the dataset, the number of surrounding zones range from 1 to 21 with an average value of 6.43.

Table 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 segment level. While we estimated spatial spill-over variables for all variables, we only present the variables that offered significant effects in the model. Several functional forms and specifications for different variables are explored. The final

specification of the model development was based on removing the statistically insignificant variables in a systematic process based on 90% significance level.

4 EMPIRICAL ANALYSIS

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 4. 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 made. 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 performs better in terms of data fit compared to other copula models in both classes (without random parameter and with random parameters). 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 parameters) provide improved data fit compared to the RPMNB model (except Joe copula without random parameters). 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.

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 6). Table 5 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 5. A positive (negative) sign for a variable in the crash count component of Table 5 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.

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.

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 5, 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.

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 (Eluru et al., 2016; 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.

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%). The result highlights how the impact of VMT can vary across zones potentially due to changes in regional characteristics (such as driver behavior, and/or geometric design) across different parts of the study region. Additionally, proportion of heavy vehicles is found to be positively associated with motorized road segment crashes.

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.

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). This is an interesting finding that highlights the varying effect (both positive and negative) of the same variable across the zones. Several possible reasons can be attributed to such variability. For example, higher number of non-motorists means higher exposure, hence in some zones it results in possibly increased number of non-motorized crashes. At the same time, in some zones drivers might drive cautiously as they expect more non-motorists, resulting in a reduced likelihood of vehicle-non motorists' collision. 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.

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 5. As highlighted earlier, in the current analysis, Clayton copula (with random parameters) 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.

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 7 summarizes the value of these measures for both RPMNB and RPCC models at a disaggregate level. As evident form Table 7, we can observe that RPCC outperforms the RPMNB model across most of the (38 out of 42) measures computed. The result clearly highlights the improved performance of the proposed hybrid 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 2a represent the ratio statistics for different crash types while in figure 2b, 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 2a) 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 in our study context.

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 represents the classification accuracy for both RPMNB and RPCC model by each quartile across different crash type. From Figure 3, 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 improved performance of the copula model in our empirical context.

6 Spatial Distribution

To illustrate the applicability of the estimated copula model, we also identify the high-crash and low-crash zones by using prediction of the estimated RPCC model. Specifically, we generate the predicted number of crashes by crash type and identify the low-crash (bottom ten percentile zones with respect to number of crashes) and high-crash 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 4. Figure 4a to 4d represents the high and low crash locations (zones) for all crash types considered while the high and low-crash zone locations for all crashes (identified based on common high/low-crash zones across all crash types) are presented in the appendix (Figure A1). From figure 4a to 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 high-crash and low-crash zones considering all crashes (Figure A1), 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.

7 CONCLUSIONS

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 performed better relative to the 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 improved performance of the RPCC-based multivariate approach over the RPMNB model in our empirical context. The RPCC based copula model is also employed to generate high-crash and low-crash location (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). Such macro level studies have mostly evolved in safety research with the target of incorporating safety considerations in the transportation planning process. Further, a regional or zonal level safety planning tool can be devised by using macrolevel study and hence are useful not only for the planners but also for the decision-makers. For example, transportation planners are required to forecast future crashes given changes in region's characteristics (population increase, addition of new facility (such as road or major facility). The proposed crash prediction models can aid the process. 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 paper is not without limitations. The reader would note that the simulation based multivariate approach (RPMNB) considered in the study employs the most commonly employed distributional assumption by characterizing the relationship between the various parameters in the form of a multivariate normal distribution. Several research studies have examined alternative distributional assumptions (such as log-normal, and triangular) in simulation-based approaches. It would be interesting as a future research direction to explore how the comparison results might alter if the alternate distributions are incorporated within simulation-based approaches. Also, the analysis and the comparison exercise are conducted using zonal level data. In the future, we can explore if and how the comparison across the various frameworks will alter in a micro level analysis. Further, the study considers the effect of observed spatial attributes, it would be beneficial to capture the spatial unobserved heterogeneity as well. Another avenue for future research would be to explore the transferability of models developed for crash type simultaneously by estimating similar models for multiple spatial units across several years.

ACKNOWLEDGMENT

The authors 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.

REFERENCES

Abdel-Aty, M., Keller, J. and Brady, P.A., 2005. Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression. Transportation Research Record, 1908, 37-45.

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

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

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

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

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

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

Bhat, C.R., 2014. The composite marginal likelihood (CML) inference approach with applications to discrete and mixed dependent variable models (No. D-STOP/2016/101). University of Texas at Austin. Data-Supported Transportation Operations and Planning Center (D-STOP).

Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. Transportation Research Part B, 45(7), 923-939.

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

Bhat, C.R. and Eluru, N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. Transportation Research Part B, 43(7), 749-765.

Bhowmik, T., 2020. Econometric Frameworks for Multivariate Models: Application to Econometric Frameworks for Multivariate Models: Application to Crash Frequency Analysis Crash Frequency Analysis.

Bhowmik, T., Yasmin, S., Eluru, N., 2019a. A multilevel generalized ordered probit fractional split model for analyzing vehicle speed. Analytic Methods in Accident Research, 21, 13–31.

