ACCOMMODATING FOR SYSTEMATIC AND UNOBSERVED HETEROGENEITY IN PANEL DATA: APPLICATION TO MACRO-LEVEL CRASH MODELING

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

Shamsunnahar Yasmin
Senior Lecturer/ Senior Research Fellow
Queensland University of Technology (QUT)
Centre for Accident Research & Road Safety – Queensland (CARRS-Q)
Brisbane, Australia
Email: shams.yasmin@qut.edu.au
Telephone: +61731384677
ORCiD number: 0000-0001-7856-5376

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

Declarations of Interest: None

*Corresponding author

ABSTRACT

The current research contributes to the burgeoning literature on multivariate models by proposing a hybrid model framework that (a) incorporates unobserved heterogeneity in a parsimonious framework and (b) allows for additional flexibility to accommodate for observed/systematic heterogeneity. Specifically, we estimate a Latent Segmentation Panel Mixed Negative Binomial (LPMNB) model to study the zonal level crash counts across different crash types. Further, we undertake a comparison exercise of the proposed hybrid LPMNB model with a Panel Mixed Negative Binomial model (PMNB) that accommodates for unobserved heterogeneity via a simulation setting. The analysis is conducted using the zonal level crash records by different crash types from Central Florida region for the year 2016 considering a comprehensive set of exogenous variables. The comparison exercise is further augmented by computing several goodness of fit measures along with elasticity analysis and the results offered by the LPMNB model highlight the value of the proposed model. Further, to offer insights on model selection incorporating computational complexity dimension along with other important attributes, we conduct a trade-off analysis considering four different attributes: (a) model fit, (b) prediction, (c) inference power and (d) computational complexity; across six different model strictures including traditional crash frequency models and our proposed LPMNB model.

Keywords: Unobserved heterogeneity; Parsimonious structure, Panel Latent segmentation; Panel mixed negative binomial; Crash type.

1 BACKGROUND

Road traffic crashes and their consequences remain a global health concern given the extent of societal, emotional and economic impacts of these unfortunate events. According to a recent report by NHTSA (NHTSA, 2018), road traffic crashes, responsible for 36,750 fatalities in the US, ranked as the third deadliest in the decade and a leading cause of death among people aged between 17 and 21 years old. The numbers while declining relative to 2016 and 2017, are still 12.2% higher than 2014 (an all time low) and warrants our attention for devising appropriate solutions for reducing the number and consequence of such unfortunate events (NHTSA, 2018). Crash frequency models are an important component for devising and evaluating road safety policies and counter measures. These models examine crashes either at the micro-level (such as an intersection or roadway segment) or at the macro-level (such as a county or Traffic Analysis Zone (TAZ)). Traditionally, earlier studies developed crash prediction models considering a single count variable (typically the total number of crashes) for a spatial unit and study the impact of exogenous variables. Econometric approaches for developing crash prediction models in a univariate setting are dominated by count regression frameworks (Poisson and negative binomial (NB)) (please see (Anastasopoulos and Mannering, 2011, 2009; Bhowmik et al., 2018; Cai et al., 2018; Lord and Mannering, 2010; Yasmin and Eluru, 2018) for a literature review).

In recent years, studies show that a single total crash model will not be able to parse the distinct crash distribution by different attributes (such as type, injury severity, and modes) and such aggregation can result in aggregation bias and loss of information available in the dataset. Hence, in recent years, safety researchers have focused on disaggregating the data by various attributes such as crash typology (Alarifi et al., 2018; Bhowmik et al., 2019a, 2018; Chen et al., 2016; Cheng et al., 2017; Intini et al., 2020; Wang et al., 2017), injury severity (Bhowmik et al., 2019a; Eluru and Bhat, 2007; Nashad et al., 2016; Serhiyenko et al., 2016; Wang et al., 2021, 2019; Yasmin et al., 2016; Yu and Abdel-Aty, 2013), vehicle involvement (Dong et al., 2014; Lee et al., 2018, 2015) and crash location (Bhowmik et al., 2021; Song et al., 2006). It is beyond the scope of our paper to review the vast literature on crash frequency (please see Bhowmik, 2020; Bhowmik et al., 2021 for a detailed documentation on crash frequency literature, particularly on crash type analysis). The disaggregation results in multiple dependent variables for each observational unit. Univariate models can be estimated for each dependent variable to address the influence of observed factors. However, accommodating for common unobserved factors across these dependent variables requires us to develop multivariate approaches. Ignoring the influence of such common unobserved factors, if present, may lead to biased and inefficient parameter estimates resulting in erroneous policy implications (see Mannering et al., 2016 for an extensive discussion).

Recognizing this drawback, recent research in safety literature has shifted toward multivariate modeling frameworks that accommodate for the influence of these common unobserved factors (Anastasopoulos, 2016; Mannering et al., 2016; Nashad et al., 2016). In these multivariate models, typically probability computation 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[1] requiring simulation. The simulation-based approaches are estimated 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). The various model structures

[1] In some cases, a parametric multivariate distributional assumption might result in closed form approaches (such as the copula based approaches (Bhowmik et al., 2021; Nashad et al., 2016) or an approximated integral is computed using quasi-likelihood methods (see Narayanamoorthy et al., 2013).

developed from multivariate models include multivariate Poisson regression model (Ye et al., 2009), multivariate Poisson lognormal model (Serhiyenko et al., 2016), multinomial-generalized Poisson model (Chiou and Fu, 2013), multivariate Poisson gamma mixture count model (Mothafer et al., 2016), multivariate Poisson lognormal spatial and/or temporal model (Cheng et al., 2017; Jonathan et al., 2016), Integrated Nested Laplace Approximation Multivariate Poisson Lognormal model (Wang et al., 2017), Bayesian latent class flexible mixture multivariate model (Heydari et al., 2017), multivariate random-parameters zero-inflated negative binomial model (Anastasopoulos, 2016), multivariate random parameter count model (Buddhavarapu et al., 2016; Huo et al., 2020), correlated grouped random parameters bivariate probit model (Fountas et al., 2019) and recently proposed fractional split approach (Afghari et al., 2020; Bhowmik, 2020; Bhowmik et al., 2019b, 2018; Yasmin et al., 2016; Yasmin and Eluru, 2018).

1.1 Contributions of the Current Study

As is evident from the discussion above, simulation-based approaches have been extensively applied for multivariate models. However, several challenges still exist with these multivariate models. The current research contributes to burgeoning literature on multivariate models by proposing a model framework that (a) incorporates unobserved heterogeneity in a parsimonious framework and (b) allows for additional flexibility to accommodate for observed/systematic heterogeneity. The proposed approach builds on previous work from two studies (Bhowmik et al., 2019a; Yasmin and Eluru, 2016).

