A Multivariate Copula Based Macro-level Crash Count Model

Shamsunnahar Yasmin
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: shamsunnahar.yasmin@ucf.edu

Salah Uddin Momtaz
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: smomtaz@Knights.ucf.edu

Tammam Nashad
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: tammam24@knights.ucf.edu

Naveen Eluru*
Associate Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-823-4815, Fax: 407-823-3315
Email: naveen.eluru@ucf.edu

November 2017

*Corresponding author

97th Annual Meeting of the Transportation Research Board, 2018, Washington DC

Submitted to: Standing Committee on Statistical Method (ABJ80) committee for **presentation and publication**

Word count: 135 abstract + 5394 texts + 1219 references+ 3 tables + 0 figures = 7498 equivalent words

ABSTRACT

The current study contributes to safety literature both methodologically and empirically by developing a macro-level multivariate copula-based crash frequency model for crash counts. The multivariate model accommodates for the impact of observed and unobserved effects on zonal level crash counts of different road user groups including car, light truck, van, other motorized vehicle (including truck, bus and other vehicles) and non-motorist (including pedestrian and bicyclist). The proposed model is estimated using Statewide Traffic Analysis Zone (STAZ) level road traffic crash data for the state of Florida. A host of variable groups including land-use characteristics, roadway attributes, traffic characteristics, socioeconomic characteristics and demographic characteristics are considered. The model estimation results illustrate the applicability of the proposed framework for multivariate crash counts. Model estimation results are further augmented by evaluation of predictive performance and policy analysis.

BACKGROUND

Road traffic crashes affect the society as a whole both emotionally and economically and are rightfully recognized as a national health problem (*1; 2*). In reducing the undue burden of road crashes and their consequences, road safety literature is focused on devising both proactive and reactive safety management policies at the user, system and/or planning level through evidence-based and data-driven strategies. Crash frequency analysis, specifically macro-level crash models, is a major component for devising and evaluating these road safety policies at a planning level. Macro-level studies have mostly evolved in safety research to incorporate safety considerations within the transportation planning process. The outcome of these models is also useful to devise safety-conscious decision support tools to facilitate a proactive approach in assessing medium and long-term policy based countermeasures. The current research effort contributes to the safety literature methodologically and empirically with specific focus on macro-level crash frequency analysis.

Econometric approaches of developing crash prediction models in safety literature are dominated by traditional count regression frameworks (Poisson and negative binomial (NB) models) in univariate modeling systems (see (*3-5*)). These studies identify a single count variable for different crash attribute levels (road user group, crash severity, crash types, or vehicles types) for a spatial unit and study the impact of exogenous variables. However, as documented in literature, crash counts across different attribute levels are likely to be dependent for the same observation resulting in a multivariate crash event set (*6*). Ignoring such correlation, if present, may lead to biased and inefficient parameter estimates resulting in erroneous policy implications (*7; 8*). To that extent, road safety researchers and analysts have estimated multivariate count models to produce more accurate predictions (see (*9*) for a detailed list of these studies).

It is beyond the scope of this paper to provide a comprehensive literature review on multivariate crash count models. For a detailed review of multivariate frameworks employed in safety, the reader is referred to recent review studies (*3; 10; 11*). Within the multivariate scheme, studies have predominantly explored crash counts by severity outcome levels and by crash types. However, multivariate crash event set may also arise when examining crash occurrences by different road user groups involved in crashes. In fact, studies have recognized this and developed multivariate crash count models for different road user groups involved in crashes – for pedestrian and bicyclists (*12; 13*), for vehicle types (*14*), for travel modes (*15*).

In these studies, the general trend is to focus entirely either on motorized road user group or on non-motorized road users (except (*15*)). However, both of these road user groups share the same travel environment within a spatial planning unit over a specific given period of time. Therefore, it is possible that the same set of observed and unobserved factors influence crash occurrences of these two different road user groups. For instance, higher number of uncontrolled intersections (usually observed to analysts) at a zonal level are likely to result in higher number of vehicular conflicts as well as higher number of pedestrian/bicyclists involved crashes. At the same time, if a zone has higher proportion of blind spots at intersections (usually unobserved to analyst) it may contribute to higher crash events involving both motorists and non-motorists. Therefore, it is important to examine crash events as a joint process considering both of these road user groups simultaneously. Further, while analyzing motorized road user groups, recognizing the implicit differences between various motorized vehicle groups is very useful. It is plausible that different exogenous variables may have distinct impact on crash occurrence across various motorized road user groups. For instance, zones with higher truck volumes may have higher number of crashes involving heavy vehicles. Moreover, it is also important to examine separate risk factors related to different types of passenger vehicles rather than considering all passenger vehicles as one category. As documented in literature, the diversity in passenger vehicle fleet has deteriorating effects on overall safety (*16*). In the United States, the sales of light truck has in fact increased 7% in 2016 relative to 2015 (*17*). The shift from light to heavy passenger vehicles are likely to result in 4.3 additional crashes (for each fatal crash that occupants of large passenger vehicles avoid) that may result in fatalities among occupants of light vehicles or non-motorists involved in crashes with these heavy passenger vehicles (*18*).

Given the potential difference in safety impacts of different types of passenger vehicles, it is important to examine separate risk factors for different types of passenger vehicles, which would allow us to devise more tangible actions and policies. The ***first contribution*** of our study is to develop multivariate crash count model for crashes involving different road user groups involved in crashes with higher resolution classification of passenger vehicle fleet. Specifically, we examine zonal level car, light truck, van, other motorized vehicles (bus, truck and other vehicles) and non-motorist (pedestrian and bicyclist) involved crash counts in a multivariate count model framework.

