Do We Need Multivariate Modeling Approaches to Model Crash Frequency by Crash Types? A Panel Mixed Approach to Modeling Crash Frequency by Crash Types

Tanmoy Bhowmik*
Doctoral Student
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

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
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
ORCiD number: 0000-0003-1221-4113

*Corresponding author

ABSTRACT

In safety literature, simulation-based multivariate framework is the most commonly employed approach for analyzing multiple crash frequency dependent variables. The current research effort contributes to literature on crash frequency analysis by suggesting an alternative and mathematically simpler approach for analyzing multiple crash frequency variables for the same study unit. The proposed recasts a multivariate distributional problem as a repeated measure univariate problem. Specifically, we employed a simpler panel random parameter based univariate model framework to analyze zonal level crash counts for different crash types. 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 performance of the proposed framework is compared with the performance of the random parameter multivariate negative binomial model (RPMNB) using a host of metrics for estimation sample and hold-out sample. The resulting goodness of fit and predictive measures clearly highlight the comparable performance offered by the proposed framework relative to the commonly used RPMNB model with substantially fewer parameters. The comparison exercise is augmented by computing aggregate level elasticity effects for both PMNB and RPMNB models. The results clearly highlight the comparable performance offered by the proposed PMNB model relative to the traditional RPMNB model. In summary, the proposed framework allows for a parsimonious specification without compromising the model explanatory power and provides similar performance as the most traditional multivariate NB model for analyzing different crash dimensions.

Keywords: Unobserved factors; panel univariate model; Multivariate negative binomial framework; Crash type.

1 INTRODUCTION

1.1 Motivation

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

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

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

1.2 Focus of the Current Study

Our proposed research attempts to contribute to simulation-based multivariate approaches by altering how the multiple dependent variables are analyzed. Prior to presenting our alternative approach, challenges with the current simulation-based multivariate approaches in estimating observed and unobserved variable effects are discussed. In multivariate approaches, a separate crash propensity equation is adopted for each crash type. Thus, if there are *D* dependent variables and *K* independent variables, the order of observed parameters estimated in the model structure is of the order of *D*K*. With increasing number of dimensions (*D)*, the number of parameters to be estimated increase rapidly. Thus, in models with *D* >3, the number of parameters to be estimated are prohibitively high. For example, consider a case of crash frequency for four crash types at an intersection (rear-end, side-swipe, angle and non-motorized). In the univariate models, for each of the crash types, Annual Average Daily Traffic (AADT) is likely to have a statistically significant impact. So, the typical multivariate model estimates 4 parameters for AADT. However, it is possible that the impact of AADT on side-swipe and angle crashes is not statistically different. Testing this is not straightforward in the multivariate model structure. The analyst will need to modify the model estimation code to restrict the parameters across the side-swipe and angle univariate models to be the same. Subsequently, the restricted model version data fit must be compared with the data fit of the unrestricted version using log-likelihood ratio (LR) test. Based on the result, the analyst can conclude if AADT does offer different impacts for side-swipe and angle crash profiles. Given the additional burden of these steps, the models employed in safety literature typically ignore if the variable impacts are really different across crash type propensities. The result is an ill-specified model structure with too many parameters. To be sure, the model estimates thus obtained are not incorrect. However, the estimation process could become inefficient particularly when sample sizes for crash frequency are small (<1000). The sample sizes for micro-level analysis can typically vary from 200-500 and the number of total parameters estimated has an impact of model estimation efficiency.

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

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

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

The rest of the paper is organized as follows: The next section presents the methodological framework adopted in the analysis while the third section provides a detailed description of the dataset. Model findings are discussed in the fourth section followed by the concluding thoughts in the last section.

2 METHODOLOGY

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

2.1 Random Parameter Multivariate NB Model

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

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

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

$$\mu_{ij} = E\left(c_{ij} | \boldsymbol{z}_{ij}\right) = exp\left((\boldsymbol{\delta}_j + \boldsymbol{\zeta}_{ij})\boldsymbol{z}_{ij} + \varepsilon_{ij} + \eta_{ij}\right) \tag{2}$$

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

$$\eta_{ij} = \boldsymbol{\gamma_j} \boldsymbol{s}_{ij} \tag{3}$$

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

In examining the model structure of crash count across different crash types, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\zeta}$ and $\boldsymbol{\gamma}$ represented by $\boldsymbol{\Omega}$. In this 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:

$$L_i = \int_{\boldsymbol{\Omega}} \prod_{j=1}^{J} \left(P(c_{ij})\right) f(\boldsymbol{\Omega}) d\boldsymbol{\Omega} \tag{4}$$

Finally, the log-likelihood function is:

$$LL = \sum_i Ln(L_i) \tag{5}$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 5. The parameters to be estimated in the multivariate NB model are: $\boldsymbol{\delta}_j$, α_j, $\boldsymbol{\pi}_j$, and $\boldsymbol{\sigma_j}$.

2.2 Panel Mixed NB Model

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

$$P(y_{ir}|v_{ir}, \lambda') = \frac{\Gamma\left(y_{ir} + \frac{1}{\lambda'}\right)}{\Gamma(y_{ir} + 1)\Gamma\left(\frac{1}{\lambda'}\right)} \left(\frac{1}{1 + \lambda' v_{ir}}\right)^{\frac{1}{\lambda'}} \left(1 - \frac{1}{1 + \lambda' v_{ir}}\right)^{y_{ir}} \tag{6}$$

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

$$v_{ir} = E(y_{ir}|\boldsymbol{x}_{ir}) = exp((\boldsymbol{\beta} + \boldsymbol{\theta}_i + \boldsymbol{\varrho}_{ir})\boldsymbol{x}_{ir} + \varepsilon_{ir}) \tag{7}$$

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

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

$$L_i = (\int_{\Psi} \prod_{r=1}^{R} \left(P(y_{ir})\right) f(\Psi) d\Psi \tag{8}$$

Finally, the log-likelihood function is:

$$LL = \sum_{i} Ln(L_i) \tag{9}$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 9.

3 DATA PREPARATION

Our study area, Central Florida region is composed of 4,747 TAZs. The study is focused on crashes involving both motor vehicles and non-motorists at a zonal level for the year 2016. The data are compiled from Florida Department of Transportation (FDOT), Crash Analysis Reporting 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; within the motorized group, the records are further classified into five categories based on the manner of crash: rear-end, angular, sideswipe, all single vehicle and other multiple vehicle crashes. Based on the crash records, crashes of different types are combined together as one category: left-turn, right-turn and angular crashes within angular class; off-road, rollover and other single vehicle in the all single vehicle category; and head-on and other multiple vehicle crashes are classified as the other multiple vehicle crash types. All the crash records are aggregated at a TAZ level using the Geographic Information System (GIS). A total of 114,458 motorized and 3,413 non-motorized crashes were reported in the Central Florida region for the year 2016. Within the motorized crashes, rear-end is found to be the most prevalent crash type (44.09%) while sideswipe is less frequent with 10.82% among all other motorized crash types. Crash statistics at a zonal level for different crash types are summarized in Table 1. From the total records, for the validation analysis, we set aside data from 932 TAZs and the remaining 3,815 TAZs are used for the estimation analysis.