Bhowmik, T., Yasmin, S. and 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. Analytic Methods in Accident Research, 24, 100107.

Bhowmik, T., Yasmin, S. and Eluru, N., 2018. A joint econometric approach for modeling crash counts by collision type. Analytic Methods in Accident Research 19, 16-32.

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

Cameron, A.C., Li, T., Trivedi, P.K. and Zimmer, D.M., 2004. Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. The Econometrics Journal, 7(2), 566-584.

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

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X. and Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accident Analysis and Prevention, 99, 330-341.

Cheng, W., Gill, G.S., Ensch, J.L., Kwong, J. and Jia, X., 2018. Multimodal crash frequency modeling: multivariate space-time models with alternate spatiotemporal interactions. Accident Analysis and Prevention, 113, 159-170.

Chiou, Y.C. and Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accident Analysis and Prevention, 50, 73-82.

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

Chiou, Y.C. and Fu, C., 2015. Modeling crash frequency and severity with spatiotemporal dependence. Analytic Methods in Accident Research, 5, 43-58.

Dong, C., Clarke, D.B., Nambisan, S.S. and Huang, B., 2016. Analyzing injury crashes using random-parameter bivariate regression models. Transportmetrica A, 12(9), 794-810.

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

Eluru, N., Paleti, R., Pendyala, R. and Bhat, C.R., 2010. Modeling injury severity of multiple occupants of vehicles: Copula-based multivariate approach. Transportation Research Record: Journal of the Transportation Research Board 2165, 1-11.

Eluru, N., Yasmin, S., Bhowmick, T. and Rahman, M., 2016. Enhancing non-motorized safety by simulating non-motorized exposure using a transportation planning approach. URL: https://rosap.ntl.bts.gov/view/dot/32653 (accessed 9.21.2018).

Heydari, S., Fu, L., Miranda-Moreno, L.F. and Jopseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. Analytic Methods in Accident Research, 13, 16-27.

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

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

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

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M. and 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. Accident Analysis and Prevention, 111, 12-22.

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

Liu, X., Saat, M.R. and Barkan, C.P., 2017. Freight-train derailment rates for railroad safety and risk analysis. Accident Analysis and Prevention, 98, 1-9.

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

Mothafer, G.I., Yamamoto, T. and Shankar, V.N., 2016. Evaluating crash type covariances and roadway geometric marginal effects using the multivariate Poisson gamma mixture model. Analytic Methods in Accident Research, 9, 16-26.

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

Nashad, T., Yasmin, S., Eluru, N., Lee, J. and Abdel-Aty, M., 2016. Joint modeling of pedestrian and bicycle crashes: copula-based approach. Transportation research record, 2601, 119-127.

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

Sener, I.N., Eluru, N. and Bhat, C.R., 2010. On jointly analyzing the physical activity participation levels of individuals in a family unit using a multivariate copula framework. Journal of Choice Modelling, 3(3), 1-38.

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

Wang, K., Bhowmik, T., Zhao, S., Eluru, N. and Jackson, E., 2021. Highway safety assessment and improvement through crash prediction by injury severity and vehicle damage using Multivariate Poisson-Lognormal model and Joint Negative Binomial-Generalized Ordered Probit Fractional Split model. Journal of Safety Research, 76, 44-55.

Wang, K., Bhowmik, T., Yasmin, S., Zhao, S., Eluru, N. and Jackson, E., 2019. Multivariate copula temporal modeling of intersection crash consequence metrics: a joint estimation of injury severity, crash type, vehicle damage and driver error. Accident Analysis and Prevention, 125,188-197.

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

Yasmin, S., Bhowmik, T., Rahman, M. and Eluru, N., 2016. Enhancing Non-Motorized Safety by Simulating Non-Motorized Exposure Using a Transportation Planning Approach. Presented at the 98th annual meeting of Transportation research Board.

Yasmin, S. and Eluru, N., 2018. A joint econometric framework for modeling crash counts by severity. Transportmetrica A, 14(3), 230-255.

Yasmin, S., Eluru, N., Pinjari, A.R. and Tay, R., 2014. Examining driver injury severity in two vehicle crashes–A copula based approach. Accident Analysis and Prevention, 66, 120-135.

Yasmin, S., Momtaz, S.U., Nashad, T., Eluru, N., 2018. A Multivariate Copula-Based Macro-Level Crash Count Model. Transportation Research Record, 2672, 64–75.

Ye, X., Pendyala, R.M., Shankar, V. and Konduri, K.C., 2013. A simultaneous equations model of crash frequency by severity level for freeway sections. Accident Analysis and Prevention 57, 140-149.

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

Zeng, Q., Huang, H., Pei, X. and Wong, S.C., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Analytic Methods in Accident Research, 10, 12-25.

Zeng, Q., Wen, H., Huang, H., Pei, X. and Wong, S.C., 2017. A multivariate random-parameters Tobit model for analyzing highway crash rates by injury severity. Accident Analysis and Prevention, 99, 184-191.

Zhan, X., Aziz, H.A. and Ukkusuri, S.V., 2015. An efficient parallel sampling technique for Multivariate Poisson-Lognormal model: Analysis with two crash count datasets. Analytic Methods in Accident Research, 8, 45-60.