Traditionally, in existing safety literature, the multivariate count models are examined by considering unobserved error components that jointly affect the dependent variables. In particular, the traditional multivariate count modeling approaches partition the error components of the dependent variables to accommodate for a common term and an independent term across dependent variables (see (*6*) for a detailed discussion of various methodologies). Thus, any probability computation, in accommodating such unobserved effect, requires integrating the probability function over the error term distribution. The exact computation is dependent on the distributional assumption and does not have a closed form expression usually. Thus, the estimation procedure requires the adoption of maximum simulated likelihood (MSL) approach in the classical approach or Markov Chain Monte Carlo (MCMC) approach in the Bayesian realm. MSL and MCMC methods provide substantial flexibility in accommodating for unobserved heterogeneity. However, the probability evaluation with high dimensional integrals is affected by the challenges in generating high dimensionality of random numbers and longer computational run times. The process of 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 (*19*) for a discussion). Within this simulated framework, the model structures employed in developing multivariate crash count model include multivariate-Poisson, multivariate Poisson-lognormal, multivariate random-parameters zero-inflated negative binomial, multinomial-generalized Poisson, multivariate conditional autoregressive, multivariate tobit and multivariate Poisson gamma mixture count models. Another multivariate count modeling approach based on the development of multivariate function has most recently been employed by Narayanamoorthy et al. (*20*). The approach circumvents the challenges associated with simulation by adopting analytical approximation of the likelihood function.

More recently, a closed form parametric formulation that obviates the need for an approximation or demanding simulation has been employed in existing econometric literature for examining joint count events. The approach, referred to as copula-based approach, allows for flexible correlation structures across joint dimensions thus enhancing the flexibility of the multivariate approach. The copula-based approach allows for analytical computation of log-likelihood based on standard maximum likelihood procedure; it is generally tractable and offers stable inference. The copula formulation allows for additional flexibility in specifying the marginal distribution. While the application of copula has seen a surge of interest in examining multivariate continuous and disaggregate discrete data, the studies employing copulas for examining aggregate level count events are relatively few (for application of copulas in continuous and disaggregate level discrete data see (*21-24*)). Copula based bivariate count model has been employed in econometrics and applied statistics (*25; 26*). To date, only one study in safety literature has employed bivariate copula count model in examining pedestrian and bicycle crash risks simultaneously (*12*).

The current study generalizes the bivariate copula count model for examining multivariate count data. Specifically, we formulate and estimate a multivariate copula count model for examining zonal level crash counts by different road user groups involved in crashes. To be sure, the application of multivariate copula count model has been demonstrated by Nikoloulopoulos and Karlis (*27*) in examining the correlation among the number of purchases of four different products (food, non-food, hygiene and fresh). In current study context, we employ multivariate copula count model for examining five different crash count dimensions – car, light truck, van, other motorized vehicle and non-motorists involved crashes. The *second contribution* of our study is to develop a closed form multivariate copula count model to accommodate for the impact of observed and unobserved effects on zonal level crash counts of different road user groups. For examining the count components of the multivariate copula-based model, we employ negative binomial (NB) regression framework. The NB model that has a built-in dispersion parameter is widely employed in safety literature. It provides a natural enhancement over the Poisson model and is easy to estimate with a closed form structure to accommodate for over-dispersion (the variance of the crash count variable usually exceeds the mean of the crash count variable). In existing safety literature, researchers have also employed count modeling frameworks accommodating the preponderance of zero count events (such as zero-inflated and hurdle models). However, NB is the most frequently used statistical technique for examining crash count events (*10*). Therefore, in our current study, we examine crash count within the proposed multivariate copula-based approach by using NB regression framework. The proposed model is estimated using Statewide Traffic Analysis Zone (STAZ) level road crash data for the state of Florida. A host of variable groups including – land-use characteristics, roadway attributes, traffic characteristics, socioeconomic characteristics and demographic characteristics are considered.

In summary, the current research effort contributes to safety literature on macro-level crash count analysis both methodologically and empirically. In terms of methodology, we formulate and estimate a multivariate copula-based count model framework to analyze the crash count events for different road user groups involved in crashes jointly, and we employ NB regression framework for examining the count components. The proposed multivariate copula count model can be employed in developing both macro and micro-level count events. In terms of empirical analysis, our study incorporates crash counts for both motorized and non-motorized road user groups while considering different types of passenger vehicles fleet categories. Specifically, we examine crash counts for car, light truck, van, other motorized vehicle and non-motorist involved crashes by employing multivariate copula count framework. Model estimation results are further augmented by evaluation of predictive performance and policy analysis.

ECONOMETRIC FRAMEWORK

The focus of our study is to propose and estimate a copula-based multivariate NB modeling framework (see (*22; 28*) for a detailed background on copula-based models and see (*27*) for a description of multivariate NB framework). The econometric framework for the joint model is presented in this section.

Let us assume that i be the index for STAZ $(i = 1,2,3, \ldots, N)$ and y_{qi} be the index for crashes occurring over a period of time in a STAZ i; q $(q = 1,2, \ldots, M, M = 5)$ be the index to represent road user group for the multivariate case examined. In this empirical study, q takes the value of 'car' $(q = 1)$, 'light truck' $(q = 2)$, 'van' $(q = 3)$, 'other motorized vehicle' $(q = 4)$ and 'non-motorist' $(q = 5)$. The NB probability expression for random variable y_{qi} can be written as:

$$P_{qi}(y_{qi}|\mu_{qi}, \alpha_q) = \frac{\Gamma(y_{qi}+\alpha_q^{-1})}{\Gamma(y_{qi}+1)\Gamma(\alpha_q^{-1})}\left(\frac{1}{1+\alpha_q\mu_{qi}}\right)^{\frac{1}{\alpha_q}}\left(1-\frac{1}{1+\alpha_q\mu_{qi}}\right)^{y_{qi}} \tag{1}$$

where, $\Gamma(\cdot)$ is the Gamma function, α_q is the NB dispersion parameter specific to road user group q and μ_{qi} is the expected number of crashes occurring in STAZ i over a given period of time for road user group q. We can express μ_{qi} as a function of explanatory variable $(\boldsymbol{x}_{qi})$ by using a log-link function as: $\mu_{qis} = E(y_{qi}|\boldsymbol{x}_{qi}) = exp(\boldsymbol{\beta}_q\boldsymbol{x}_{qi})$, where $\boldsymbol{\beta}_q$ is a vector of parameters to be estimated specific to road user group q.