3.1 Variables Considered

A comprehensive set of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics are considered in the current research effort. Information about these variables are collected from different data sources including FDOT

Transportation Statistics Division, US Census Bureau, American Community Survey and Florida Geographic Data Library databases. Similar to the 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. While the information about land use category including area of urban, residential, industrial, institutional, recreational, office and land use mix are provided in the land use attributes, built environment characteristics mainly reflects the information about the number of business center, commercial center, school, hospital, recreational center, restaurant and shopping center are collected. 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) proportion of people by age and race and proportion of household by vehicle ownership level (1, 2, 3 and 4 or more) are included.

Table 2 summarizes sample characteristics of the explanatory variables with the appropriate definition considered for final model estimation along with the minimum, maximum and mean values at a zonal level. In estimating the model, several functional forms, combination of variables and interaction terms are considered and those that provides the best fit are retained in the final specification. The final specification of the model was based on removing the statistically insignificant variables in a systematic process based on 90% confidence level.

4 EMPIRICAL ANALYSIS

4.1 Model Specification and Overall Measure of Fit

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

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

4.2 Model Estimation Result

This section presents a detailed discussion of the factors affecting crash count components across different crash types. Table 3 presents the model estimation results for the proposed panel mixed NB model. The estimation results of the multivariate NB model are presented in Table 4 for comparison. For the sake of brevity, we do not discuss these parameter estimates.

As discussed before, in presenting our model results, we have selected a representation that provides results similar to the traditional model approach i.e. present the net effect of each exogenous variable in the crash propensity equation. For example, consider the constants estimated in the various crash type propensity equations. The proposed estimated the base effect as -1.074 and the deviations across crash type as – rear-end 0.000, angular -0.716, Sideswipe -0.907, All single vehicle 2.137, Other multiple vehicle -1.172, and Non-motorized -2.109. The reader would note that the "rear-end" crash type served as the base. The model results presented compute the net effect for each crash type. For non-motorized crash type this would be computed as -1.704 (base) + -2.109 (non-motorized deviation) = -3.841. The consolidation of parameters in this manner allows an easy comparison with the traditional approach. The consolidation of parameters in this manner allows an easy comparison with the traditional approach. At the same time, to highlight the gains in parameters if any, we identify the number of parameters estimated across the crash types (range between 1 and 6). In cases where the deviation for a crash type was insignificant, the reader would notice a common coefficient across 2 or more crash types. The number of distinct parameters estimated provides a guide to the improvement in model estimation attained by the proposed model structure. For instance, the variable length of divided roads offers an important comparison across the two models (see Table 3 and 4). In our proposed model, we estimated a single parameter across 5 crash types while the same variable results in five distinct parameters across 5 crash types in the traditional multivariate model. The variable impact illustrates how our proposed approach allows for parsimonious specification while not compromising on model explanatory power. Finally, the reader would note that for some exogenous variables, a common base effect might not be statistically significant. In such cases, the exogenous variable is considered by crash type to test for the variable impact.

A positive (negative) sign for a variable in the crash count component of Table 3 indicates that an increase in the variable is likely to result in more (less) crashes.

4.2.1 Crash Specific Constants

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

4.2.1 Roadway Attributes

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

In terms of length of divided roads, the variable is found to have the same positive effect on all crash types except non-motorized crashes. Signal intensity in the zone reveals a negative association with sideswipe and all single vehicle specific crashes indicating a reduced occurrence of sideswipe and all single vehicle crashes in a zone with higher number of signals. This is expected because, vehicles are likely to drive at a lower speed in the location with higher number of signals and as a result, the risk of motorized off-road crashes reduces. Average outside shoulder width has a negative influence on crash risk propensity for rear-end, angular, sideswipe and other multiple vehicle crashes which is perhaps indicating greater safety margins for vehicular maneuverability. The estimated results show that a TAZ with higher proportion of roads over 55mph speed limit is likely to experience increased number of rear-end, sideswipe and all single vehicle crashes while a negative effect is observed for angular and non-motorized crashes. Further, we found that proportion of road over 55mph has significant variability specific to angular crashes as indicated by the standard deviation parameter. The reader would note that the distributional parameter indicates that the overall impact of the variable on angular crashes is likely to be negative (80%). With respect to sidewalk width, the variable is found to be significant in rear-end crash component with a positive impact while a negative association is observed for the non-motorized crashes. The results are contrary to some of the earlier studies (Aguero-Valverde and Jovanis, 2006; Cai et al., 2016; Dong et al., 2014). However, there is a reasonable explanation for the effects identified. Increasing sidewalk width is a surrogate for non-motorized activity in the zone. The presence of non-motorists can potentially increase rear-end crashes at as vehicles might stop abruptly to allow for non-motorist movement increasing rear-end crash risk. Also, the presence of a wider side walk provides additional margin of safety for non-motorists from colliding with a motorized vehicle and thus results in reduced risk for non-motorized users in the zone.

4.2.2 Traffic Characteristics

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

4.2.3 Land-use Attributes

From Table 3, we can observe that TAZs with higher urbanized and office areas are likely to experience more crashes specific to all crash types. This is expected as urban area serves as an additional surrogate for exposure for traffic. Moreover, the impact of urban area specific to rear-end crash is of higher magnitude relative to other crash types signifying that rear-end crash is a prominent safety issue in urban areas. Institutional areas are associated with increased crash risks for rear-end, angular, other multiple vehicle and non-motorized crash. The variable also illustrated the advantages of our proposed approach. Specifically, in our proposed framework, we estimate a total of two parameters for the variable. However, in the traditional multivariate structure, four distinct parameters were estimated. Residential area has a significant negative impact for rear-end, angular and sideswipe crashes.

4.2.4 Built Environment Characteristics

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

4.2.5 Sociodemographic Characteristics

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

4.2.7 Unobserved Heterogeneity

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

5 MODEL COMPARISON EXERCISE

5.1 Predictive Performance

In an effort to assess the predictive performance of the estimated models, we compute several goodness fit of measures at disaggregate level including MPB (Mean prediction bias), MAD (mean absolute deviation), MAPE (mean absolute percentage error), RMSE (Root mean square error) and predictive BIC (please see (Bhowmik et al., 2018) for a discussion on estimating these measures). Specifically, we employ these measure on two datasets: 1) in-sample dataset: for the records used in the model estimation (sample size = 3,815 TAZs) and 2) holdout sample: records that are set aside for validation analysis (sample size = 932 TAZs). The reader would note that model with lower value of predictive measures and BIC will reflect better performance in terms of prediction and statistical fit relative to the observed data. Table 5 presents the values of these measures for Random parameter multivariate NB and Panel mixed NB models for both in-sample and holdout-sample measures. From Table 5, we can observe that the performance of the two models across various prediction measures are quite similar even though there is a large difference in the number of parameters between the two specifications (92 vs 61). Further, RPMNB model performs marginally better than the proposed framework for the deviation measures with respect to angular, sideswipe, other multiple vehicle and non-motorized crashes while in terms of rear-end and all single vehicle crashes, the proposed approach offers better performance (for both in-sample and holdout samples). These deviation measures do not consider the difference in number of parameters across the two models. The BIC measure that penalizes additional parameters clearly shows that the proposed panel model structure offers improved statistical fit. In summary, the resulting goodness of fit measures clearly highlight the comparable performance offered by the proposed framework compared to the commonly used RPMNB model even with substantially fewer parameters.