Bhowmik and colleagues proposed a parsimonious model structure for multivariate models by recasting the multivariate crash frequency modeling problem as a pooled univariate crash frequency analysis problem (with unobserved heterogeneity accommodated). To elaborate, instead of considering the crash frequency by crash type as a multivariate distribution, the authors represent it as repeated measures of crash frequency while recognizing that each repetition represents a different crash type. The recasting process allows for the estimation of a parsimonious model system by allowing for an improved specification testing of variable impacts across different crash types (see (Bhowmik et al., 2019a) for details). Using this consideration, the proposed model system enhances the efficiency of estimation through a single crash frequency model while also allowing for parameter effects to vary across different crash types through crash type specific deviation terms. Further, as only one propensity equation is to be estimated, it allows for reduction in parameters especially for unobserved factors resulting in substantial improvements in model efficiency and computational times.

Bhowmik et al. 2019a approach offered significant enhancement to the state of the art multivariate model estimation. However, the study implicitly started with a population homogeneity assumption for the estimated parameters i.e. the influence of exogenous variables was assumed to be the same across the dataset (Eluru et al., 2012; Yasmin and Eluru, 2016). While the assumption was partially relaxed through random parameters estimated for each parameter the process is computationally intensive and simulation reliant. Furthermore, the approach completely focuses on the unobserved error component of the propensity. To address this restriction, in our proposed study, we bridge the panel recasting approach with the latent segmentation based approach employed for crash frequency modeling (Park et al., 2010; Park and Lord, 2009; Zou et al., 2014; Yasmin and Eluru, 2016; Fountas et al., 2018; Yu et al., 2019). In a latent segmentation model, TAZs are allocated probabilistically to different segments and a segment specific model is estimated for each segment. The probabilistic assignment explicitly acknowledges the role played by unobserved factors in moderating the impact of observed exogenous variables. Further, the approach provides valuable insights on how the exogenous variables affect segmentation. To the best of authors' knowledge, this study is the first of its kind to develop a latent class count model considering multiple dependent

variables (different crash types) while simultaneously accommodating potential correlations resulting from unobserved factors across the count dimensions.

To summarize, our current study contributes to crash frequency literature both methodologically and empirically by estimating a Latent Segmentation Panel Mixed Negative Binomial (LPMNB) to study the zonal level crash counts across different crash types. The newly formulated model will allow us to partition the TAZs into segments based on their attributes and estimate the influence of exogenous variables on crash counts of different crash types. From a *methodological* perspective, the current research makes a threefold contribution to literature on crash frequency analysis: First, the recasting allows us to estimate a parsimonious model system and reduce the computational time for estimating parameters associated with unobserved factors. Second, by introducing the latent class version of the PMNB model, we allow for both observed/systematic and unobserved heterogeneity relaxing the homogeneity assumption of the traditional count models. Third, we allow for a flexible segment membership function and test for the presence of multiple segments in the model estimation. *Empirically*, the research contributes to our understanding of analyzing zonal level crashes for both motorized and non-motorized road user group while considering different crash types within the motorized category including rear-end, angular, sideswipe, single vehicle and head-on crashes.

The analysis is conducted using the zonal level crash records from Central Florida for the year 2016 considering a comprehensive set of exogenous variables. Further, we undertake a comparison exercise of the proposed LPMNB model with its' counterpart proposed in previous work by Bhowmik and colleagues (Bhowmik et al., 2019a).

2 METHODOLOGY

The focus of our study is to estimate a Latent Segmentation based Panel Mixed NB modeling framework and compare its performance with previously proposed Panel Mixed NB model (PMNB). In this section, we restrict ourselves to the latent model system (please refer to section 4.1 in the current paper and our earlier work (Bhowmik et al., 2019a) for details on PMNB model). The general structure for latent segmentation-based count models involves specifying these two components: (1) assignment component and (2) segment specific count model component. For the ease of presentation, we describe modeling framework by the components.

2.1 Model Structure

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

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

where, P_{is} is the probability of TAZ_i to be assigned to segment s, $\boldsymbol{z}_s$ is a vector of attributes and $\boldsymbol{\alpha}_s$ is a conformable parameter vector to be estimated. Segment Specific Count Component

The probability equation of the NB formulation can be rewritten as follow:

$$P_{is}(y_{ir}|s) = \frac{\Gamma\left(y_{ir}+\frac{1}{\lambda'}\right)}{\Gamma\left(y_{ir}+1\right)\Gamma\left(\frac{1}{\lambda'}\right)}\left(\frac{1}{1+\lambda' v_{ir}}\right)^{\frac{1}{\lambda'}}\left(1-\frac{1}{1+\lambda' v_{ir}}\right)^{y_{ir}} \quad (2)$$

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

$$v_{ir} = E(y_{ir}|\boldsymbol{x}_{ir}) = exp((\boldsymbol{\beta} + \boldsymbol{\theta}_i + \boldsymbol{\varrho}_{ir})\boldsymbol{x}_{ir} + \varepsilon_{ir}) \quad (3)$$

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

2.2 Model Estimation

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

$$L_i = \left(\int_{\Psi} \sum_{s=1}^{S}\left[\prod_{r=1}^{R}\left((P_{is}) \times \left(P_i(y_{ir}|s)\right)\right)\right] f(\Psi) d\Psi \quad (4)$$

Further, we apply simulation techniques to approximate the integrals in the likelihood function and maximize the logarithm of the resulting simulated likelihood function across individuals with respect to Ψ . The simulation technique approximates the likelihood function in Equation (4) by computing the L_i for each TAZ_i at different realizations drawn from a multivariate normal distribution, and averaging it over the different realizations (see (Eluru and Bhat, 2007) for detail). Notationally, if DL_i is the realization of the likelihood function in the cth draw (c = 1, 2, …, C), then the observational likelihood function is approximated as:

$$DL_i = \frac{1}{C}\sum_{c=1}^{C}(\mathrm{DL}_{\mathrm{i}}^{c}) \quad (5)$$

Finally, the log-likelihood function is:

$$LL = \sum_{i} Ln(DL_i) \quad (6)$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 6.

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

3 DATA PREPARATION

Our study area includes the Central Florida region associated with 4,747 zones encompassing a total of 11 counties in the state of Florida and covers an area of approximately 11,150 mile^2 with a population of around 8.2 million. The empirical analysis focused on crashes by different types involving both motor vehicles and non-motorists at a zonal level for 2016. For processing the data, crash data were sorted into two classes based on the road user group: motorist and non-motorist[2]; within the motorized group, the records are further classified into five categories based on the manner of crash: rear-end, angular, sideswipe, head-on and single vehicle crashes. Thus, the number of dependent variables to be analyzed in the current study is six. Crash records for different crash types are sourced from Florida Department of Transportation (FDOT), Crash Analysis Reporting System (CARS) and Signal Four Analytics (S4A) databases. 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, and rollover in the single vehicle crash category. All the crash records are finally aggregated at a TAZ level using Geographic Information System (GIS).