The correlation or joint behavior of random variables y_{1i}, y_{2i},…y_{Mi} 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 ((*22*) (*29*) provide a detailed description of the copula approach). In constructing the copula dependency, let us assume that $\Lambda_1(y_{1i})$, $\Lambda_2(y_{2i})$ … $\Lambda_M(y_{Mi})$ are the marginal distribution functions of the random variables y_{1i}, y_{2i},…y_{Mi}, respectively; and $\Lambda_{12\ldots M}(y_{1i}, y_{2i}, \ldots y_{Mi})$ is the M variate joint distribution with corresponding marginal distributions. Subsequently, the M variate distribution $\Lambda_{12\ldots M}(y_{1i}, y_{2i}, \ldots y_{Mi})$ can be generated as a joint cumulative probability distribution of uniform [0, 1] marginal variables U_1, U_2 … U_M as below:

$$\begin{aligned}\Lambda_{12\ldots M}(y_{1i}, y_{2i}, \ldots y_{Mi}) &= Pr(U_1 \le y_{1i}, U_2 \le y_{2i} \ldots, U_M \le y_{Mi}) \\ &= Pr[\Lambda_1^{-1}(U_1) \le y_{1i}, \Lambda_2^{-1}(U_2) \le y_{2i} \ldots, \Lambda_M^{-1}(U_M) \le y_{Mi}] \\ &= Pr[U_1 < \Lambda_1(y_{1i}), U_2 < \Lambda_2(y_{2i}) \ldots, U_M < \Lambda_M(y_{Mi})]\end{aligned} \tag{2}$$

The joint distribution (of uniform marginal variable) in equation 2 can be generated by a function $C_{\theta i}(.,.)$ (*30*), such that:

$$\Lambda_{12\ldots M}(y_{1i}, y_{2i}, \ldots y_{Mi}) = C_{\theta i}(U_1 = \Lambda_1(y_{1i}), U_2 = \Lambda_2(y_{2i}) \ldots, U_M = \Lambda_M(y_{Mi})) \tag{3}$$

where, $C_{\theta i}(.,.)$ is a copula function and θ_i is the dependence parameter defining the link between $y_{1i}, y_{2i}, \ldots y_{Mi}$. In the case of continuous random variables, the joint density can be derived from partial derivatives. However, in our study, y_{qi} are nonnegative integer valued events. For such count data, following (*26*), the probability mass function $(\zeta_{\theta i})$ is presented (instead of continuous derivatives) by using finite differences of the copula representation as follows:

$$\zeta_{\theta i}\big(\Lambda_1(y_{1i}),\Lambda_2(y_{2i})\ldots\ \Lambda_M(y_{Mi})\big)$$
$$=\sum_{a_1=1}^{2}\sum_{a_2=1}^{2}\ldots\sum_{a_M=1}^{2}(-1)^{a_1+a_2+\ldots\ a_M}\left[C_{\theta i}(\Lambda_1(y_{1i}+a_1-2),\Lambda_2(y_{2i}+a_2-2)\ldots\Lambda_M(y_{Mi}+a_M-2)\ ;\ \theta_i)\right] \tag{4}$$

The reader would note the probability in Equation 4 is written in terms of 2^M copula evaluations (see (*31; 32*) for a similar derivation). The number of computations increases rapidly with the number of dependent variables (M), but this is not much of a problem when the dependent variable number M is 6 or less because of the closed-form structures of the copula function evaluation. Given the above setup, we specify $\Lambda_1(y_{1i})$, $\Lambda_2(y_{2i})\ \ldots\ \Lambda_M(y_{Mi})$ as the cumulative distribution function (cdf) of the NB distribution. The cdf of NB probability expression (as presented in Equation 1) for y_{qi} can be written as:

$$\Lambda_q\big(y_{qi}|\mu_{qi}\,,\alpha_q\big)=\sum_{k=0}^{y_{qi}}P_{qi}\big(y_{qi}|\mu_{qi}\,,\alpha_q\big) \tag{5}$$

Thus, the log-likelihood function (LL) with the joint probability expression in Equation 5 can be written as:

$$LL=\sum_{i=1}^{N}\ln\Big(\zeta_{\theta i}\big(\Lambda_1(y_{1i}),\Lambda_2(y_{2i})\ldots\ \Lambda_M(y_{Mi})\big)\Big) \tag{6}$$

In the current empirical study, we employ Archimedean copulas that span the spectrum of different kinds of dependency structures including Clayton, Gumbel, Frank, and Joe copulas (see (*22*) 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. The parameters are estimated using maximum likelihood approach. The model estimation is achieved through the log-likelihood functions programmed in GAUSS.

DATA DESCRIPTION

Our study area includes the state of Florida with 8,518 STAZ. Similar to the rest of the North-American transportation trends, the ethos of travel in Florida is also predominantly auto-oriented. The state has nearly 100,000 more crashes in 2015 than in 2011 with higher number of non-motorist fatalities (*33*). These numbers clearly signify that it is important to identify critical factors contributing to road traffic crashes at a planning level for all road user groups to improve overall road safety situation. The traffic crash records are collected and compiled from Florida Department of Transportation (FDOT) Crash Analysis Reporting System (CARS) database for the year 2014. The geocoded crash data are aggregated at the level of STAZ for each road user group. Thus, the dependent variable of the empirical study is zonal level number of crash counts involving car, light truck, van, other motorized vehicles and non-motorist.