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

[1] The same correlation structure was revealed from the traditional multivariate model structure (as shown in Table 4).

and we predict 60 and 50 TAZs from RPMNB and PMNB model respectively, then the estimated ratio of these models will be 0.6 (60/100) and 0.5 (50/100) respectively. For both models, two box plots are generated using all the data samples (for every count event, there are 20 points) by each count group and crash type. Figure 1a to 1c represents the ratio statistics for different crash types. From Figure 1, we can see that while the models might under-predict or over-predict crash counts, the performance of the two models are quite similar. Thus, one can conclude that the proposed approach has offered equivalent predictions relative to the multivariate NB model despite with substantially fewer model parameters (31 less parameters to be precise).

5.2 Elasticity Effects

The parameters of the exogenous variables in Table 3 and 4 do not directly provide the exact magnitude of the effects of variables on the zonal level crash counts across different crash types. However, it might be possible that the effects (exact magnitude) of some attributes could differ considerably across the two frameworks. To evaluate this, we compute aggregate level elasticity effects for both PMNB and RPMNB models. For this purpose, we identify a subset of exogenous variables including proportion of arterial roads, length of divided roads, proportion of roads over 55mph, institutional areas and number of non-motorist commuters. In our study, we investigate the effect as percentage change in the expected zonal level crash counts in response to the increase of the explanatory variable by 10% (see Eluru and Bhat, 2007 for a discussion on the methodology for computing elasticities). The numbers in Figure 2 can be interpreted as the percentage change in the expected crash counts (increase for positive sign and decrease for negative sign) due to the change in the exogenous variable for different crash types. For instance, the elasticity estimates generated from the proposed PMNB (RPMNB) model for proportion of arterial roads variable in rear-end crashes indicates that the expected mean rear-end crash will increase by 0.656% (1.038%) for an 10% increase in the proportion of arterial roads.

Several observations can be made based on the elasticity effects presented in Figure 2. First, in general, we do not observe any large differences in the elasticity effects of the two models across different crash types. From the five variables considered for our elasticity exercise, a substantial number of the effects (14 out of 22) offer very little differences. Second, the PMNB model with fewer parameters is able to represent the substantial differences in the elasticity effects for the same variable across different crash types. For instance, the elasticity effect for length of divided roads variable is different across the five crash types despite estimating a single parameter (same impact in magnitude) for the variable across the five crash types. Third, for some variables, we found substantial differences in the elasticity effects across the two frameworks for different crash types. For example, in case of rear-end crashes, the proposed PMNB model predicts an 0.65% increase in the expected mean for 10% increase in the proportion of roads over 55mph while we found an increase of 0.92% from the RPMNB model. Such differences could be attributed to the non-linearity embedded within the two model structures estimated with similar data fit. In summary, the proposed framework allows for a parsimonious specification without compromising the model explanatory power and provides similar performance (most of the times) as the most traditional multivariate NB model.

6 CONCLUSIONS

The most common approach employed to address correlation across multiple crash frequency dependent variables in safety literature is the development of simulation-based multivariate frameworks. However, with higher dimensions, the multivariate model estimation infrastructure can get computationally demanding in terms of the number of observed and unobserved parameters to estimate. In this context, our proposed research attempts to contribute to simulation-based multivariate approaches by altering how the multiple dependent variables are analyzed. Specifically, instead of considering the crash frequency by crash type as a multivariate distribution, we represent it as a repeated measures of crash frequency while recognizing that each repetition represents a crash type specific to a zone. Thus, in this process we cast a multivariate distribution as a univariate distribution with repeated measures. The recasting allows us to estimate parsimonious model systems as well as simplify the specification process. This simplification leading to parsimonious specification can reduce the computational time for estimating parameters associated with unobserved factors. To the best of authors' knowledge, this study is the first of its kind to simplify current modeling infrastructure for multivariate analysis in safety literature.

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

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

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

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

Abdelwahab, H. and Abdel-Aty, M., 2002. Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transportation Research Record: Journal of the Transportation Research Board 1784, 115-125.

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.

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.

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.

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.

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.

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.

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., and Bhat, C.R., 2007. A joint econometric analysis of seat belt use and crash-related injury severity. Accident Analysis and Prevention, 39 (5), 1037–1049.

Eluru, N., Paleti, R., Pendyala, R. and Bhat, C., 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.

Geedipally, S., Patil, S. and Lord, D., 2010. Examination of methods to estimate crash counts by collision type. Transportation Research Record: Journal of the Transportation Research Board 2165 (1), 12-20.

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.

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.

Milton, J.C., Shankar, V.N. and Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40(1), 260-266.

Nashad, T., Yasmin, S., Eluru, N., Lee, J. and Abdel-Aty, M.A., 2016. Joint modeling of pedestrian and bicycle crashes: copula-based approach. Transportation Research Record: Journal of the Transportation Research Board 2601, 119-127.

National Highway Traffic Safety Administration, 2017. 2016 fatal motor vehicle crashes: overview. Traffic Safety Facts-Research Note (DOT HS 812 456). URL https://crashstats. nhtsa. dot. gov/Api/Public/Publication/812456 (accessed 7.26.2018).

Reynolds, C.C., Harris, M.A., Teschke, K., Cripton, P.A. and Winters, M., 2009. The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature. Environmental health 8(1), 47.

Wang, K., Ivan, J.N., Ravishanker, N. and Jackson, E., 2017. Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways. Accident Analysis and Prevention 99, 6-19.

Wang, K., Yasmin, S., Konduri, K.C., Eluru, N. and Ivan, J.N., 2015. Copula-based joint model of injury severity and vehicle damage in two-vehicle crashes. Transportation Research Record: Journal of the Transportation Research Board 2514, 158-166.

Wang, X. and Abdel-Aty, M., 2008. Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. Accident Analysis and Prevention 40(5), 1674-1682.

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.

Yan, X., Radwan, E. and Mannila, K.K., 2009. Analysis of truck-involved rear-end crashes using multinomial logistic regression. Advances in Transportation Studies 17, 39-52.

Yasmin, S., Momtaz, S.U., Nashad, T., Eluru, N., 2018. A Multivariate Copula-Based Macro-Level Crash Count Model. Transportation Research Record, 2672 (30), 64–75.

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

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.

LIST OF FIGURES

FIGURE 1a Predicted to Observed Ratio for Rear-end and Angular Crashes

FIGURE 1b Predicted to Observed Ratio for Sideswipe and All Single Vehicle Crashes.

FIGURE 1c Predicted to Observed Ratio for Other Multiple Vehicle and Non-motorized Crashes.