A total of 114,458 motorized (ranging from 0 to 243) and 3,413 non-motorized crashes (ranging from 0 to 12) were reported in the Central Florida region for the year 2016. Within the motorized crashes, rear-end crash is found to be the most prevalent crash type (44.09%) while sideswipe crash is less frequent with 10.82% among all other motorized crash types. The crash counts for each crash type are presented in the top row panel in table 1. Further, we present the distribution of the crashes (across different count categories) corresponding to each crash types in Figure 1. As expected, we see a strong clustering for the non-motorized and head-on crashes around the lower values (<=5 crashes). From the total records, we have partitioned the zonal level records into two datasets as: 1) 3,815 TAZs for estimation analysis and 2) 932 TAZs set aside for validation.

3.1 Explanatory Variables Considered

In addition to the crash records, a number of zonal level attributes are considered for the current analysis including roadway, built environment, land-use, traffic and sociodemographic characteristics. 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. In order to access the

[2] Motorized crashes involved one or multiple motor vehicles and the non-motorized crashes are defined by the collision between a motor vehicle and one or multiple non-motorists (pedestrian/bicyclist).

roadway attributes, 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, average posted 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 are considered in the current study. 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 centers, schools, hospitals, recreational centers, restaurants and shopping centers are collected. Further to accommodate for traffic attributes, we consider 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. Finally, the sociodemographic attributes take into account the population and household density, proportion of means of transportation used by commuter for their work trips (car, motorcycle, transit, bike and walk) proportion of people by age and race and proportion of household by vehicle ownership level (0, 1, 2, and 3 or more).

Table 1 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 and combination of variables 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 Model Specification and Overall Measure of Fit

As discussed earlier, out of 4,747 TAZs, 3,815 TAZs are randomly selected for model estimation and the remaining 932 TAZs are set aside for validation purpose. The number of count dependent variables (crash types) to be analyzed in the current study is six and so every TAZ is repeated six times recognizing that each repetition represents a different crash type (Bhowmik et al., 2019a). Thus, the estimation sample has 22,890 (3,815*6) records and the validation sample has 5,592 (932*6) data records. The empirical analysis involved a series of model estimations. First, we estimated six separate independent NB model for six crash types to establish a benchmark for comparison. Second, we estimated a parsimonious model structure (Panel independent NB model) using the same independent model system while restricting the parameters across different crash types considered. To elaborate, we estimate a base effect for each exogenous variable that is common across the crash types and estimate deviations for each crash types relative to the base effect. If a deviation is insignificant, it concludes that there is no significant difference in effect for that particular variable between the base crash type and crash type for which the deviation was computed (see (Bhowmik et al., 2019a) for more details). Thus, the model estimated in such a panel formulation results in fewer parameters. Third, we estimated a latent class version of the panel negative binomial (LPNB) model to capture the potential variation in the impact of exogenous variables. Fourth, within the Panel NB model and Latent Panel NB model, we consider unobserved heterogeneity in terms of correlation (across the crash count dimensions) and random parameters (within the crash count propensity). Prior to presenting the comparison exercise, we will briefly summarize the panel recasting approach followed by the latent segmentation model in the following section.

4.1 Panel Recasting Approach

The recasting approach involves four steps: First, we estimate the traditional univariate/multivariate NB model for all six crash types with six separate propensity equation. Second, we restructure the data so that each TAZ has repeats six records (same as the number

of crash types). With respect to exogeneous variables, we compute a base effect that will remain the same across the crash types and then estimate deviation terms (interaction between the variable and indicator variables for crash type) for each crash type relative to the base effect. The reader would note that our goal is to first replicate the traditional NB model results and then move to a more parsimonious system. So, the base variable will be defined based on the traditional univariate NB model specification. For instance, let us assume the variable AADT has a significant impact on the crash propensity across all the six crash types. Given the 6 significant estimates, for our recasting approach, there will be 1 base AADT variable which will be common across the six crash types and 5 deviation terms for five crash types (the sixth one will serve as the base). On the other hand, let's say from the univariate NB model results we find that number of intersections in a zone has significant effect on the crash propensity of three crash types say rear-end, sideswipe and angular crashes. For this case, the base variable for the number of intersections will be common across these three crash types and for the other remaining three crash types, the parameter value would be 0. Subsequently, we will estimate two deviation terms for two crash types (with the third set as the base). Third, we drop the statistically insignificant deviation terms for each variable. This is one of the advantages of the recasting approach. In traditional model, the analyst needs to conduct a log-likelihood ratio test for identifying the difference in parameter estimates across the crash types whereas in our system, we can easily identify whether a variable effect is significantly different or not across the crash types based on the t-statistics of the deviation term. After dropping the insignificant deviation terms, we will now have a panel negative binomial model specification in a more parsimonious system (please see (Bhowmik et al., 2019a) for more details on this approach). Fourth, we follow the same process (step 2 and 3) within a latent segmentation approach on a segment-by-segment basis to capture the population heterogeneity.

4.2 Latent Segmentation Approach: Determining Appropriate Number of Segments

In case of latent models, determining the appropriate number of segments is a critical issue with respect to interpretation and inferences. The estimation process for such latent class model begins with the independent model considering two segments. Then we continued adding additional segments until further addition does not enhance intuitive interpretation and data fit (Eluru et al., 2012). For identifying the appropriate number of segments for the latent class model, we employ the Bayesian Information Criterion (BIC) as it offers higher penalty on over-fitting. Specifically, we estimated independent latent NB model with different number of segments (2, 3…) and selected the model with the lowest BIC value. Once, the independent latent model is finalized with appropriate number of segments, we estimated the mixed version of the corresponding independent model.

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

4.3 Comparison Between Models

We estimated five models in two regimes: a) unsegmented models including: 1) Independent NB model; 2) Panel independent NB model (PNB); 3) Panel Mixed NB model (PMNB); and b) segmented model including: 4) Latent Segmentation Panel Independent NB model with two segments (LPNB II) and 5) Latent Segmentation Panel Mixed NB model with two segments (LPMNB II). Finally, we compare the unsegmented models with the latent segmentation based count models in order to assess the importance of accounting for population heterogeneity in estimating zonal level crash frequency models. The reader would note that all the models mentioned above are non-nested in nature and so, we employ several goodness of fit measures including the Akaike information criterion (AIC), corrected AIC (see (Fountas et al., 2020) for details of corrected AIC) and Bayesian Information Criterion (BIC) measures for the comparison exercise.

The results from the various model systems – convergence log-likelihood, number of parameters and the model fit measures are presented in Table 2. Based on the measures in table 2, several observations can be made. First, the PNB model that accounts for penalty for additional parameters provide improved data fit compared to the independent NB model. This supports our hypothesis that the impact of some variables may not differ across the crash types and through the recasting, we can have a parsimonious model system with improved parameter efficiency. Second, the segmented independent LPNB II model performs better relative to the PNB model. This result provides strong evidence in favour of our hypothesis that crash counts by different crash types can be investigated in a more efficient way through the segmentation of the TAZs. Third, models accommodating unobserved effects perform better than their corresponding independent models in both unsegmented (PMNB vs PNB) and segmented regimes (LPMNB II vs LPNB II) highlighting the importance of accommodating for unobserved heterogeneity in examining crash count by different crash types. Fourth, within the mixed models, the unsegmented model (PMNB) provides improved data fit relative to the segmented model (LPMNB II). Based on the results provide above, we can conclude that the segmented model is a preferred choice as long as the framework is estimated in a closed form structure (independent models that do not account for unobserved heterogeneity). However, when we rely on simulation in the latent segmentation model system (LPMNB II) and the panel negative binomial model system (PMNB) for capturing the unobserved effects, the PMNB model outperforms its segmented counterparts.