In addition to the crash database, the explanatory attributes considered in the empirical study are also aggregated at the STAZ level. The selected explanatory variables can be grouped into five broad categories: land-use characteristics, roadway attributes, traffic characteristics, socioeconomic characteristics and demographic characteristics. These variables are collected and compiled from different data sources including: 2010 US census data, 2009-2013 American Community Survey (ACS), Florida Geographic Data Library (FDGL) databases. *Land-use characteristics* included shopping centers, restaurants, park/recreational centers and proportion of urban area. *Roadway attributes* included proportion of local roads and proportion of major roads length. *Traffic characteristics* included annual average daily traffic (AADT) and truck AADT. *Socioeconomic characteristics* included proportion of industrial jobs, proportion of retail jobs, proportion of households with no vehicle and proportion of households with one vehicle. Finally, *Demographic characteristics* included proportion of Hispanic population and proportion of Caucasian population.

Table 1 offers a summary of the sample characteristics of the exogenous factors in the estimation dataset along with the descriptive statistics of the dependent variables. The table represents the definition of variables considered for final model estimation along with the zonal minimum, maximum, average values and standard deviation. The final specification of the model development was based on removing the statistically insignificant variables in a systematic process based on statistical significance (95% significance level). The specification process was also guided by prior research and parsimony considerations. In estimating the models, several functional forms and variable specifications were explored. The functional form that provided the best result was used for the final model specifications and, in Table 1, the variable definitions are presented based on these final functional forms.

TABLE 1 Sample Statistics for the State of Florida

Variable names	Variable description	Zonal			
		Minimum	Maximum	Mean	Standard deviation
DEPENDENT VARIABLES					
Car crashes per STAZ	Total number of car involved crashes per STAZ	0.000	628.000	22.487	45.824
Light truck crashes per STAZ	Total number of light truck involved crashes per STAZ	0.000	81.000	4.237	7.561
Van crashes per STAZ	Total number of van involved crashes per STAZ	0.000	62.000	1.859	3.833
Other motorized vehicle crashes per STAZ	Total number of other motorized vehicle involved crashes per STAZ	0.000	148.000	3.195	6.102
Non-motorist crashes per STAZ	Total number of non-motorist involved crashes per STAZ	0.000	187.000	6.557	12.830
INDEPENDENT VARIABLES					
Land-use characteristics					
Shopping centers	Count of shopping centers in STAZ/10	0.000	18.900	0.583	1.049
Restaurants	Count of restaurants in STAZ/10	0.000	11.000	0.410	0.679

Park and recreational centers	Count of park and recreational centers in STAZ/10	0.000	5.400	0.079	0.167
Proportion of urban area	Total urban Area in STAZ/Total area of STAZ	0.000	1.000	0.731	0.425
Roadway attributes					
Proportion of local roads	Length of local roads in STAZ/Total length of roads in STAZ	0.000	1.000	0.085	0.192
Proportion of major roads	Length of major road in STAZ/Total length of roads in STAZ	0.000	1.000	0.544	0.371
Traffic characteristics					
AADT	Total annual average daily traffic (AADT) of STAZ/100,000	0.000	6.044	0.721	0.732
Truck AADT	Total truck AADT of STAZ/100,000	0.000	0.611	0.028	0.044
Socioeconomic characteristics					
Proportion of industrial jobs	Total number of industrial jobs in STAZ/Total number of jobs in STAZ	0.000	1.000	0.506	0.484
Proportion of retail jobs	Total number of retail jobs in STAZ/Total number of jobs in STAZ	0.000	1.000	0.276	0.277
Proportion of households with zero vehicle	Total number of households with no vehicles in STAZ/Total number of households in STAZ	0.000	1.000	0.094	0.111
Proportion of households with one vehicle	Total number of households with one vehicle in STAZ/Total number of households in STAZ	0.000	0.967	0.411	0.135
Demographic characteristics					
Proportion of Hispanic population	Total number of Hispanic population in STAZ/Total number of population in STAZ	0.000	1.000	0.171	0.202
Proportion of Caucasian population	Total number of Caucasian population in STAZ/Total number of population in STAZ	0.000	1.000	0.621	0.278

EMPIRICAL ANALYSIS

Model Selection

The empirical analysis involves estimation of four different multivariate count models including: 1) Clayton, 2) Gumbel, 3) Frank, and 4) Joe copulas. We also estimate an independent copula model (separate NB models for crash counts involving different road user groups) to establish a benchmark for comparison. A comparison exercise is undertaken to identify the most suitable copula model (including the independent copula model). The alternative copula models estimated are non-nested and hence cannot be tested using traditional log-likelihood ratio test. We employ the Bayesian Information Criterion (BIC) to determine the best model among all copula models

(*12; 21; 29; 32*). The model with the <u>*lower*</u> BIC is the preferred copula model. The BIC value for independent copula model is 192617.83. The BIC values for the estimated multivariate copula models are: Clayton - 166929.78, Gumbel - 165722.37, Frank - 167869.31, and Joe - 167572.03. From the BIC values, we can see that the estimated copula models provide improved data fit relative to independent model. However, copula model with Gumbel distribution outperforms all other copula models. The BIC comparisons confirm the importance of accommodating dependence among crash count events of different road user groups in the macro-level analysis.