LIST OF TABLES

TABLE 1: Descriptive Statistics of Dependent Variables

TABLE 2: Summary Statistics of Exogenous Variables (Zonal Level)

TABLE 3: Panel Mixed NB Model Estimation Results

TABLE 4: Random Parameter Multivariate NB Model Estimation Results

TABLE 5: Predictive Performance Measures of Two Models

TABLE A.1: Independent NB Model Estimation Results

TABLE A.2: Independent Panel NB Model Estimation Results

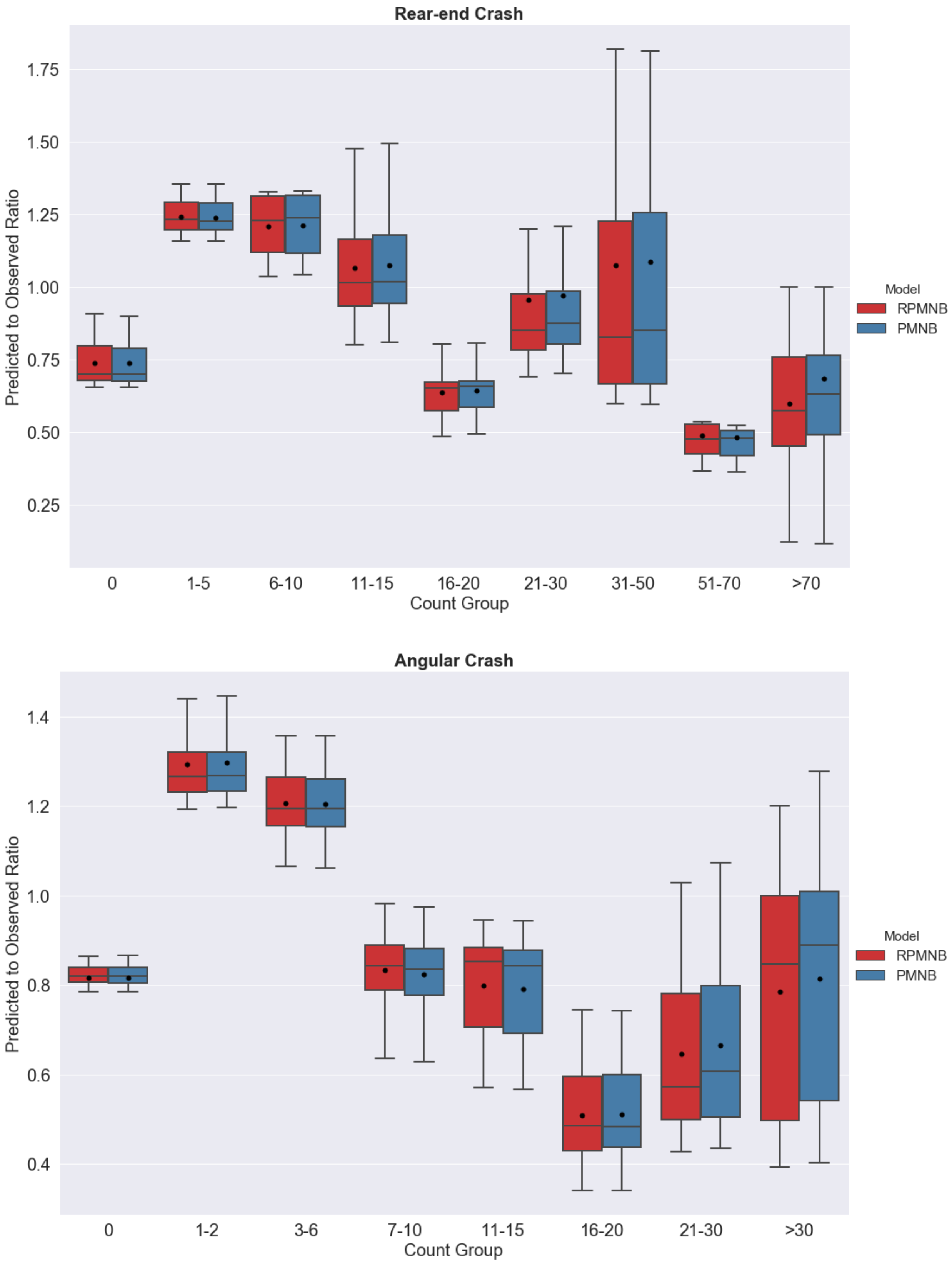

FIGURE 1a Predicted to Observed Ratio for Rear-end and Angular Crashes.

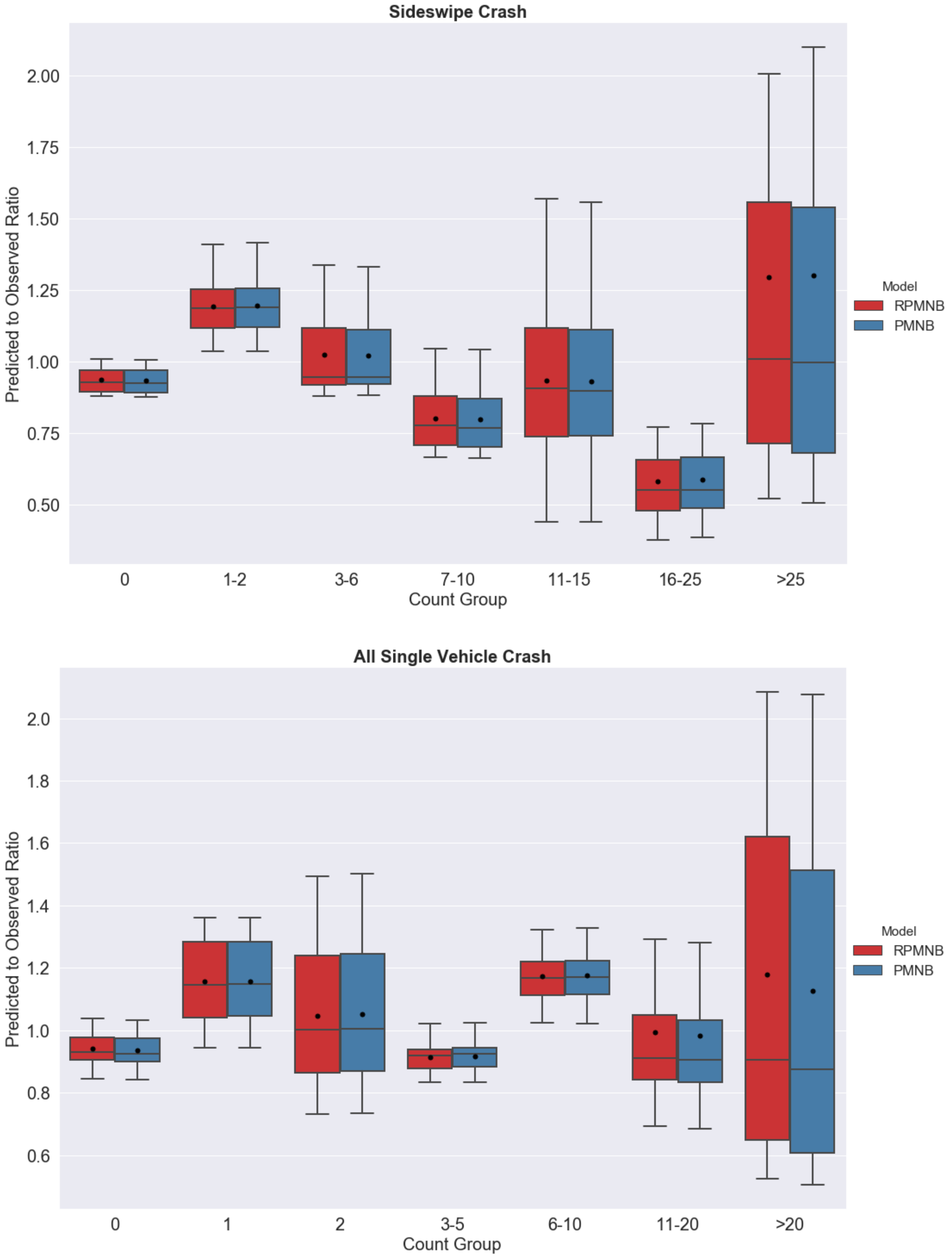

FIGURE 1b Predicted to Observed Ratio for Sideswipe and All Single Vehicle Crashes.