5 Estimation Results

This section offers a detailed discussion of exogenous variable effects on the crash count outcome for different crash types. Table 4 presents the model estimation results for the proposed Latent Panel Mixed NB model (LPMNB II). The estimation results of the PMNB model are presented in Table 5 for comparison. In discussing the model results, for the sake of brevity, we will restrict ourselves to the discussion of the LPMNB II model only (see Appendix for the results of independent models, PNB and LPNB II models). For the ease of presentation, we first present an intuitive discussion of the segmentation component followed by the segment specific count component by different variable groups.

5.1 Segmentation Component

5.1.1 Descriptive Characteristics of the Segments

To delve into the segmentation characteristics, the model estimates are used to generate information on two criterion including: 1) percentage TAZ share across the two segments, and 2) expected mean of crash count events of different crash types within each segment (see (Eluru et al., 2012) for detail). Table 3 provides these estimates. From the estimates, it is clear that the

likelihood of a TAZ being assigned to segment 1 is substantially higher than the likelihood of being assigned to segment 2 (0.74 vs 0.26). Further, the expected number of crash counts by different crash types conditional on their belonging to a particular segment offer contrasting results indicating that the two segments exhibit distinct crash risk profiles for different crash types in the current study. As evident form table 3, we can observe that relative to observed sample mean, the expected mean crash counts by different crash types is higher in segment 1 (except head -on) while in segment 2, the expected mean is lower for every crash types except head-on crashes. Interestingly, segment 2 has higher risk for head-on crashes relative to segment 1. Based on overall results, it is clear that a TAZ, if allocated to segment 1 is likely to experience higher number of crashes by most of the crash types than if allocated to segment 2. Thus, we may label segment 1 as the "high risk segment" and segment 2 as the "low risk segment".

5.1.2 Segment Membership Component

The latent segmentation component determines the relative prevalence of each segments, as well as the likelihood of a TAZ being allocated to one of the two segments based on zonal level exogenous variables. In our analysis, we find that segment share is influenced by zonal level roadway and land use attributes. In particular, number of intersections, average outside should width, urban area and residential area in a zone affect the assignment of a TAZ to a segment. The first row panel of Table 4 represents the effect of these control variables. In the segmentation component, one of the segments must be the base for every variable for the sake of identification. In our current analysis, the high risk segment (segment 1) is chosen to be the base and the coefficients presented in the table correspond to the propensity for being a part of the low risk segment (Segment 2). Thus, a positive (negative) sign for a variable in the segmentation component indicates that TAZs with the variable characteristics are more (less) likely to be assigned to the low risk segment relative to the high risk segment.

The positive sign on the constant does not have any substantive interpretation after the introduction of other independent variables. From the estimated results, we can observe that higher number of intersections in a zone increase the likelihood of assigning the TAZ to the high risk segment while TAZ with wider shoulder width have a higher probability to be allocated to the low-risk segment. TAZ with more urbanized area are more likely to be assigned to the high-risk segment. On the other hand, with increase in residential area, the likelihood of a TAZ to being allocated in the low risk segment increases. Based on these results, we can argue that high risk segment consists of urbanized zone having higher number of intersections with narrow average outside shoulder and less residential area. On the other hand, zones within segment 2 are more likely to be characterized by rural area with less intersections, wider average outside shoulder width and more residential area.

5.2 Segment Specific Count Component

The coefficients in Table 4 represent the effect of exogenous variables on the frequency component of each crash type within each segment. The reader would note that, within each segment, the variables in the crash count component of Table 4 with positive (negative) sign indicates that an increase in the variable is likely to result in more (less) crashes. In the subsequent sections, we provide a discussion of model results for different crash types by segment groups.

5.2.1 High Risk Segment (Segment 1)

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

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.

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

In terms of proportion of roads over or equal 55mph speed limits, we find contrasting results across different crash types within the high risk segment. For instance, the positive coefficient offered by the variable on rear-end, sideswipe and single vehicle crashes (same effect) indicates an increased likelihood of these crash types in a TAZ having higher percentage of roads over 55mph speed limit. On the other hand, the estimated results show that TAZ with more high-speed roads (≥55mph) results in reduced incidence of angular, head-on and non-motorized crashes. The result is expected since high speed roads are usually straight (less curvature) with a divider or median which reduce the risk of angular and head-on crashes. Further, we found that the impact of the proportion of road over 55mph has significant variability on angular crashes (indicated by the standard deviation parameter) which implies that the overall impact is most likely to be negative (98%).

Land-use Characteristics:
Within the high risk segment, the only land use characteristic influencing crash risk by different crash types is the amount of office area in a zone. As evident from Table 4, we can see that office area is positively associated with rear-end, sideswipe and non-motorized crashes indicating a higher likelihood of these crash types in a TAZ with increased office areas. This variable basically reflects the presence of higher vehicular and non-motorist interactions and in turn, higher exposure for both road user groups.

Built Environment Characteristics:
In terms of built environment attributes, we considered a number of variables, among which only number of restaurants and shopping centers have significant impact on zonal level crash risks within the high risk segment. In particular, higher number of restaurant and shopping centers in a TAZ results in higher incidence of rear-end and sideswipe crashes perhaps due to the higher density of traffic volume for these zones. With respect to non-motorized crashes, number of restaurants is found to be a significant determinant with a positive impact (see (Yasmin et al., 2021)for similar result).

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

Socio-demographic Characteristics:
For socio-demographic attributes, we consider the number of non-motorists (walk/bike) and transit commuters in a zone as additional exposure measures for the crash risk model. As evident from table 4, our analysis shows that TAZ with increased number of non-motorist commuters is likely to experience increased number of rear-end, sideswipe, non-motorized and angular crashes. In fact, the reader would note that the magnitude of these impacts is same across the three crash types (rear-end, sideswipe and non-motorized) while a more larger impact is observed for the angular crashes. On the other hand, the likelihood of being involved in a rear-end and non-motorized crashes increases with increasing share of transit commuters in a zone.

Unobserved Common Factors:
The final set of variables in Table 4 correspond to the potential correlation affecting zonal level crash counts by different crash types simultaneously. The reader would note that, in estimating the model, we found significant impact of two common unobserved components[3] including (1) common unobserved factors affecting rear-end and non-motorized crashes and (2) common unobserved factors affecting angular, sideswipe and all single vehicle crashes. Overall, the results clearly indicate the presence of common unobserved heterogeneity across different crash types within the high risk segment. As explained earlier, though we consider both road user groups, all the crash types considered in the analysis involved motor vehicles and this common involvement might be a possible reason for the significant correlation across these crash types.