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

LIST OF FIGURES

LIST OF TABLES

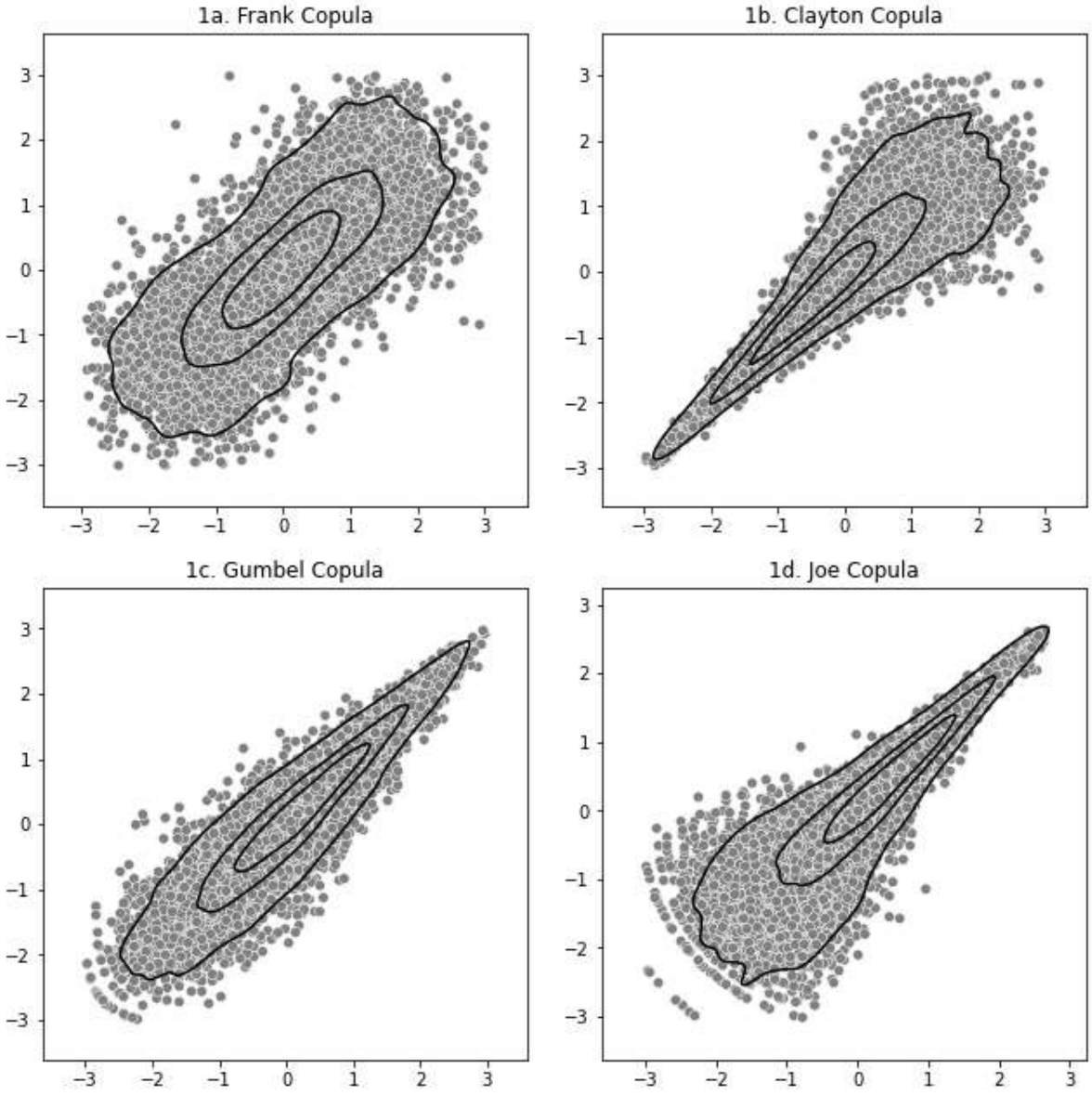


Figure 1: Dependency Structure of Different Archimedean Copulas*

*The correlation across all the pairs (4 in our case) are same. Hence, for the ease of presentation, we present the scatter plot across the two pairs. In the figure, the contours are employed to partition the random draws into quartiles. The first contour envelops the first quartile. The intermediate space between the second contour and the first contour represents the second quartile and so on. The shape of the contours illustrates the dependency structure such as radial symmetry for Frank.