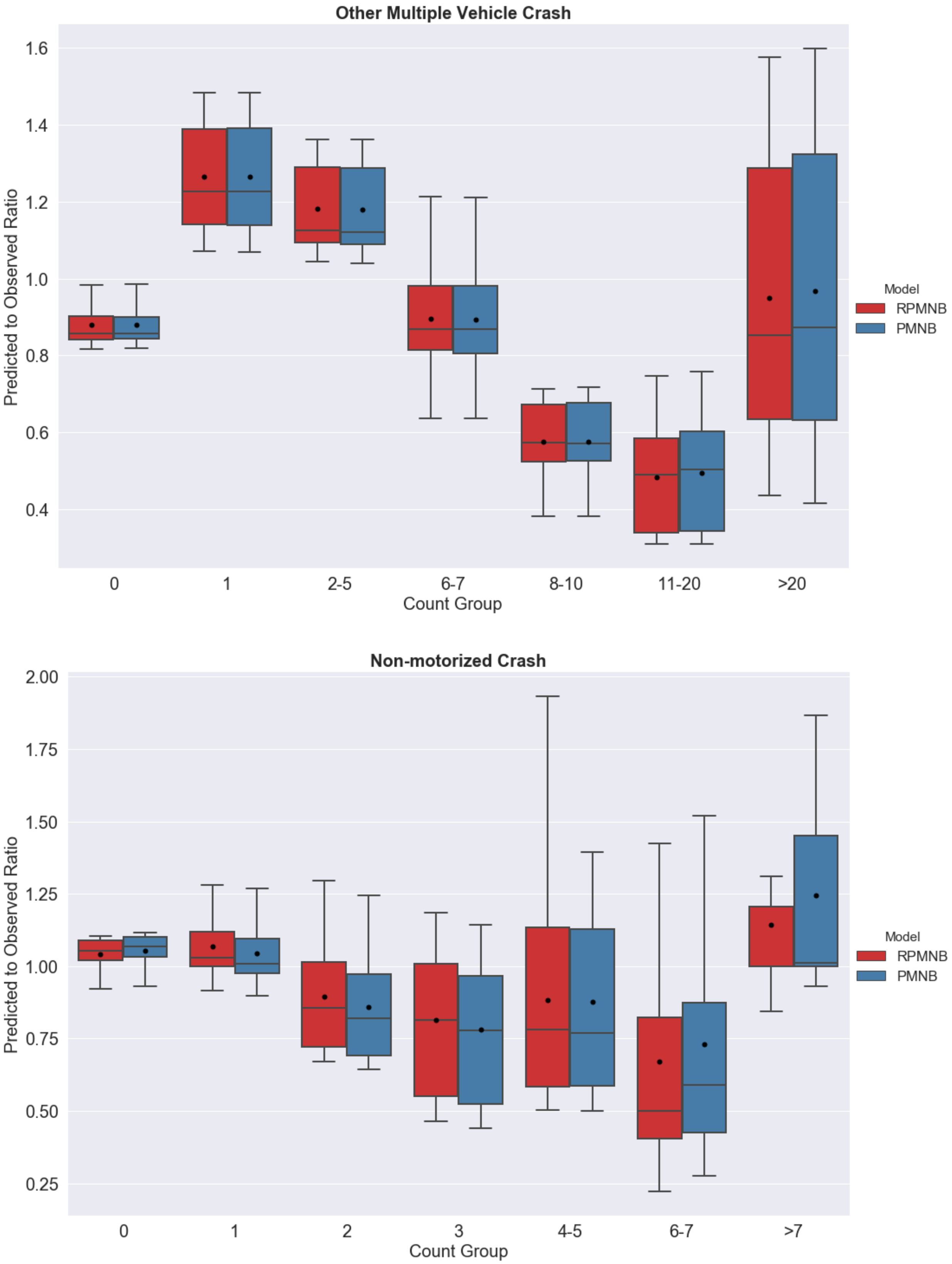

FIGURE 1c Predicted to Observed Ratio for Other Multiple Vehicle and Non-motorized Crashes.

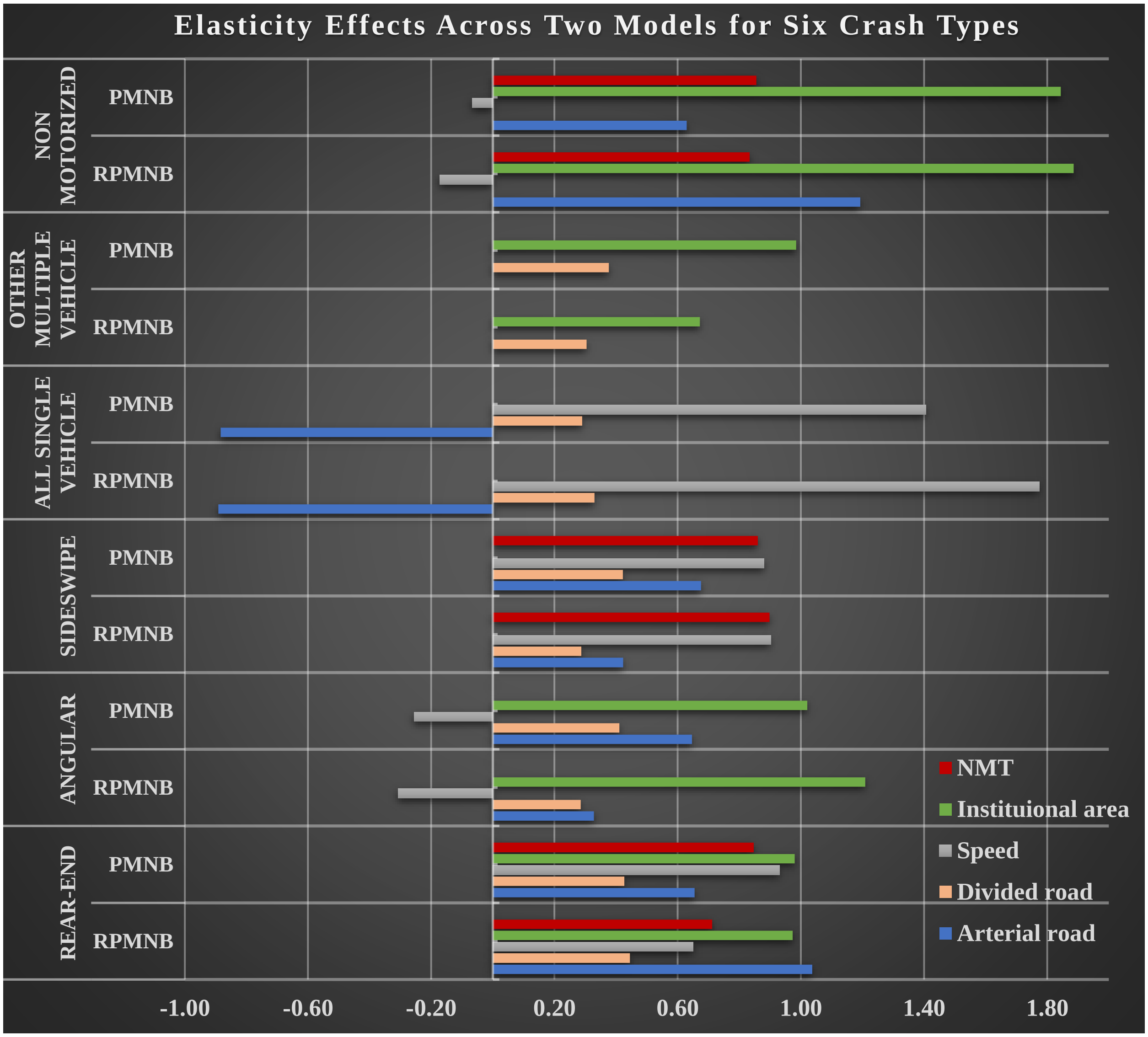

FIGURE 2 Elasticity Effects Across Two Models (PMNB and RPMNB) for Six Crash Types