Estimation Results

In presenting the effects of exogenous variables in the multivariate model specification, we will restrict ourselves to the discussion of the Gumbel Copula specification. The estimation results of the multivariate (Gumbel Copula) model are presented in Table 2.We include car, light truck, van, other motorized vehicle and non-motorist crash count components in second, third, fourth, fifth and sixth column panels, respectively. For brevity, results are discussed together for different road user groups in the following section by variable groups.

TABLE 2 Multivariate Count Model Estimation Results – Gumbel Copula

Variable names	**Car**	**Light truck**	**Van**	**Other motorized vehicle**	**Non-motorist**
	Estimate	**Estimate**	**Estimate**	**Estimate**	**Estimate**
	t-stat	**t-stat**	**t-stat**	**t-stat**	**t-stat**
Constant	1.311	0.680	-0.571	0.051	0.474
	23.940	21.571	-14.563	2.042	16.514
Land-use characteristics					
Shopping centers	0.159	0.097	--*	--	0.140
	8.402	6.575	--	--	9.180
Restaurants	--	--	0.080	0.127	--
	--	--	4.297	7.374	--
Park and recreational centers	0.541	0.347	0.427	--	0.599
	6.941	5.092	5.584	--	8.579
Proportion of urban area	0.570	0.074	0.562	--	0.444
	20.962	3.381	20.352	--	17.996
Roadway attributes					
Proportion of local roads	--	--	-0.298	--	--
	--	--	-4.928	--	--
Proportion of major roads	0.312	0.293	--	0.468	0.373
	10.994	10.810	--	17.745	13.489
Traffic characteristics					
AADT	0.851	0.660	0.643	0.553	0.809
	29.823	26.507	27.542	24.236	31.662
Truck AADT	-2.187	-2.225	--	2.876	-2.645
	-6.340	-7.676	--	10.052	-8.325

Socioeconomic characteristics					
Proportion of industrial jobs	-0.066 -3.335	0.148 7.527	-- --	-- --	-- --
Proportion of retail jobs	0.395 9.146	0.600 14.451	0.543 11.727	0.388 9.891	0.478 11.001
Proportion of households with zero vehicle	-- --	-- --	-- --	-- --	0.395 4.703
Proportion of households with one vehicle	0.752 9.442	-- --	0.715 9.130	0.888 15.494	-- --
Demographic characteristics					
Proportion of Hispanic population	0.645 9.434	0.124 2.971	-- --	-- --	0.578 11.405
Proportion of Caucasian population	-0.143 -3.473	-- --	-- --	-- --	-- --
Overdispersion parameter	1.864 76.517	1.877 63.394	2.028 48.290	1.887 58.978	1.840 64.363
Correlation parameter	3.395 122.559				

*variable insignificant at 90% significance level

Land-use characteristics: Among different points of interest considered, the copula model results reveal a higher probability of car, light truck and non-motorists crashes in the STAZs with higher number of shopping centers. The results indicate that the presence of more restaurants in a STAZ is positively associated with van and other motorized vehicle crashes. The results associated with parks/recreational centers show positive association with road traffic related crash risk. However, the variable effect is not significant in NB model component for other motorized vehicle crashes. Further, proportion of urban area is found to be significant in the count model component for car, light truck, van and non-motorist. In zones with larger proportion of urban areas, higher traffic related crashes are likely to occur, plausibly indicating higher density of transport activities and in turn higher traffic conflicts within an urbanized environment (see (*34; 35*) for similar results).

Roadway attributes: In the crash count component for van, we find that in the presence of more local roads in a STAZ, the possibility of crash risk for van decreases. At the same time, the results associated with roadway class show that car, non-motorist, light truck and other motorized vehicles' crash risks are positively correlated with higher proportion of major road.

Traffic characteristics: With respect to traffic characteristics, both AADT and truck AADT are found to have significant influence on crash occurrences for different road user groups. The model estimation results indicate that traffic related crashes are positively associated with higher AADT at the zonal level for both motorists and non-motorists road user groups. The result is in line with previous studies and can be attributable to higher exposure and/or adaptation of road users to different levels of traffic volume (see (*14; 36; 37*) for similar results). Further, the effect of zonal level truck AADT has significant influence on all NB model components other than van.

The model results reveal that the higher truck AADT at the zonal level are likely to reduce crash propensities for car, light truck and non-motorists (see (*38*) for similar result). The result may be explained by the overall cautious travel behavior of different road user groups in the presence of high heavy vehicle volume. On the other hand, the model estimation shows a positive correlation between truck AADT and crashes involving other motorized vehicles; perhaps indicating higher exposure of heavy vehicle and bus in these zones.

Socioeconomic characteristics: In terms of proportion of jobs by industry, the result associated with industrial jobs indicates that zones with higher proportion of industrial jobs increases the likelihood of light trucks' crash risk. It is likely that zones with higher proportion of industrial jobs experience higher usage of light truck for industrial job related activities. On the other hand, an increase in proportion of industrial jobs in a zone decreases the likelihood of crash risk for auto group of road users. Zones with higher number of retail jobs are likely to result in higher traffic crashes involving both motorized and non-motorized road user groups. Levine et al. (*39*) found similar impact of retail jobs on total crash count events. Further, non-motorists' crash risks are found to be higher for STAZs with higher proportion of households without access to private vehicles. The variable is a surrogate indicator for non-motorists exposure. Household members with no private vehicles are likely to walk/bike for daily activities resulting in higher exposure and in turn higher potential of crash risk. As expected, car, van and other motorized vehicle crash risks are found to be higher for STAZs with higher proportions of households with one available private vehicle.

Demographic characteristics: From Table 2, we can see that proportion of zonal level population by ethnicity are found to be significant determinants of zonal level crash risk in count model components for car, light truck and non-motorist crashes. Road traffic crashes for car, light truck and non-motorist increase with increasing proportion of Hispanic population, a result also observed in Lee et al. (*40*). On the other hand, the estimation results indicate that STAZs with greater proportion of Caucasian population are likely to experience less auto crashes.