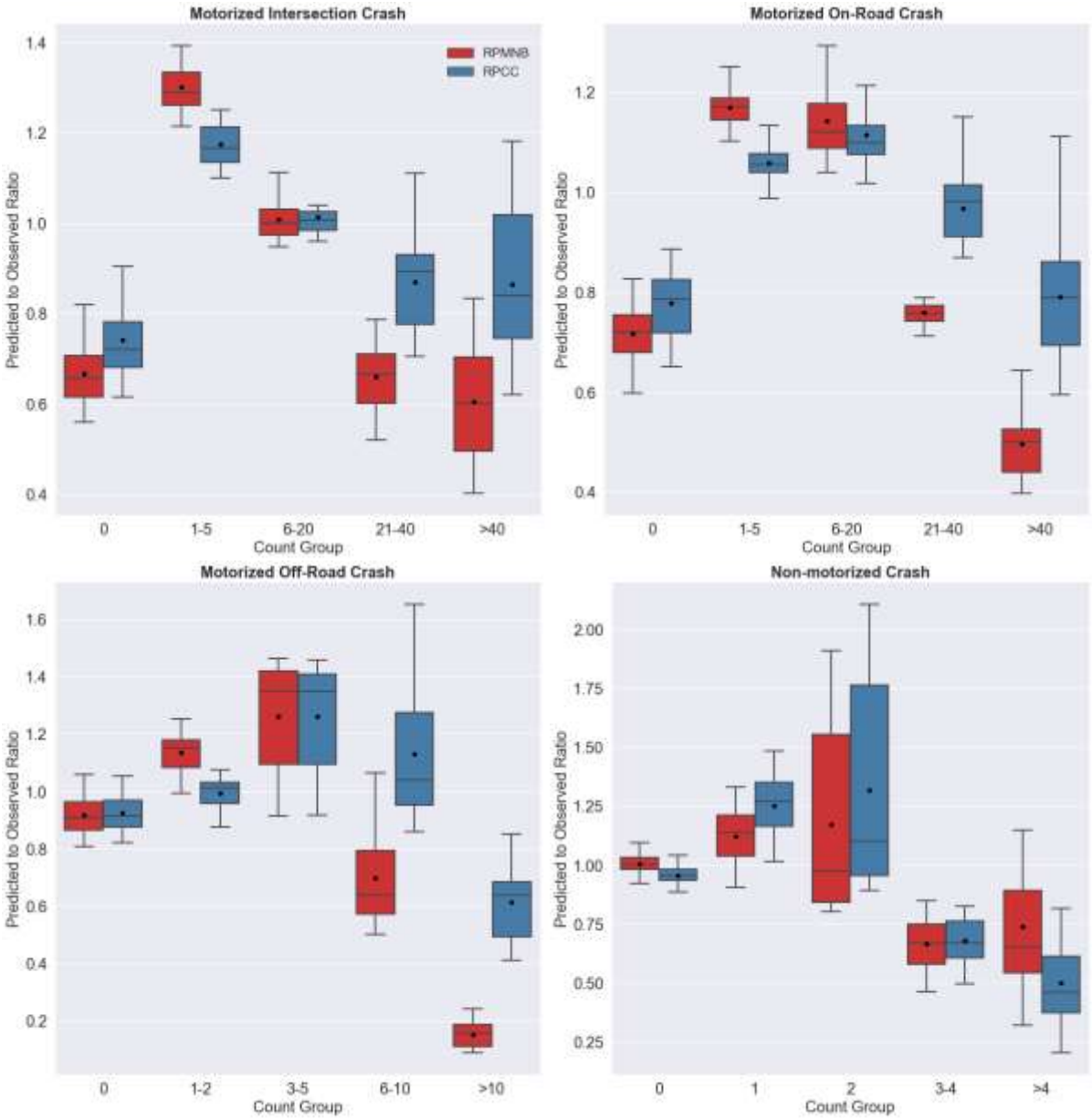


Figure 2a: Predicted to Observed Ratio for Different Crash Types

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

The x-axis in the figure represents the count group. For example, 1-5 represents the number of TAZs with 1-5 crashes. The y-axis presents the ratio of number of TAZs predicted to be within this group to the number of TAZs observed to be within this group. To generate the box plot, the ratio computation is repeated 20 times using samples of 250 records from the validation set. The black dot in the middle represents the mean predicted to observed ratio for the 20 different samples. In our ratio measure, model that offers proximity to 1 offers better performance while a perfect prediction would be represented by a value of 1.