TABLE 1 Descriptive Statistics of Dependent Variables

Variable Names	Definition	Zones (N=4,747)			
		Minimum	Maximum	Mean	Standard Deviation
Rear-end Crash (motorized)	Total number of rear-end crash (motorized) occurred in a TAZ	0.000	243.000	10.948	18.517
Angular Crash (motorized)	Total number of left turn, right turn and angular crash (motorized) occurred in a TAZ	0.000	104.000	4.216	6.817
Sideswipe Crash (motorized)	Total number of sideswipe crash (motorized) occurred in a TAZ	0.000	66.000	2.686	5.228
All Single Vehicle Crash (motorized)	Total number of off-road, rollover and other-single vehicle crash (motorized) occurred in a TAZ	0.000	62.000	3.317	4.480
Other-multiple Vehicle Crash (motorized)	Total number of head-on and other-multiple vehicle crash (motorized) occurred in a TAZ	0.000	112.000	2.945	4.549
Non-motorized Crash	Total number of non-motorized (pedestrian and bicycle) crash in a TAZ	0.000	12.000	0.719	1.318
Total Crash	Total number of crash (motorized and non-motorized) in a TAZ	0.000	413.000	24.831	35.326

TABLE 2 Summary Statistics of Exogenous Variables (Zonal Level)

Variables	Definition	Zonal (N=4,747)			
		Minimum	**Maximum**	**Mean**	**Std. Deviation**
Roadway Characteristic					
Proportion of rural roads	(Rural roads length/total road length)	0.000	1.000	0.121	0.309
Proportion of urban roads	(Urban roads 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[2]: No of business center	-0.138	19.664	0.000	1.000
No of commercial center	Z score: No of commercial center	-0.270	9.521	0.000	1.000
No of educational center	Z score: No of educational center	-0.487	11.610	0.000	1.000
No of recreational center	Z score: No of park and recreational center	-0.475	16.678	0.000	1.000
No of restaurant	Z score: No of restaurant	-0.464	11.021	0.000	1.000
No of shop	Z score: No of shopping center	-0.442	19.728	0.000	1.000
Traffic Characteristics					
VMT	Vehicle miles travelled	0.000	15.026	7.914	3.368
Truck VMT	Tuck vehicle miles traveled	0.000	13.049	3.474	2.864
Proportion of heavy 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	0.000	8.556	0.902	0.878
Average TAZ income	Ln (Average TAZ income+1)	0.000	12.534	11.065	0.386
Proportion of commuter	Total number of commuter/total population	0.000	0.778	0.408	0.085
Non-motorist commuter	Ln (NMT means to work for a TAZ)	0.000	5.261	1.278	1.098
Proportion of senior people	Total number of people over 65 years/total population in TAZ	0.000	0.821	0.206	0.114
Proportion of African-American people	Total number of African-American people /total population in TAZ	0.000	0.969	0.142	0.159
Proportion of household with no vehicle	Number of household with no vehicle/total household	0.000	0.471	0.069	0.065

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

TABLE 3 Panel Mixed NB Model (PMNB) Estimation Results

Variables[3]	No. of Param*	Rear End	Angular	Sideswipe	All single vehicle	Other multiple vehicle	Non-motorized
		Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)
Constant	6	-1.074 (-12.165)	-1.790 (-22.389)	-2.697 (-25.745)	-0.560 (-8.412)	-1.732 (-21.046)	-3.841 (-34.651)
Roadway Characteristics							
Proportion of arterial roads	2	0.134 (5.545)	0.134 (5.545)	0.134 (5.545)	-0.230 (-5.266)	--	0.134 (5.545)
Number of intersections	2	--[4]	0.305 (13.006)	--	--	0.173 (6.503)	0.305 (13.006)
Variance of speed	2	0.031 (6.845)	0.031 (6.845)	0.072 (6.379)	0.031 (6.845)	0.031 (6.845)	--
Length of divided roads	1	0.456 (6.346)	0.456 (6.346)	0.456 (6.346)	0.456 (6.346)	0.456 (6.346)	--
Signal intensity	1	--	--	-0.585 (-5.613)	-0.585 (-5.613)	--	--
Average outside shoulder width	3	-0.386 (-11.366)	-0.166 (-4.576)	-0.386 (-11.366)	--	-0.099 (-2.633)	--
Road length over 55mph	3	1.039 (21.596)	-0.516 (-4.090)	1.039 (21.596)	1.039 (21.596)	--	-0.139 (-1.717)
Standard deviation	1	--	0.622 (3.040)	--	--	--	--
Sidewalk width	2	0.089 (3.401)	--	--	--	--	-0.085 (-4.136)
Traffic Characteristic							
VMT	4	--	0.065 (7.910)	0.211 (21.727)	-0.118 (-5.491)	0.087 (9.454)	0.065 (7.910)
Truck VMT	2	0.209 (18.563)	--	--	0.332 (13.182)	--	--
Land-use attributes							
Urban area	5	0.173 (13.355)	0.125 (9.322)	0.134 (8.675)	0.060 (7.964)	0.092 (7.916)	0.173 (13.345)
Office area	4	0.234 (30.359)	0.234 (30.359)	0.234 (30.359)	0.083 (6.931)	0.169 (13.301)	0.161 (8.505)
Institutional area	2	0.063 (7.291)	0.063 (7.291)	--	--	0.063 (7.291)	0.109 (5.942)
Residential area	2	-0.085 (-14.223)	-0.023 (-2.668)	-0.085 (-14.223)	--	--	--
Built environment characteristic							
No. of restaurants	4	0.241 (19.756)	0.241 (19.756)	0.301 (17.306)	0.101 (5.017)	0.265 (19.626)	0.241 (19.756)

[3] Please see Table 3 for variable definitions and units

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

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

Note: *No. of Param = Number of parameters estimated for the corresponding variable.

TABLE 4 Random Parameter Multivariate NB (RPMNB) Model Estimation Results

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

[5] Please see Table 3 for variable definitions and units

No of shopping center	2	0.041 (1.859)	0.021 (1.721)	--	--	--	--
Socio-demographic characteristics							
Population density	6	0.246 (10.682)	0.133 (12.254)	0.144 (10.490)	0.027 (3.243)	0.114 (10.808)	0.128 (10.153)
Non-motorist commuter	3	0.034 (1.883)	--	0.044 (2.152)	--	--	0.042 (1.752)
Proportion of household without vehicle	3	-0.674 (1-.748)	--	-1.143 (-3.077)	--	--	2.491 (6.084)
Over dispersion	6	0.522 (24.872)	0.179 (9.849)	0.291 (11.565)	0.491 (23.614)	0.098 (5.921)	0.033 (2.152)
Unobserved Heterogeneity							
Correlation 1	1	0.669 (27.067)	--	--	--	--	0.669 (27.067)
Correlation 2	1	--	0.772 (48.990)	0.772 (48.990)	0.772 (48.990)	--	--
Total number of parameters= 92, Log-likelihood: -43,597.82; AIC: 87,379.64; BIC:87,954.34							

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

TABLE 5 Predictive Performance Measure of Two Models

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

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

*Red colours are the one where multivariate NB model performs better

APPENDIX A

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

The reader would note, for simplicity in comparison, we do not add unobserved parameters in the models provided in Appendix A.