5.2.2 Low Risk Segment (Segment 2)
The crash risk component for different crash types within the low risk segment (segment 2) is discussed in this section by variable groups. Similar to the high-risk segment, the effect observed for different attributes on different crash types are also intuitive in the low risk segment. As evident form table 4, we can see that the crash count propensity for different crash types for the "low risk" segment provides variable impacts that are significantly different, in magnitude (for a few variables), from the impacts offered by the exogenous variables in "high risk" segment. Additionally, the number of variables influencing the zonal level crash frequency by different crash types are significantly lower in the low risk segment relative to the high risk segment which further highlights the difference between the two segments.

Crash Specific Constants:
Similar to the high risk segment, the crash specific constants in the low risk segments also represent the intercept of crash propensity after adding the various exogenous variables and do not have any substantive interpretation.

[3] The same correlation structure was revealed from the PMNB model also (as shown in Table 5).

Roadway Characteristics:
As in the high risk segment, proportion of arterial roads offers a negative influence on single vehicle crashes in the low risk segment also (same reasoning as segment 1) though the magnitude is much higher in the low risk segment. One possible explanation can be attributed to the fact that segment 2 consists of zone with wider outside average shoulder width. Outside shoulder width in a road reflects the extra margin of safety for vehicular maneuvers and thus reduce the potential for single vehicle crashes. Further, the parameter associated with signal intensity offers contrasting effects on different crash types. While an increase in the variable positively influences the rear-end, sideswipe and non-motorized crashes, a negative association is observed for single vehicle crashes. This is intuitive as with more signals on the road, the traffic density increases thus results in increased conflicts between vehicles to vehicles and vehicles to non-motorists. At the same time, these conflicts result in lower operating speed which in turn reduce the potential for single vehicle crashes. Interesting thing to note is that the influence of signal intensity is not different for the three crash types (rear-end, sideswipe and non-motorized).

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

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

Built Environment Characteristics:
We did not find any variable specific to build environment characteristics to significantly affect the zonal level crash counts of different crash types in the low risk segment.

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

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

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

6 COMPARISON EXERCISE

6.1 Predictive Performance

In an effort to assess the predictive performance of the estimated models, we compute several goodness fit of measures at disaggregate level including MPB (Mean prediction bias), MAD (mean absolute deviation), MAPE (mean absolute percentage error), RMSE (Root mean square error) and predictive log-likelihood (see 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 the model with lower value of predictive measures and higher value of predictive log-likelihood will reflect better performance in terms of prediction and statistical fit relative to the observed data. Table 6 presents the values of these measures for PMNB and LPMNB models for both in-sample and holdout-sample measures.

Several observations can be made based on the measures presented in Table 6. First, a total of 70 prediction measures are estimated considering six crash types and total crash counts in both estimation and validation sample. Out of these 70 measures, LPMNB model provide improved predictive performance for most of the measures (52). Second, whenever PMNB model performs better, the differences are not substantially large. For example, the RMSE value estimated for sideswipe crashes form PMNB model is 4.171 (for estimation sample) while LPMNB model provides a RMSE value of 4.216. On the other hand, for rear-end, the RMSE value found from PMNB is 38.098 (for estimation sample) whereas for LPMNB, it is only 18.682. This clearly indicates the improved predictive power of the segmented model over its' unsegmented counterpart. Third, with respect to predictive log-likelihood, again LPMNB model performs better in most of the crash types (10 out of 14). The reader would note that, there is a difference between estimated and predicted log-likelihood. When we estimate our model considering correlation and unobserved effects, for every observation unit (TAZ), we get a joint probability for log-likelihood estimation. However, in terms of prediction, we want to see the difference in likelihood across crash types and thus need to estimate probability by crash type. Though PMNB model provides improved data fit in terms of model estimation (estimated log-likelihood, discussed in section 4.1.2), it falls short in prediction (based on predictive log-likelihood). In summary, the resulting goodness of fit measures and predictive log-likelihood offer by the LPMNB model clearly highlight its improved performance over the PMNB model.

6.2 Elasticity Effects

The parameters of the exogenous variables in Table 4 and 5 do not directly provide the exact magnitude of the effects of variables on the zonal level crash counts across different crash types. However, it might be possible that the effects (exact magnitude) of some attributes could differ considerably across the two frameworks. To evaluate this, we compute aggregate level

elasticity effects for both PMNB and LPMNB models. In particular, we estimate the percentage change in the expected zonal level crash counts for every crash types in response to the increase of the explanatory variable by 10% (see (Eluru and Bhat, 2007; Kabli et al., 2020) for a discussion on the methodology for computing elasticities). For this purpose, we identify a subset of exogenous variables including proportion of arterial roads, variance of speed limit, proportion of roads over 55mph and proportion of roads with separated median. Further, for the LPMNB model, we estimate the aggregate level elasticities for the overall sample as well as for each segment separately to emphasize policy repercussions based on most critical contributory factors. For the overall sample, we took the segmentation probabilities into consideration. Table 7 provides the elasticity results across the crash types for both PMNB and LPMNB models. Further, to generate a distribution of the elasticity effects, we employ 50 realizations of the parameters from both model employing a normal distribution assumption based on the parameter and its standard error from the corresponding model. The generated confidence band of the elasticity effects will allow us to test if the elasticity effects significantly differ across the two models (LPMNB and PMNB) for same variable. Figure 2 represents the confidence bands for 4 different variables across six crash types generated for the two models based on the results from 50 realizations.

Several observations can be made based on the elasticity effects presented in Table and Figure 2. First, from the elasticity effects presented in table, we can clearly see some significant differences across two segments for some variables which highlights the importance of allowing for population heterogeneity in examining aggregate level crash counts across different crash types. For instance, due to the 10% increase in proportion of arterial roads, the expected mean of single vehicle crashes will reduce by 0.97% in the high risk segment whereas the effect is larger in low risk segment with a reduction rate of 1.66%. Such differences can also be observed for other variables including variance of speed limit on rear-end, angular and sideswipe crash counts; and proportion of roads over 55mph on single vehicle crashes. Second, interestingly, with respect to the variables present in both segments, TAZs assigned to low risk segment have higher elasticities relative to the high risk segment. Third, in terms of comparison across the two models adopted in the study (from Figure 2), we found substantial differences in elasticities. Specifically, the confidence band for the two models are quite different (with the exception of the proportion of arterial roads). For example, across rear-end, angular and sideswipe crashes, LPMNB model has narrower band relative to PMNB model for the proportion of road over 55mph speed. On the other hand, for the same three crash types, LPMNB model provides wider confidence band for the variable that corresponds to proportion of roads with separated median.

An examination of the mean elasticity values indicates that for the proportion of roads over 55mph speed, the PMNB model predicts an increase of 0.88% in expected mean for single vehicle crashes while LPMNB model predicts 1.16%. Similarly, with a 10% increase in the proportion of roads with separated median, PMNB model predicts a 0.74% increase in expected mean for rear-end crashes whereas the elasticity value is almost doubled (1.62% increase) in LPMNB model. Thus, it is evident that allowing for a flexible specification (population heterogeneity) of observed and unobserved factors provides representative variable impacts.