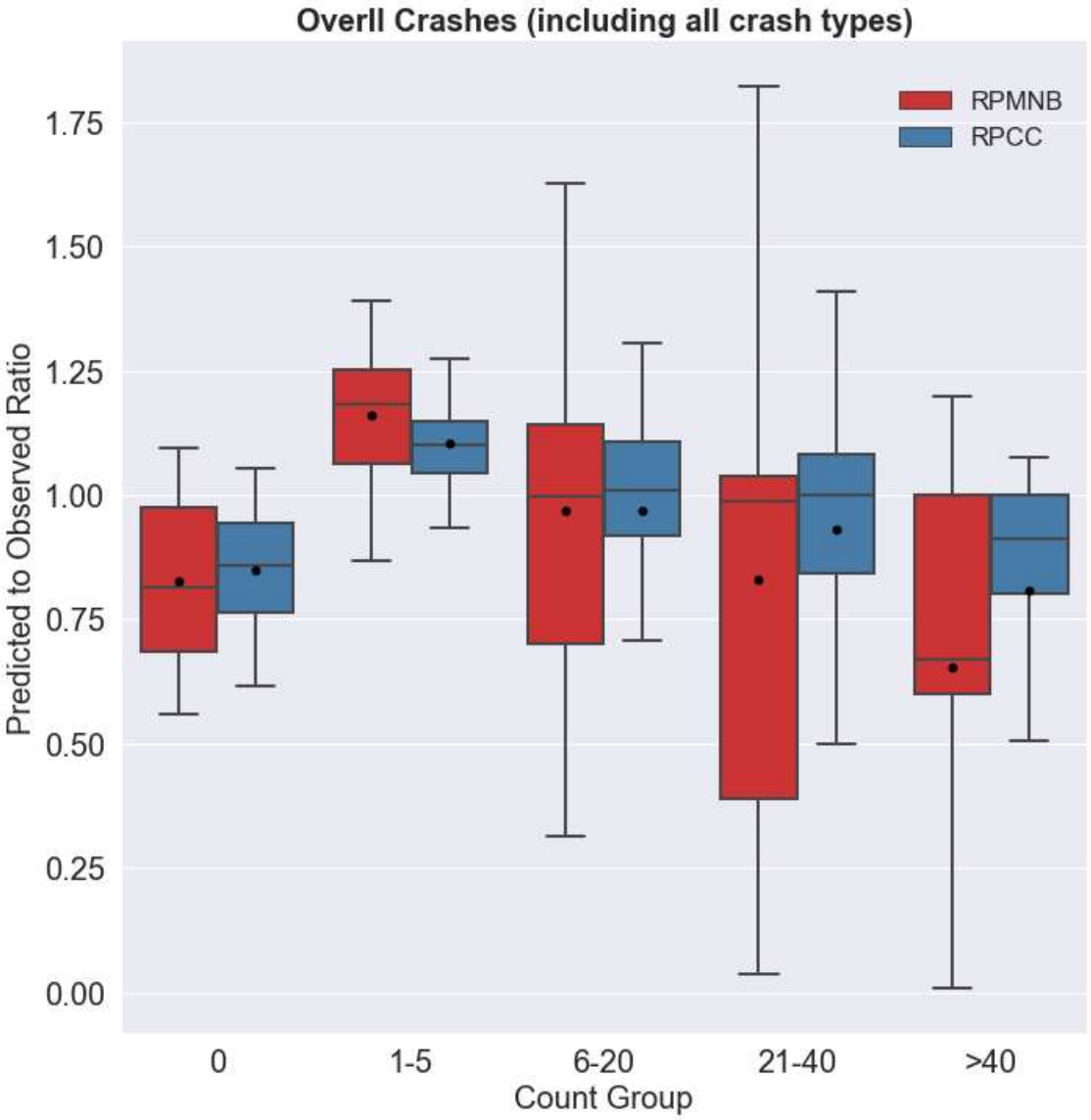


Figure 2b: Predicted to Observed Ratio for Overall Crashes

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