TABLE A.1 Model 1: Traditional Multivariate Model with Distinct Propensity Equations

Variables[6]	**Rear End**		**Angular**		**Sideswipe**		**All single vehicle**		**Other multiple vehicle**		**Non-motorized**	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Constant	-0.770	-10.181	-1.301	-14.573	-2.161	-16.951	-0.612	-7.364	-1.261	-14.781	-3.391	-25.776
Roadway Characteristics												
Proportion of arterial roads	0.232	5.769	0.114	2.082	0.123	1.770	-0.232	-4.805	--	--	0.265	3.757
Number of intersections	--	--	0.235	7.037	--	--	--	--	0.087	3.147	0.243	5.239
Variance of speed	0.037	3.560	0.040	3.263	0.075	5.153	0.021	1.971	0.032	2.830	--	--
Length of divided road	0.478	2.427	0.357	1.978	0.361	1.668	0.512	3.590	0.458	2.528	--	--
Signal intensity	--	--	--	--	-0.753	-3.330	-0.632	-2.704		--	--	--
Average outside shoulder width	-0.420	-7.493	-0.135	-3.078	-0.321	-5.840	--	--	-0.072	-1.891	--	--
Road length over 55mph	0.900	7.911	-0.424	-2.509	1.165	6.711	1.245	10.174	--	--	-0.469	-1.923
Sidewalk width	0.104	3.859	--	--	--	--	--	--	--	--	-0.071	-2.874
Traffic Characteristic												
VMT	--	--	0.060	4.504	0.187	12.395	-0.111	-4.127	0.094	7.979	0.061	3.300
Truck VMT	0.183	20.085	--	--	--	--	0.325	11.169	--	--	--	--
Land-use attributes												
Urban area	0.169	17.080	0.117	8.812	0.132	7.968	0.063	6.286	0.082	6.304	0.158	7.629
Office area	0.201	15.566	0.226	14.091	0.221	10.625	0.087	6.602	0.157	9.691	0.158	7.414
Institutional area	0.046	3.342	0.079	4.996	--	--	--	--	0.054	3.892	0.113	5.683
Residential area	-0.064	-7.251	-0.025	-2.128	-0.103	-6.621	--	--				
Built environment characteristic												
No. of restaurant	0.226	8.275	0.222	8.124	0.318	11.882	0.102	6.062	0.292	11.228	0.212	9.009
No. of shopping center	0.074	2.842	0.067	1.721	--	--	--	--	--	--	--	--

[6] Please see Table 3 for variable definitions and units

Socio-demographic characteristics												
Population density	0.148	15.432	0.127	16.045	0.129	11.311	0.027	3.675	0.105	14.110	0.126	11.010
Non-motorist commuter	0.037	2.096	--	--	0.055	2.381	--	--	--	--	0.041	1.770
Proportion of household without vehicle	-0.463	-1.683	--	--	-0.646	-1.871	--	--	--	--	2.508	6.609
Over dispersion	0.943	36.926	0.729	23.693	0.946	20.026	0.491	20.471	0.557	18.801	0.427	9.019
Total number of parameters = 89, Log-likelihood: -44,791.53; AIC: 89,761.07; BIC:90,317.02												

TABLE A.2 Model 2: Panel Model with Same Specification as Model 1

Variables[7] (Base in Overall Crash Risk Component)	Overall Crash Risk	Deviation					
		Rear End (1)	Angular (2)	Sideswipe (3)	All single Vehicle (4)	Other Multiple Vehicle (5)	Non-motorized (6)
	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)
Constant	-3.390 (-22.411)	2.617 (13.917)	2.091 (11.546)	1.229 (6.025)	2.778 (15.951)	2.129 (11.989)	--
Roadway Characteristics							
Proportion of arterial roads (1)	0.232 (3.969)	--[*]	-0.118 (-1.287)	-0.110 (-0.927)	-0.464 (-5.913)	N/I[**]	0.033 (0.336)
Number of intersections (6)	0.243 (5.228)	N/I	-0.008 (-0.126)	N/I	N/I	-0.156 (-2.541)	--
Variance of speed (1)	0.037 (2.633)	--	0.003 (0.159)	0.038 (1.639)	-0.016 (-0.890)	0.005 (0.260)	N/I
Length of divided roads (1)	0.476 (1.723)	--	-0.117 (-0.278)	-0.112 (-0.235)	0.034 (0.093)	-0.016 (-0.046)	N/I
Signal intensity (3)	-0.752 (-2.821)	N/I	N/I	--	0.122 (0.368)	N/I	N/I
Average outside shoulder width (1)	-0.421 (-6.389)	--	0.286 (3.513)	0.100 (1.040)	N/I	0.349 (4.038)	N/I
Roads length over 55mph (1)	0.903 (5.407)	--	-1.328 (-5.063)	0.261 (0.862)	0.343 (1.179)	N/I	-1.367 (-4.487)
Sidewalk width (1)	0.105 (3.629)	--	N/I	N/I	N/I	N/I	-0.176 (-6.962)
Traffic Characteristic							
VMT (2)	0.060 (5.128)	N/I	--	0.128 (6.975)	-0.170 (-6.620)	0.035 (1.911)	-0.002 (-0.071)
Truck VMT (1)	0.183 (15.736)	--	N/I	N/I	0.142 (4.699)	N/I	N/I
Land-use Attributes							
Urban area (1)	0.170	--	-0.053	-0.038	-0.106	-0.087	-0.012

[7]

	(13.060)		(-2.620)	(-1.746)	(-5.839)	(-5.139)	(-0.418)
Office area (1)	0.201 (10.952)	--	0.025 (1.327)	0.020 (0.833)	-0.114 (-4.775)	-0.044 (-1.811)	-0.044 (-1.612)
Institutional area (1)	0.046 (2.675)	--	0.034 (1.156)	N/I	N/I	0.008 (0.330)	0.068 (2.318)
Residential area (1)	-0.063 (-5.321)	--	0.038 (2.098)	-0.040 (-1.212)	N/I	N/I	N/I
Built Environment Characteristic							
No. of restaurant (1)	0.226 (5.106)	--	0.003 (0.044)	0.092 (1.651)	-0.124 (-2.424)	0.067 (1.758)	-0.014 (-0.279)
No of shopping center (1)	0.074 (1.802)	--	-0.007 (-0.105)	N/I	N/I	N/I	N/I
Socio-demographic Characteristics							
Population density (1)	0.148 (10.789)	--	-0.021 (-1.107)	-0.019 (-0.750)	-0.121 (-7.447)	-0.043 (-2.534)	-0.022 (-1.121)
Non-motorist commuters (1)	0.036 (2.494)	--	N/I	0.019 (0.401)	N/I	N/I	0.005 (0.145)
Proportion of household without vehicle (1)	-0.437 (-1.730)	--	N/I	-0.213 (-0.267)	N/I	N/I	2.935 (4.606)
Over dispersion	--	0.943 (32.137)	0.729 (24.060)	0.946 (21.171)	0.491 (23.583)	0.557 (21.641)	0.427 (9.987)
Total number of parameters = 89, Log-likelihood: -44,791.53; AIC: 89,761.07; BIC:90,317.02							