Dependence effect: As indicated earlier, the estimated Gumbel copula-based multivariate NB model provides the best fit in incorporating the correlation among different road user groups' crash count events. An examination of the copula parameter presented in the last row panel of Table 2 highlights the presence of common unobserved factors affecting zonal level crash counts of different road user groups considered in current study context. For the Gumbel copula, the dependency is entirely positive and the coefficient sign and magnitude reflects whether a variable increases or reduces the dependency and by how much. The proposed framework by allowing for such correlation allows us to improve data fit.

Predictive Performance Evaluation and Policy Analysis

In an effort to assess the predictive performance of the estimated models (Gumbel copula and independent models), we also perform computation of several in-sample goodness-of-fit measures. To evaluate the predictive performance of these models, we employ two different fit measures: mean prediction bias (MPB) and mean absolute deviation (MAD) both at the aggregate and disaggregate level (see (*10*) for a discussion on computing these measures). At the aggregate/disaggregate level, the computed values of MPB (MAD) for copula and independent models are 4.890 (9.071)/24.448 (45.356) and 9.529 (15.019)/47.645 (75.096), respectively. The resulting fit measures for comparing the predictive performance clearly indicate that multivariate copula count model offers superior predictions compared to independent count model both at the aggregate and disaggregate levels in the current study context.

The parameter effects of exogenous variables in Table 2 do not directly provide the magnitude of the effects on zonal level crash counts across different road user groups involved in crashes. For this purpose, we compute aggregate level "elasticity effects" for all the exogenous variables by using the Gumbel copula model estimates. We investigate the effect as the percentage change in the expected total zonal crash counts across different road user groups due to the change in exogenous variable. Road user group specific elasticities would allow us to identify policy measures targeting each group separately. However, it might also be useful in identifying contributions of exogenous variables on total crashes considering contributions from all count components. Total and group specific elasticity effects would allow us to prioritize the safety improvement programs based on the level (all groups need attention in a specific area) and type (a specific group needs attention in a specific area) of safety issues. Therefore, we also present the overall total crash elasticities in our current study. Total crash elasticities are computed by considering the change in exogenous variables across all count components simultaneously.

The computed elasticities are presented in the first row panel of Table 3 (see (*41*) for a discussion on the methodology for computing elasticities). In calculating the expected percentage change of crash counts, we increase the value of variables by 10% for each STAZ. The numbers in Table 3 may be interpreted as the percentage change in the expected total zonal crash counts due to the change in exogenous variable. For instance, the elasticity effects for shopping centers in car model for in-sample data indicates that, the expected mean car crashes will increase by 3.074% with an increase in 10% of shopping centers. To emphasize policy repercussions based on most critical contributory factors, we also rank each variables based on their contribution in increasing the elasticity effects – with 1 as the highest contributor and 14 as the lowest contributor across different variables considered. The results of this ranking is presented in second row panel of Table 3.

The following observations can be made based on the results presented in Table 3. First, the most significant variable in terms of increase in the expected number of crashes across all road user groups is AADT, which is also the most important contributor for overall road traffic crashes. Second, the ranking of variables are different across the different road user groups, which illustrating that the relative contributions of different exogenous variables are substantially different across different road user groups. This has important implications in identifying critical factors for crash occurrences at a zonal level. For instance, targeted policy measures should be implemented to reduce overall crashes for zones with higher AADT. However, to improve car, van and other motorized vehicle safety, zones with higher proportions of households with one vehicle should be the major focus. On the other hand, to improve non-motorists safety, zones with higher shopping centers should be targeted. While for improved safety of light truck traffic, zones with higher proportion of retail jobs should be the focus. Moreover, the results indicate that there are considerable differences in the elasticity effects across different road user groups, thus illustrating the value of examining separate risk factors for different road user groups. Third, the impacts, in magnitude, are substantially different in crash count events among different passenger vehicles (car, light truck and van) for many variables. The effects are different in magnitude and sign for proportion of industrial jobs. These differences clearly highlight that each road user group has distinct critical risk factors underscoring the importance of examining the effect of various exogenous variables on zonal level crash count events by different road user groups. Finally, the elasticity analysis assists in providing a clear picture of attribute impacts on zonal level crash counts for different road user groups. The elasticity analysis conducted provides an illustration on how the proposed model can be applied to determine the critical factors contributing to increase in crash counts.

TABLE 3 Elasticity Effects for Multivariate Copula Count Model for Florida

ELASTICITY EFFECTS						
Variables	**Car**	**Light truck**	**Van**	**Other motorized vehicle**	**Non-motorist**	**Total**
AADT	17.371	9.483	9.986	8.467	15.055	60.362
Proportion of households with one vehicle	3.354	--	3.169	3.927	--	10.450
Shopping centers	3.074	1.193	--	--	2.400	6.668
Proportion of major roads	1.253	1.083	--	1.788	1.480	5.605
Proportion of retail jobs	0.946	1.260	1.156	0.826	1.108	5.296
Proportion of Hispanic population	1.541	0.240	--	--	1.313	3.095
Park and recreational centers	0.928	0.422	0.590	--	1.015	2.954
Restaurants	--	--	0.640	0.979	--	1.618
Proportion of urban area	0.207	0.030	0.225	--	0.171	0.634
Proportion of households with zero vehicle	--	--	--	--	0.382	0.382
Proportion of industrial jobs	-0.022	0.229	--	--	--	0.207
Proportion of local roads	--	--	-0.174	--	--	-0.174
Proportion of Caucasian population	-0.698	--	--	--	--	-0.698
Truck AADT	-1.324	-0.994	--	1.900	-1.451	-1.869