6.3 Trade-off

The earlier sections presented the comparison of the various model frameworks in terms of model fit and predictive power. However, other considerations such as inference power and computational complexity also influence model selection. It is quite possible that the model that has a greater inference and prediction capability (say RPMNB) can be computationally resource intensive while a model that is simple to estimate has a moderate prediction power but fail to discover the underlying factors properly (Mannering et al., 2020). Based on the

application purpose, there is possible variation in the "best" model selection. To offer insights on model selection incorporating computational complexity dimension along with other important attributes, we conduct a multi-attribute comparison across the six different models comparing four attributes: (a) model fit, (b) prediction, (c) inference power and (d) computational complexity. We will illustrate the trade-off across the mentioned measures considering six different models along the three streams: 1. Traditional count models: a) univariate NB model and b) random parameter multivariate NB model; 2. Panel recasting Model: a) Panel negative binomial model (PNB) and b) Panel mixed negative binomial model (PMNB) and 3. Latent segmentation panel recasting model: a) Latent segmentation panel NB model (LPNB) and b) Latent segmentation panel mixed NB model (LPMNB).

While several potential measures can be generated as surrogates for these attributes, we employed the following measures in our comparison: (a) Model fit is measured employing BIC, (b) prediction capability is evaluated using RMSE, (c) inference power was measured based on the number of distinct independent variables in the model and (d) computation complexity is measured in run times. The measures are defined such that the best performing model to has a value of 1 and the corresponding measures are generated relative to the best model. For BIC, for each model, the distance from the lowest fit model is measured and a ratio is computed as the ratio of the distance of the model from the lowest fit model to the corresponding model distance from the lowest fit model (we add 1 to both the numerator and the denominator for each ratio to avoid the 0/0 issue for the lowest fit model). Similarly, a normalization process has been applied to other measures as follows: 1) RMSE ratio: RMSE of best model/ RMSE of each model (the model with lowest RMSE will have a ratio of 1 and other models provides a RMSE ratio less than 1_; 2) Parameter ratio: total distinct independent variables in each model / total distinct parameters in the best model; and 3) Run time ratio: model with lowest run time/run times corresponding to each model (model with the fastest run times will have a ratio of 1 and other models will provide a run time ratio less than 1). Figure 3 presents attribute measures for the six model systems.

Several observations can be made from Figure 3. First, the models accommodating for unobserved heterogeneity always provide superior performance in terms of prediction, model fitness and inference power relative to its simpler counterparts (RPMNB vs UNB; PMNB vs PNB and LPMNB vs LPNB), however these models are usually associated with increased computational burden as indicated by the higher complexity in the figure. Second, among the simpler models (that do not accommodate for unobserved factors; UNB, PNB and LPNB); the latent segmentation model (LPNB) has the best goodness of fit, prediction and inference accuracy while having a slightly higher complexity rate relative to the other two models (UNB and PMNB). Third, interestingly, within the models accommodating for unobserved factors, the traditional RPMNB model provides inferior performance across all the measures. The recasted PMNB model is usually easy to estimate and also results in good prediction and inference power (this finding is supported by our previous work). However, our proposed LPMNB model provides the best inference and prediction capability, however this model comes with a moderate complexity rate and high run times (still lower than traditional RPMNB model). For instance, with six dependent variable and 3,815 observations, the RPMNB model took around 32 hours to converge while the LPMNB and PMNB model took 27 hours and 21 hours respectively.

Based on our multi-attribute analysis, we provide the following concluding thoughts: a) if an analyst wants to maximize the prediction and inference power irrespective of the complexity of the model, then the LPMNB model would be the preferred framework; b) If an analyst needs to check for complexity while not loosing prediction and inference power significantly, PMNB model would be the suitable one ; c) however, if the focus is entirely on model complexity i.e. an analyst wants to minimize the complexity while having relatively

good model fitness, prediction and inference power, the independent LPNB model is a good choice.

7 CONCLUSIONS

The current research contributes to burgeoning literature on multivariate models by proposing a model framework that (a) incorporates unobserved heterogeneity in a parsimonious framework and (b) allows for additional flexibility to accommodate for observed/systematic heterogeneity. Specifically, we extend our previous work (Bhowmik et al., 2019a) by addressing the population homogeneity assumption with the latent segmentation based approach employed for crash frequency modeling. Our current study contributes to crash frequency literature both methodologically and empirically by estimating a latent segmentation-based Panel Negative Binomial (LPNB) to study the zonal level crash counts across different crash types. Finally, we undertake a comparison exercise of the proposed LPMNB model with its' counterpart PMNB model proposed in our previous work to assess the importance of accounting for population heterogeneity in estimating zonal level crash frequency models. The analysis is conducted using the zonal level crash records from Central Florida for the year 2016 considering a comprehensive set of exogenous variables.

Based on the statistical data fit, we can conclude that the segmented model is a preferred choice as long as the framework is estimated in a closed form structure (independent models that do not account for unobserved heterogeneity; no need for simulation). However, when we rely on simulation for capturing the unobserved effects, the unsegmented model outperforms its' segmented counterparts. In an effort to assess the predictive performance of the estimated models, we compute several goodness fit of measures at disaggregate level including MPB (Mean prediction bias), MAD (mean absolute deviation), MAPE (mean absolute percentage error), RMSE (Root mean square error) and predictive log-likelihood for a discussion on estimating these measures). Specifically, we employ these measure on two datasets: 1) in-sample dataset: for the records used in the model estimation (sample size = 3,815 TAZs) and 2) holdout sample: records that are set aside for validation analysis (sample size = 932 TAZs). The resulting goodness of fit measures and predictive log-likelihood highlight the improved performance of LPMNB model over the PMNB model. Further, we compute aggregate level elasticity effects for both PMNB and LPMNB models to quantify whether the effect of variables significantly differs across the two frameworks. For this purpose, we identify a subset of exogenous variables including proportion of arterial roads, variance of speed limit, proportion of roads over 55mph and proportion of roads with separated median in a zone. For the LPMNB model, we estimate the aggregate level elasticities for the overall sample as well as for each segment separately to emphasize policy repercussions based on most critical contributory factors. From the elasticity effects, we can clearly see some significant differences across two segments for some variables which highlights the importance of allowing for population heterogeneity. Further, in terms of comparison across the two models adopted in the study, we found differences in elasticities across the two regimes. From the elasticity results, it is evident that allowing for a flexible specification (population heterogeneity) of observed and unobserved factors provides representative variable impacts. Further, to offer insights on model selection incorporating computational complexity dimension along with other important attributes, we conduct a multi-attribute comparison across the six different models comparing four attributes: (a) model fit, (b) prediction, (c) inference power and (d) computational complexity. The results highlight that our proposed LPMNB model provides the best inference and prediction capability, with a moderate complexity and higher run times.

Finally, the paper is not without its limitations. We evaluate zonal level (aggregate) crash counts for different crash types, and it would be useful to consider spatial correlation for such aggregate level planning analysis which could further improve the estimation process.