Note: *-- defines the base; ** N/I denotes that the variable does not have any impact on the particular crash types

TABLE A.3 Model 3: Parsimonious Model Specification Dropping Insignificant Variables from Model 2

Variables[8] (Base in Overall Crash Risk Component)	**Overall Crash Risk**	**Deviation**					
		Rear End (1)	**Angular (2)**	**Sideswipe (3)**	**All single Vehicle (4)**	**Other Multiple Vehicle (5)**	**Non-motorized (6)**
	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)	Estimate (t-stat)
Constant	-3.448 (-33.249)	2.699 (24.465)	2.113 (16.499)	1.245 (8.111)	2.867 (22.165)	2.187 (15.938)	--
Roadway Characteristics							
Proportion of arterial roads (1-3,6)	0.179 (9.674)	--[*]	--	--	-0.403 (-7.413)	N/I[**]	--
Number of intersections (2,6)	0.242 (9.471)	N/I	--	N/I	N/I	-0.159 (-3.273)	--
Variance of speed (1,2,4,5)	0.032 (7.566)	--	--	0.044 (2.389)	--	--	N/I
Length of divided roads (1-5)	0.451 (9.257)	--	--	--	--	--	N/I
Signal intensity (3-4)	-0.685 (-6.538)	N/I	N/I	--	--	N/I	N/I
Average outside shoulder width (1,3)	-0.351 (10.229)	--	0.223 (3.895)	--	N/I	0.278 (4.712)	N/I
Roads length over 55mph (1,3,4)	1.109 (21.579)	--	-1.489 (-9.475)	--	--	N/I	-1.494 (-4.487)
Sidewalk width (1)	0.076 (3.088)	--	N/I	N/I	N/I	N/I	-0.151 (-6.958)
Traffic Characteristic							
VMT (2,6)	0.060 (7.079)	N/I	--	0.126 (8.932)	-0.175 (-7.508)	0.035 (2.332)	--
Truck VMT (1)	0.186 (18.694)	--	N/I	N/I	0.141 (5.173)	N/I	N/I
Land-use Attributes							

[8] Please see Table 3 for variable definitions and units

Urban area (1,6)	0.174 (17.160)	--	-0.056 (-3.283)	-0.049 (-2.531)	-0.113 (-8.161)	-0.092 (-6.200)	--
Office area (1-3)	0.216 (31.945)	--	--	--	-0.132 (-8.670)	-0.061 (-3.337)	-0.061 (-2.784)
Institutional area (1,2,5)	0.063 (9.772)	--	--	N/I	N/I	--	0.068 (2.318)
Residential area (1,3)	-0.079 (-13.480)	--	0.059 (4.234)	--	N/I	N/I	N/I
Built Environment Characteristic							
No. of restaurant (1,2,6)	0.219 (14.863)	--	--	0.099 (2.336)	-0.117 (-4.296)	0.074 (2.498)	--
No of shopping center (1,2)	0.076 (6.523)	--	--	N/I	N/I	N/I	N/I
Socio-demographic Characteristics							
Population density (1,2,3,6)	0.134 (34.649)	--	--	--	-0.109 (-11.870)	-0.029 (-2.412)	--
Non-motorist commuters (1,3,6)	0.043 (5.163)	--	N/I	--	N/I	N/I	--
Proportion of household without vehicle (1,3)	-0.476 (-2.444)	--	N/I	--	N/I	N/I	3.044 (6.321)
Over dispersion	--	0.948 (32.462)	0.731 (24.381)	0.951 (21.434)	0.490 (23.749)	0.557 (21.905)	0.433 (10.217)
Total number of parameters = 58, Log-likelihood: -44,808.32; AIC: 89,732.64; BIC:90,094.95							

Note: *-- defines the base; ** N/I denotes that the variable does not have any impact on the particular crash types

TABLE A.4 Model 4: Parsimonious Model Specification with Net Effect of Each Exogenous Variable from Model 3

Variables[9]	Rear End		Angular		Sideswipe		All single vehicle		Other multiple vehicle		Non-motorized	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Constant	-0.749	-8.787	-1.335	-16.767	-2.203	-25.154	-0.581	-11.104	-1.261	-17.328	-3.448	-33.249
Roadway Characteristics												
Proportion of arterial roads	0.179	9.674	0.179	9.674	0.179	9.674	-0.224	-4.996	--*	--	0.179	9.674
Number of intersections	--	--	0.242	9.471	--	--	--	--	0.083	2.651	0.242	9.471
Variance of speed	0.032	7.566	0.032	7.566	0.076	4.888	0.032	7.566	0.032	7.566	--	--
Length of divided road	0.451	9.257	0.451	9.257	0.451	9.257	0.451	9.257	0.451	9.257	--	--
Signal intensity	--	--	--	--	-0.685	-6.537	-0.685	-6.537	--	--	--	--
Average outside shoulder width	-0.351	-10.231	-0.127	-3.228	-0.351	-10.231	--	--	-0.072	-1.742	--	--
Road length over 55mph	1.109	21.584	-0.380	-2.902	1.109	21.584	1.109	21.584	--	--	-0.385	-1.666
Sidewalk width	0.076	3.091	--	--	--	--	--	--	--	--	-0.075	-3.882
Traffic Characteristic												
VMT	--	--	0.060	7.079	0.186	18.632	-0.115	-5.367	0.095	9.848	0.060	7.079
Truck VMT	0.186	18.692	--	--	--	--	0.327	13.070	--	--	--	--
Land-use attributes												
Urban area	0.174	17.153	0.118	9.913	0.125	8.902	0.061	8.189	0.082	8.483	0.174	17.153
Office area	0.216	31.945	0.216	31.945	0.216	31.945	0.085	6.993	0.156	11.014	0.156	7.848
Institutional area	0.063	9.777	0.063	9.777	--	--	--	--	0.063	9.777	0.111	5.414
Residential area	-0.079	-13.484	-0.020	-1.944	-0.079	-13.484	--	--	--	--	--	--
Built environment characteristic												
No. of restaurant	0.219	14.860	0.219	14.860	0.318	9.567	0.103	4.997	0.293	14.170	0.219	14.860

[9] Please see Table 3 for variable definitions and units

No. of shopping center	0.076	6.523	0.076	6.523	--	--	--	--	--	--	--	--
Socio-demographic characteristics												
Population density	0.134	34.659	0.134	34.659	0.134	34.659	0.024	3.100	0.105	10.690	0.134	34.659
Non-motorist commuter	0.043	5.163	--	--	0.043	5.163	--	--	--	--	0.043	5.163
Proportion of household without vehicle	-0.476	-2.446	--	--	-0.476	-2.446	--	--	--	--	2.568	6.599
Over dispersion	0.948	32.464	0.731	24.381	0.951	21.434	0.490	23.749	0.557	21.905	0.433	10.217
Total number of parameters = 58, Log-likelihood: -44,808.32; AIC: 89,732.64; BIC:90,094.95												

Note: *-- = attribute insignificant at 90% significance level