See discussion above (figure 1a) for the figure

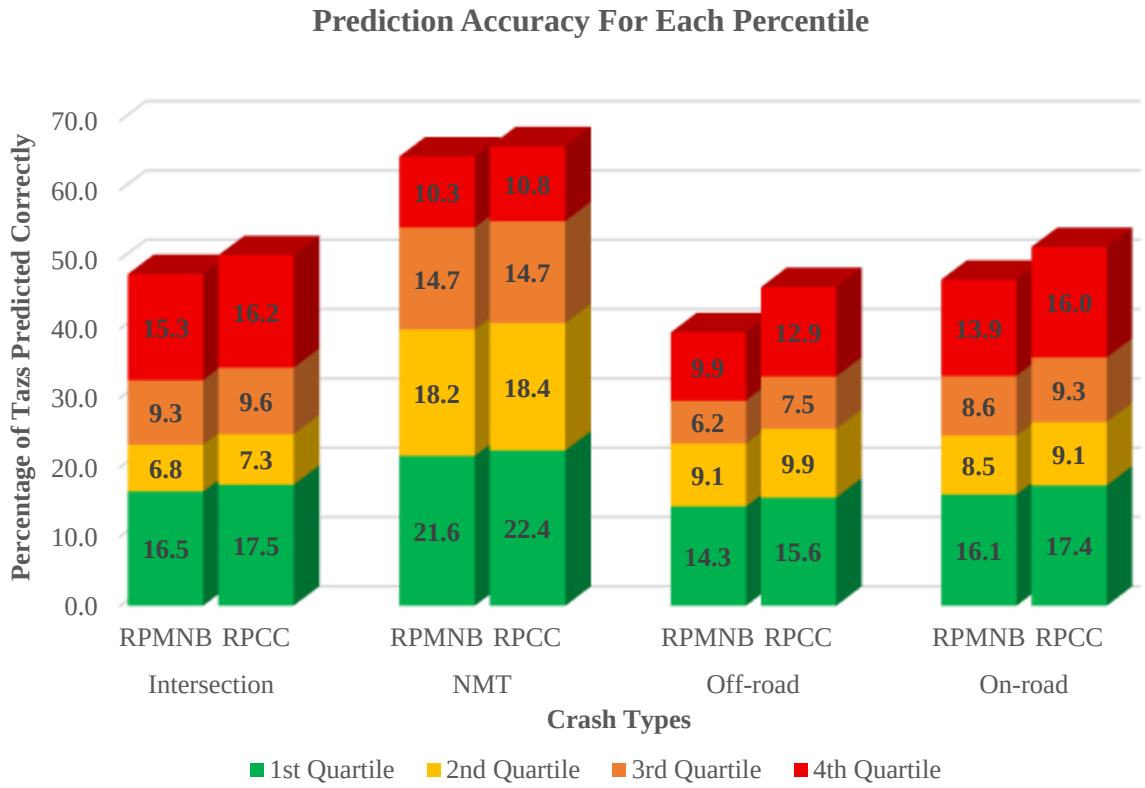
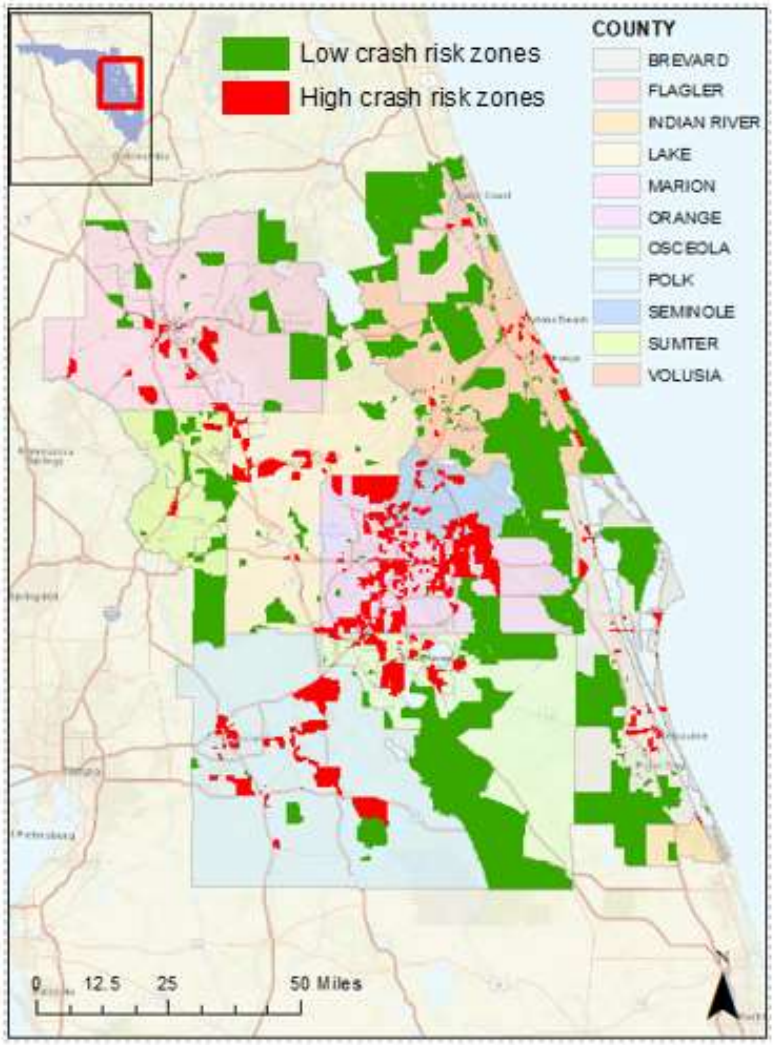
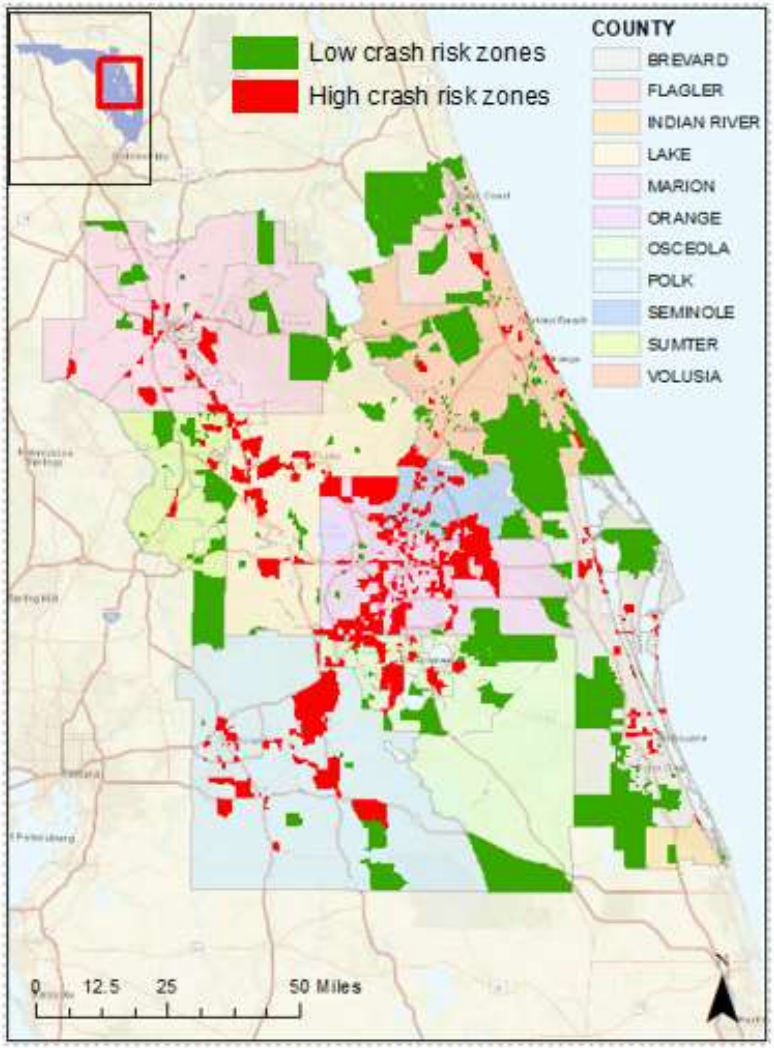