Moreover, it would be interesting to see if the findings are consistent with other spatial units and temporal periods.

ACKNOWLEDGMENT

The authors would also like to gratefully acknowledge Signal Four Analytics, Florida Department of Transportation and Department of Revenue for providing access to Florida crash data, geospatial data and land-use data.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: Tanmoy Bhowmik, Naveen Eluru, Shamsunnahar Yasmin; data collection: Tanmoy Bhowmik, Shamsunnahar Yasmin; model estimation and validation: Tanmoy Bhowmik, Shamsunnahar Yasmin, Naveen Eluru; analysis and interpretation of results: Tanmoy Bhowmik, Naveen Eluru, Shamsunnahar Yasmin; draft manuscript preparation: Tanmoy Bhowmik, Naveen Eluru, Shamsunnahar Yasmin. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

Afghari, A.P., Haque, M.M., Washington, S., 2020. Applying a joint model of crash count and crash severity to identify road segments with high risk of fatal and serious injury crashes. Accident Analysis and Prevention 144, 105615.

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. h

Alarifi, S.A., Abdel-Aty, M., Lee, J., 2018. A Bayesian multivariate hierarchical spatial joint model for predicting crash counts by crash type at intersections and segments along corridors. Accident Analysis and Prevention 119, 263–273.

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

Anastasopoulos, P.C., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. Accident Analysis and Prevention 43, 1140–1147.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41, 153–159.

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

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

Barua, S., El-Basyouny, K., Islam, M.T., 2014. A Full Bayesian multivariate count data model of collision severity with spatial correlation. Analytic Methods in Accident Research 3–4, 28–43.

Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B 35, 677–693.

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

Bhowmik, T., Rahman, M., Yasmin, S., Eluru, N., 2021. Exploring Analytical, Simulation-Based, And Hybrid Model Structures For Multivariate Crash Frequency Modeling.

Analytic Methods in Accident Research 100167.
Bhowmik, T., Yasmin, S., Eluru, N., 2019a. Do we need multivariate modeling approaches to model crash frequency by crash types? A panel mixed approach to modeling crash frequency by crash types. Analytic Methods in Accident Research 24.
Bhowmik, T., Yasmin, S., Eluru, N., 2019b. A multilevel generalized ordered probit fractional split model for analyzing vehicle speed. Analytic Methods in Accident Research 21, 13–31.
Bhowmik, T., Yasmin, S., Eluru, N., 2018. A joint econometric approach for modeling crash counts by collision type. Analytic Methods in Accident Research 19, 16–32.
Buddhavarapu, P., Scott, J.G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. Transp. Res. Part B. 91, 492–510. https://doi.org/10.1016/j.trb.2016.06.005
Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., Wang, X., 2018. Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. Analytic Methods in Accident Research 19, 1–15.
Chen, Y., Wang, K., King, M., He, J., Ding, J., Shi, Q., Wang, C., Li, P., 2016. Differences in factors affecting various crash types with high numbers of fatalities and injuries in China. PLoS One 11, 158559.
Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X., Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accident Analysis and Prevention 99, 330–341.
Chiou, Y.C., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accident Analysis and Prevention 50, 73–82.
Dey, B.K., Anowar, S., Eluru, N., Hatzopoulou, M., 2018. Accommodating exogenous variable and decision rule heterogeneity in discrete choice models: Application to bicyclist route choice. PLoS One 13.
Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accident Analysis and Prevention 70, 320–329.
Eluru, N., Bagheri, M., Miranda-Moreno, L.F., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. Accident Analysis and Prevention 47, 119–127.
Eluru, N., Bhat, C.R., 2007. A joint econometric analysis of seat belt use and crash-related injury severity. Accident Analysis and Prevention 39, 1037–1049.
Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis and Prevention 40, 1033–1054.
Fountas, G., Anastasopoulos, P.C., Mannering, F.L., 2018. Analysis of vehicle accident-injury severities: A comparison of segment- versus accident-based latent class ordered probit models with class-probability functions. Analytic Methods in Accident Research 18, 15–32.
Fountas, G., Fonzone, A., Gharavi, N., Rye, T., 2020. The joint effect of weather and lighting conditions on injury severities of single-vehicle accidents. Analytic Methods in Accident Research 27, 100124.
Fountas, G., Pantangi, S.S., Hulme, K.F., Anastasopoulos, P.C., 2019. The effects of driver fatigue, gender, and distracted driving on perceived and observed aggressive driving behavior: A correlated grouped random parameters bivariate probit approach. Analytic Methods in Accident Research 22, 100091.
Heydari, S., Fu, L., Miranda-Moreno, L.F., Joseph, L., 2017. Using a flexible multivariate

latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. Analytic Methods in Accident Research 13, 16–27.

Huo, X., Leng, J., Hou, Q., Zheng, L., Zhao, L., 2020. Assessing the explanatory and predictive performance of a random parameters count model with heterogeneity in means and variances. Accident Analysis and Prevention 147, 105759.

Intini, P., Berloco, N., Fonzone, A., Fountas, G., Ranieri, V., 2020. The influence of traffic, geometric and context variables on urban crash types: A grouped random parameter multinomial logit approach. Analytic Methods in Accident Research 28, 100141.

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

Kabli, A., Bhowmik, T., Eluru, N., 2020. A multivariate approach for modeling driver injury severity by body region. Analytic Methods in Accident Research 28, 100129.

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

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., Cai, Q., 2018. Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects. Accident Analysis and Prevention 111, 12–22.

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

Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic Methods in Accident Research 25, 100113.

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

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

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

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

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

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97, 246–273. https://doi.org/10.1016/j.jmva.2005.03.007

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

Wang, K., Bhowmik, T., Zhao, S., Eluru, N., Jackson, E., 2021. Highway safety assessment and improvement through crash prediction by injury severity and vehicle damage using Multivariate Poisson-Lognormal model and Joint Negative Binomial-Generalized Ordered Probit Fractional Split model. Journal of Safety Research76, 44–55.

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

and Prevention 99, 6–19.
Wang, Y., 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.
Wedel, M., Desarbo, W.S., Bult, J.R., Ramaswamy, V., 1993. A latent class poisson regression model for heterogeneous count data. Journal of Applied Econometrics 8, 397–411.
Xiong, Y., Tobias, J.L., Mannering, F.L., 2014. The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity. Transportation Research Part B 67, 109–128.
Yasmin, S., Bhowmik, T., Rahman, M., Eluru, N., 2021. Enhancing non-motorist safety by simulating trip exposure using a transportation planning approach. Accident Analysis and Prevention 156, 106128.
Yasmin, S., Eluru, N., 2018. A joint econometric framework for modeling crash counts by severity. Transportmetrica A 14, 230–255.
Yasmin, S., Eluru, N., 2016. Latent segmentation based count models: Analysis of bicycle safety in Montreal and Toronto. Accident Analysis and Prevention 95, 157–171. h
Yasmin, S., Eluru, N., Lee, J., Abdel-Aty, M., 2016. Ordered fractional split approach for aggregate injury severity modeling. Transportation Research Reecord 2583, 119–126.
Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Science 47, 443–452.
Yu, H., Li, Z., Zhang, G., Liu, P., 2019. A latent class approach for driver injury severity analysis in highway single vehicle crash considering unobserved heterogeneity and temporal influence. Analytic Methods in Accident Research 24, 100110.
Yu, R., Abdel-Aty, M., 2013. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. Accident Analysis and Prevention 58, 97–105.