RANKING OF VARIABLES BASED ON CONTRIBUTION IN ELASTICITY EFFECTS						
Variables	**Car**	**Light truck**	**Van**	**Other motorized vehicle**	**Non-motorist**	**Total**
AADT	1	1	1	1	1	1
Proportion of households with one vehicle	2	9	2	2	9	2
Shopping centers	3	3	7	9	2	3
Proportion of major roads	5	4	8	4	3	4
Proportion of retail jobs	6	2	3	6	5	5
Proportion of Hispanic population	4	6	9	10	4	6
Park and recreational centers	7	5	5	7	6	7
Restaurants	9	10	4	5	10	8
Proportion of urban area	8	8	6	8	8	9
Proportion of households with zero vehicle	11	12	11	12	7	10
Proportion of industrial jobs	12	7	10	11	11	11
Proportion of local roads	10	11	14	14	13	12
Proportion of Caucasian population	13	13	12	13	12	13

Truck AADT	14	14	13	3	14	14

CONCLUSIONS

The current study contributes to safety literature both methodologically and empirically. In terms of methodology, the study generalized the bivariate copula count model for examining multivariate crash count data by formulating and estimating a multivariate copula count model. For examining the count components of the multivariate copula model, we employed negative binomial (NB) regression framework. The empirical contribution of our study was to incorporate crash counts for both motorized and non-motorized road user groups while considering different types of passenger vehicles fleet categories. Specifically, we examined crash counts for car, light truck, van, other motorized vehicle (including truck, bus and other vehicles) and non-motorist (pedestrian and bicyclist) involved crashes by employing multivariate copula count framework.

The proposed model was estimated using zonal level (Statewide Traffic Analysis Zone (STAZ) level) road traffic crash data for the state of Florida. A host of variable groups including – land-use characteristics, roadway attributes, traffic characteristics, socioeconomic characteristics and demographic characteristics were considered. The empirical analysis involved estimation of four different multivariate copula count models including: 1) Clayton, 2) Gumbel, 3) Frank, and 4) Joe copulas. The Gumbel copula model offered the most superior fit to our data. Further, the comparison between copula and the independent models confirmed the importance of accommodating dependence among crash count events of different road user groups in the macro-level analysis. Further, an in-sample validation exercise was conducted to compare the performance of the independent and copula model based on different fit measures. The resulting fit measures for comparing the predictive performance clearly indicate that multivariate copula count model offered superior predictions compared to independent count model both at the aggregate and disaggregate levels in the current study context.

In our research, to further understand the impact of various exogenous factors, aggregate level elasticity effects were computed for all the exogenous variables by using the estimates from multivariate copula-based count model. To emphasize policy repercussions based on most critical contributory factors, we also generated a rank for each variable based on their contribution in influencing crash frequency. The elasticity effects clearly indicated that there are considerable differences across different road user groups for the same variable, thus illustrating the value of examining separate risk factors for different road user groups. Further, the impacts were substantially different in crash count events among different passenger vehicles (car, light truck and van). The elasticity analysis conducted provides an illustration of how the proposed model can be applied to determine the critical factors contributing to increase in crash counts.

The study is not without limitations. In modeling zonal level crash risks, we did not have access to exposure data for different road user groups considered. It would be useful to compile zonal level exposure data for different motorized and non-motorized road user groups to enhance the model frameworks developed in our work. Further, it would be interesting to examine the count components within the copula-based multivariate approach by using Zero-inflated or Hurdle models in accommodating the preponderance of zero counts (if present) while also considering more flexible copula structures. A comparison of the proposed copula-based model with other multivariate modeling approaches will be a useful future research direction.

Acknowledgements
The authors would also like to gratefully acknowledge Florida Department of Transportation (FDOT) for providing access to Florida crash data.

REFERENCES

[1] Subramanian, R. Motor Vehicle Traffic Crashes as a Leading Cause of Death in the United States, 2002. *Young*, Vol. 1, 2005, p. 3.
[2] Blincoe, L., A. Seay, E. Zaloshnja, T. Miller, E. Romano, S. Luchter, and R. Spicer. The economic impact of motor vehicle crashes, 2000.In, Washington, DC, National Highway Traffic Safety Administration, 2002.
[3] Lord, D., and F. Mannering. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, Vol. 44, No. 5, 2010, pp. 291-305.
[4] Mannering, F. L., and C. R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic methods in accident research*, Vol. 1, 2014, pp. 1-22.
[5] Yasmin, S., and N. Eluru. Latent Segmentation Based Count Models: Analysis of Bicycle Safety in Montreal and Toronto. *Accident Analysis & Prevention (Forthcoming)*, 2016.
[6] Mannering, F. L., V. Shankar, and C. R. Bhat. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic methods in accident research*, Vol. 11, 2016, pp. 1-16.
[7] Chamberlain, G. Analysis of Covariance with Qualitative Data. *The Review of Economic Studies*, Vol. 47, No. 1, 1980, pp. 225-238.
[8] Bhat, C. R. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, Vol. 35, No. 7, 2001, pp. 677-693.
[9] Barua, S., K. El-Basyouny, and M. T. Islam. Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic methods in accident research*, Vol. 9, 2016, pp. 1-15.
[10] Yasmin, S., and N. Eluru. A joint econometric framework for modeling crash counts by severity. *Transportmetrica A: Transport Science*, 2017, pp. 1-26.
[11] Anastasopoulos, P. C. 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*, Vol. 11, 2016, pp. 17-32.
[12] Nashad, T., S. Yasmin, N. Eluru, J. Lee, and M. A. Abdel-Aty. Joint Modeling of Pedestrian and Bicylist Crashes: A Copula Based Approach.In *Transportation Research Record (Forthcoming)*, 2016.
[13] Heydari, S., L. Fu, L. F. Miranda-Moreno, and L. Jopseph. Using a flexible multivariate latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. *Analytic methods in accident research*, Vol. 13, 2017, pp. 16-27.
[14] Dong, C., D. B. Clarke, X. Yan, A. Khattak, and B. Huang. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. *Accident Analysis & Prevention*, Vol. 70, 2014, pp. 320-329.
[15] Huang, H., H. Zhou, J. Wang, F. Chang, and M. Ma. A multivariate spatial model of crash frequency by transportation modes for urban intersections. *Analytic methods in accident research*, Vol. 14, 2017, pp. 10-21.
[16] Jacobsen, M. R. Fuel economy and safety: The influences of vehicle class and driver behavior. *American Economic Journal: Applied Economics*, Vol. 5, No. 3, 2013, pp. 1-26.