Figure 3 Prediction Accuracy for Two Frameworks by Crash type Quartile

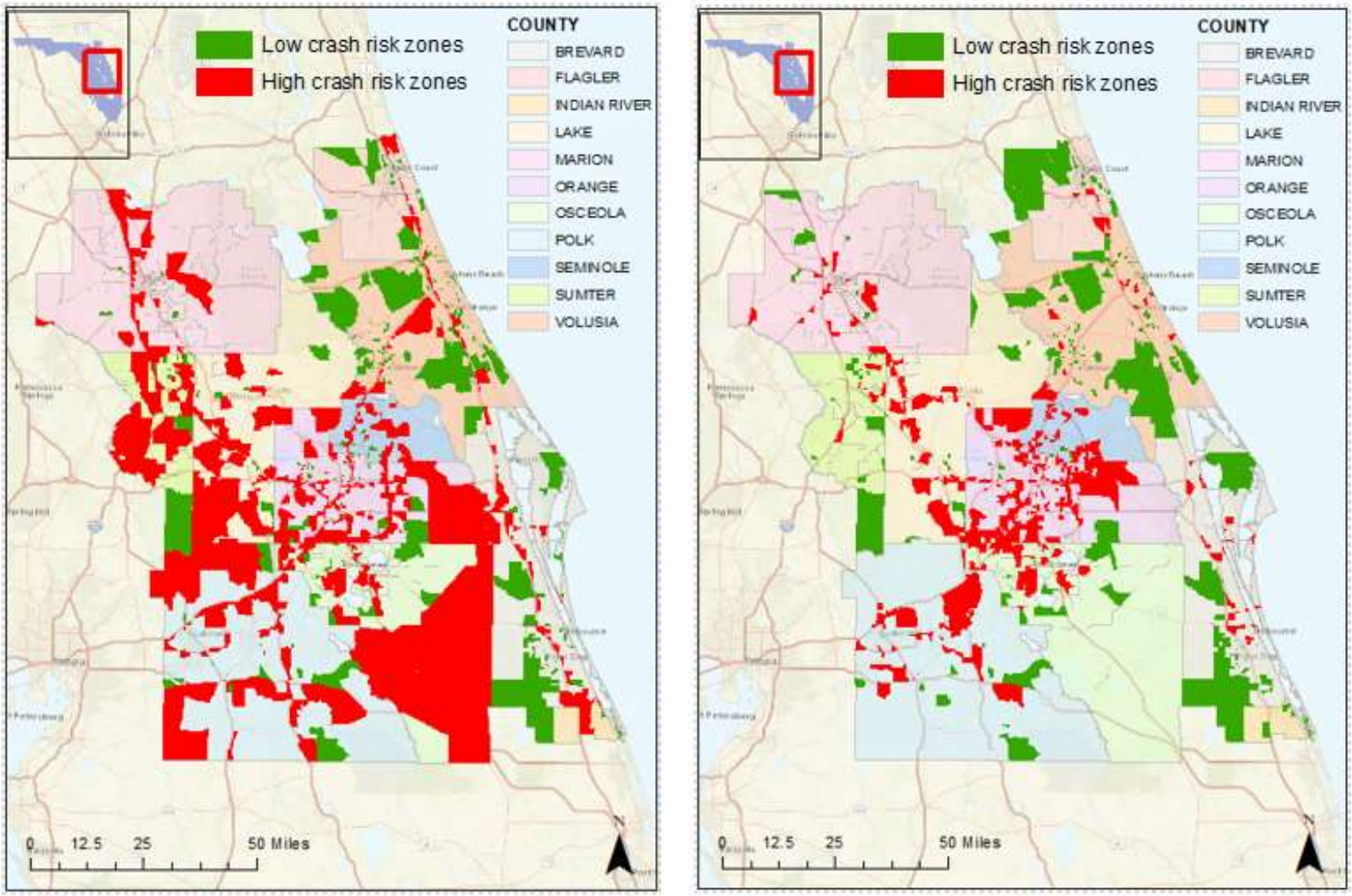
Note: *RPMNB=Random parameter multivariate negative binomial model, RPCC =Random Parameter Clayton Copula model



a. Intersection Crash



b. Non-motorized Crash



c. Off-road Crash

d. On-road Crash

Figure 4: Spatial Distribution for Every Crash Types

TABLE 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)[1]	Micro	Multivariate tobit regression	MSL*	Rates and counts 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

[1] *Anastasopoulos et al. (2012) and Anastasopoulos (2016)) considered the random parameters within a multivariate Tobit framework with a Copula framework. However, it is important to recognize that the Copula considered in these studies is the traditional multivariate normal distribution based dependency without any additional flexibility for non-normal marginals and non-normal multivariate couplings. This also explains why the models were only applied to crash rates. For crash frequency models in the Anastasopoulos (2016) the marginals do not take the normal form and would require additional transformations (as considered in our study) for applying more flexible copula methods.*

(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	Rates and counts 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
(Bhowmik et al., 2019b)	Macro	Panel Mixed Negative Binomial Model	MSL	by crash type - rear-end, head-on, angular, off-road, sideswipe, non-motorized	6
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
(Yasmin and Eluru, 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
(Wang et al., 2021)	Micro	Multivariate Poisson log-normal model, Joint Negative Binomial-Generalized Ordered Probit Fractional Split (NB-GOPFS) Model	ML, QMCSL	by crash severity - (1) fatal and incapacitating injury, (2) non-incapacitating and possible injury, (3) property damage only; by vehicle damage – (1) severe damage, (2) moderate damage, (3) minor damage	3,3
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 negative binomial model	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 2 Descriptive Statistics of Dependent Variables

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 3 Summary Statistics of Exogenous Variables (Zonal Level)

Variables (4747)	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 roads	(Arterial roads 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 intersection	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 55mph	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: 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 shopping center	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 vehicles	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 TAZ in acre	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 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 zones	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
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

TABLE 4 Summary of Statistical Data Fit from Different Model Systems

Model (Sample Size = 3,800)		**Log-Likelihood**	**No. of Parameter**	**AIC**	**BIC**
Independent Model		-33108.8	51	66319.6	66637.9
RPMNB		-32541.4	54	65190.7	65527.8
Copula Without random effect	*Frank*	-32330.7	53	64767.3	65098.1
	Clayton	-32285.6	53	64677.1	65007.9
	Gumbel	-32477.9	52	65059.9	65384.6
	Joe	-32609.3	52	65322.5	65647.2
Copula With random effect	*Frank*	-32324.9	54	64757.9	65095.0
	Clayton	-32269.3	55	64648.5	64991.9
	Gumbel	-32437.4	53	64980.8	65311.7
	Joe	-32345.8	54	64799.6	65136.8

TABLE 5 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 6 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 7 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 performance (lower measure)

APPENDIX A

TABLE A1 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								

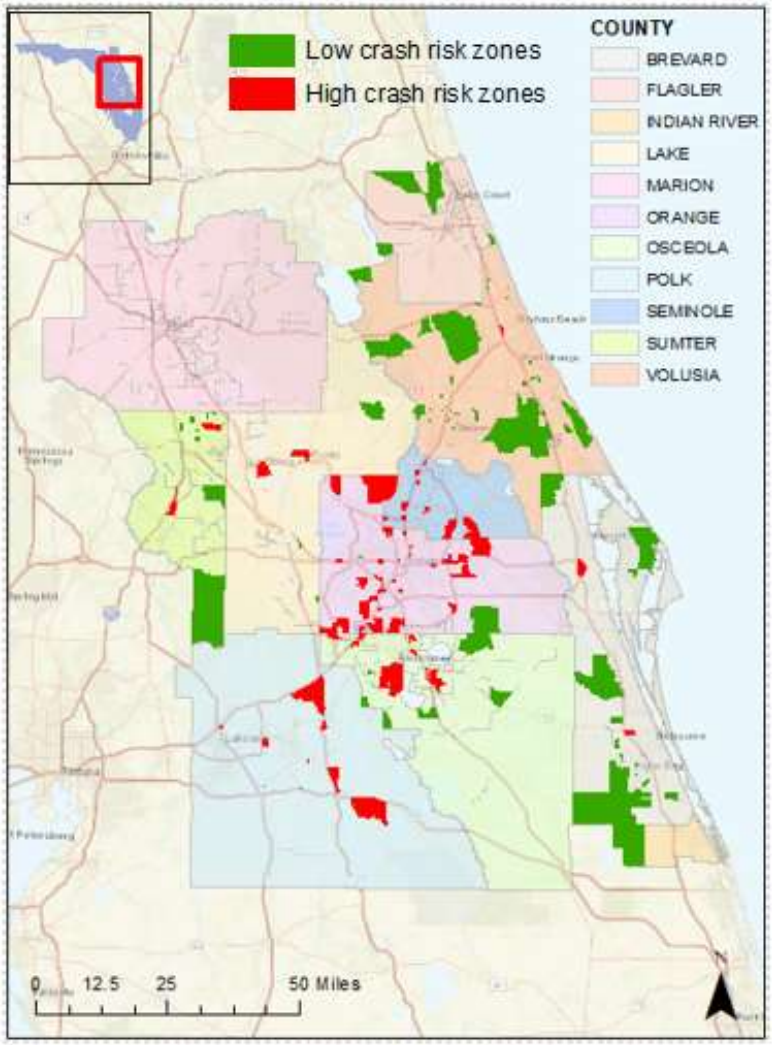


Figure A1 Spatial Distribution for Overall Crashes (Considering all crash types together)