[17] Alliance, A. *Light trucks help fuel 2016 sales*. https://autoalliance.org/economy/facts-about-auto-sales/. Accessed July, 2017.
[18] White, M. J. The “arms race” on American roads: The effect of sport utility vehicles and pickup trucks on traffic safety. *The Journal of Law and Economics,* Vol. 47, No. 2, 2004, pp. 333-355.
[19] Bhat, C. R. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological,* Vol. 45, No. 7, 2011, pp. 923-939.
[20] Narayanamoorthy, S., R. Paleti, and C. R. Bhat. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. *Transportation Research Part B: Methodological,* Vol. 55, 2013, pp. 245-264.
[21] Yasmin, S., N. Eluru, A. R. Pinjari, and R. Tay. Examining driver injury severity in two vehicle crashes – A copula based approach. *Accident Analysis & Prevention,* Vol. 66, 2014, pp. 120-135.
[22] Bhat, C. R., and N. Eluru. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological,* Vol. 43, No. 7, 2009, pp. 749-765.
[23] Wang, C., M. A. Quddus, and S. G. Ison. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis & Prevention,* Vol. 43, No. 6, 2011, pp. 1979-1990.
[24] Ayuso, M., L. Bermúdez, and M. Santolino. Copula-based regression modeling of bivariate severity of temporary disability and permanent motor injuries. *Accident Analysis & Prevention,* Vol. 89, 2016, pp. 142-150.
[25] Nikoloulopoulos, A. K., and D. Karlis. Regression in a copula model for bivariate count data. *Journal of Applied Statistics,* Vol. 37, No. 9, 2010, pp. 1555-1568.
[26] Cameron, A. C., T. Li, P. K. Trivedi, and D. M. Zimmer. Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal,* Vol. 7, No. 2, 2004, pp. 566-584.
[27] Nikoloulopoulos, A. K., and D. Karlis. Modeling multivariate count data using copulas. *Communications in Statistics-Simulation and Computation,* Vol. 39, No. 1, 2009, pp. 172-187.
[28] Yasmin, S., N. Eluru, A. R. Pinjari, and R. Tay. Examining driver injury severity in two vehicle crashes–A copula based approach. *Accident Analysis & Prevention,* Vol. 66, 2014, pp. 120-135.
[29] Trivedi, P. K., and D. M. Zimmer. Copula modeling: an introduction for practitioners. *Foundations and Trends® in Econometrics,* Vol. 1, No. 1, 2007, pp. 1-111.
[30] Sklar, A. Random variables, joint distribution functions, and copulas. *Kybernetika,* Vol. 9, No. 6, 1973, pp. (449)-460.
[31] Sener, I. N., N. Eluru, and C. R. Bhat. On jointly analyzing the physical activity participation levels of individuals in a family unit using a multivariate copula framework. *Journal of Choice Modelling,* Vol. 3, No. 3, 2010, pp. 1-38.
[32] Eluru, N., R. Paleti, R. Pendyala, and C. Bhat. Modeling injury severity of multiple occupants of vehicles: Copula-based multivariate approach. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2165, 2010, pp. 1-11.
[33] FIRES. *Quick Stats: Crash summary.* Florida's Integrated Report Exchange Sytem (FIRES). https://firesportal.com/Pages/Public/QuickStats.aspx. Accessed July, 2017.

[34] Huang, H., M. Abdel-Aty, and A. Darwiche. County-level crash risk analysis in Florida: Bayesian spatial modeling. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2148, 2010, pp. 27-37.
[35] Siddiqui, C., M. Abdel-Aty, and K. Choi. Macroscopic spatial analysis of pedestrian and bicycle crashes. *Accident Analysis & Prevention*, Vol. 45, 2012, pp. 382-391.
[36] Milton, J. C., V. N. Shankar, and F. L. Mannering. Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis & Prevention,* Vol. 40, No. 1, 2008, pp. 260-266.
[37] Hadayeghi, A., A. Shalaby, and B. Persaud. Macrolevel accident prediction models for evaluating safety of urban transportation systems. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1840, 2003, pp. 87-95.
[38] Cai, Q., M. Abdel-Aty, J. Lee, and N. Eluru. Comparative analysis of zonal systems for macro-level crash modeling. *Journal of Safety Research,* Vol. 61, No. Supplement C, 2017, pp. 157-166.
[39] Levine, N., K. E. Kim, and L. H. Nitz. Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators. *Accident Analysis & Prevention,* Vol. 27, No. 5, 1995, pp. 675-685.
[40] Lee, J., M. Abdel-Aty, and X. Jiang. Development of zone system for macro-level traffic safety analysis. *Journal of transport geography,* Vol. 38, 2014, pp. 13-21.
[41] Eluru, N., and C. R. Bhat. A joint econometric analysis of seat belt use and crash-related injury severity. *Accident Analysis & Prevention,* Vol. 39, No. 5, 2007, pp. 1037-1049.