Table 1 Summary Statistics of Exogenous Variables (Zonal Level)

Variables	Definition	Zonal (N=4,747)			
		Minimum	Maximum	Mean	Std. Deviation
Dependent Variables					
Rear-end	Number of rear-end crashes in a zone	0.000	243.000	10.947	18.517
Angular	Number of angular crashes in a zone	0.000	104.000	4.215	6.816
Sideswipe	Number of sideswipe crashes in a zone	0.000	66.000	2.686	5.228
Head-on	Number of head-on crashes in a zone	0.000	24.000	0.344	1.028
Single Vehicle	Number of single vehicle crashes in a zone	0.000	51.000	2.366	3.573
Non-motorized	Number of non-motorized crashes in a zone	0.000	12.000	0.719	1.318
Roadway Characteristic					
Proportion of rural road	(Rural road length/total road length)	0.000	1.000	0.121	0.309
Proportion of urban road	(Urban road length/total road length)	0.000	1.000	0.806	0.381
Proportion of arterial road	(Arterial road length/total road length)	0.000	1.000	00377	0.393
Number of Intersection	Ln (no of intersection)	0.000	4.682	1.921	1.053
Signal intensity	Total number of traffic signal per 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[4]: 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
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

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

Table 2 Measure of Fit for Different Models

Model	*Log-Likelihood*	*No. Of Parameters*	*AIC*	*Corrected AIC*	*BIC*
Model without Unobserved heterogeneity					
Univariate NB Model	-39954.90	68.00	80045.80	80048.30	80592.41
Panel independent NB model (PNB)	-39961.82	52.00	80027.64	80029.10	80352.47
Latent Segmentation Independent NB with two segments (LPNB II)	-39890.40	57.00	79894.81	79896.57	80250.87
Latent Segmentation Independent NB with three segments (LPNB III)	-39839.82	58.00	79795.63	79797.45	80157.94
Model with Unobserved heterogeneity					
Panel Mixed NB (PMNB)	-39235.75	53.00	78577.50	78579.02	78908.57
Latent Segmentation Mixed NB with two segments (LPMNB II)	-39352.26	57.00	78818.52	78820.28	79174.58

Table 3 Segment Characteristics for LPMNB model

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

Table 4 LPMNB Model Results

Segment Component				
Variables	***Segment 1***		***Segment 2***	
	Coeff.	T-stat	Coeff.	T-stat
Constant	--	--	1.532	11.898
Number of intersections	--	--	-0.660	-14.394
Average outside shoulder width	--	--	0.897	13.163
Urban Area (acre)	--	--	-0.534	-21.139
Residential area	--	--	0.056	2.932
Crash Count Component				
Crash Specific Characteristic				
Rear-end	-0.171	-3.372	-3.298	-13.797
Angular	-1.654	-27.320	-4.363	-13.680
Sideswipe	-0.325	-6.400	-4.225	-11.594
Single Vehicle	-0.345	-8.048	-3.185	-14.410
Head-on	-2.882	-18.544	-4.227	-12.654
Non-motorized	-2.040	-15.908	-5.338	-14.331
Roadway Characteristics				
Proportion of arterial roads				
Rear-end+angular+NMT	0.166	4.933	--	--
All single vehicle	-0.260	-4.087	-0.472	-3.312
Signal Intensity				
Rear-end+sideswipe+NMT	--	--	2.350	3.479
Single vehicle	--	--	-1.760	-1.686
Variance of speed limit				
Rear-end+sideswipe+NMT	0.036	5.167	0.133	5.244
Road length over 55mph				
Rear-end+sideswipe	0.846	12.212	--	--
Angular	-2.058	-11.470	--	--
Standard Deviation	0.452	1.904	--	--
Single vehicle	0.846	12.212	0.753	2.921
Standard Deviation	--	--	0.930	3.149
Head-on	-2.103	-4.559	--	--
Non-motorized	-1.900	-6.312		
Roads with separated median				
Rear-end+angular+sideswipe	--	--	0.925	6.286
Head-on	--	--	-0.276	-1.138
Land Use Characteristics				
Office area (acre)				
Rear-end+sideswipe	0.195	20.947	--	--
Non-motorized	0.169	6.687	--	--
Built Environment Characteristics				
Number of restaurants				
Rear-end+sideswipe	0.192	13.919	--	--
Non-motorized	0.190	6.635		
Number of shopping centers				

Rear-end+sideswipe	0.034	2.676	--	--
Traffic Characteristics				
VMT				
Angular+sideswipe	0.147	45.205	--	--
Head-on	0.171	10.550	--	--
Non-motorized	0.102	8.365		
Truck VMT				
Rear-end	--	--	0.418	14.386
Single vehicle	--	--	0.554	21.272
Socio-economic Characteristics				
Non-motorist commuter				
Rear-end+sideswipe+NMT	0.076	3.924	--	--
Angular	0.170	8.790		
Transit commuter	--	--	--	--
Rear-end+ Non-motorized	0.217	11.883	0.576	8.584
Over Dispersion Parameter				
Rear-end	0.279	9.521	0.965	11.865
Angular	0.190	7.825	1.512	3.064
Sideswipe	0.284	8.294	0.965	11.865
Single Vehicle	0.726	17.746	0.115	1.554
Head-on	0.190	7.825	1.512	3.064
Non-motorized	0.279	9.521	0.965	11.865
Correlations				
Rear-end+NMT	0.679	23.659	--	--
Angular+sideswipe+single vehicle	0.840	34.284	1.245	7.913
Log-likelihood at zero: -44378.90; log likelihood at convergence: -39890.40				

Table 5 PMNB Model Results

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

Socio-demographic characteristics												
Non-motorists (3)	0.070	3.408	0.148	6.956	0.164	6.505	--	--	--	--	0.070	3.408
Transit users (1)	0.239	13.596	--	--	--	--	--	--	--	--	0.239	13.596
Over dispersion (6)	0.396	31.904	0.384	14.952	0.396	31.904	0.384	14.952	0.700	22.059	0.396	31.904
Unobserved Effects												
Correlation 1 (1)	0.741	33.753	--	--	--	--	--	--	--	--	0.741	33.753
Correlation 2 (1)	--	--	0.936	40.216	0.936	40.216	0.936	40.216	--	--	--	--
Log-likelihood at zero: -44541.65; log likelihood at convergence: -39235.75												

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

Table 6 Predictive Performance Measure of Two Models (PMNB and LPMNB)

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

Table 7 Elasticity Effects Across Two Models (PMNB and LPMNB)

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

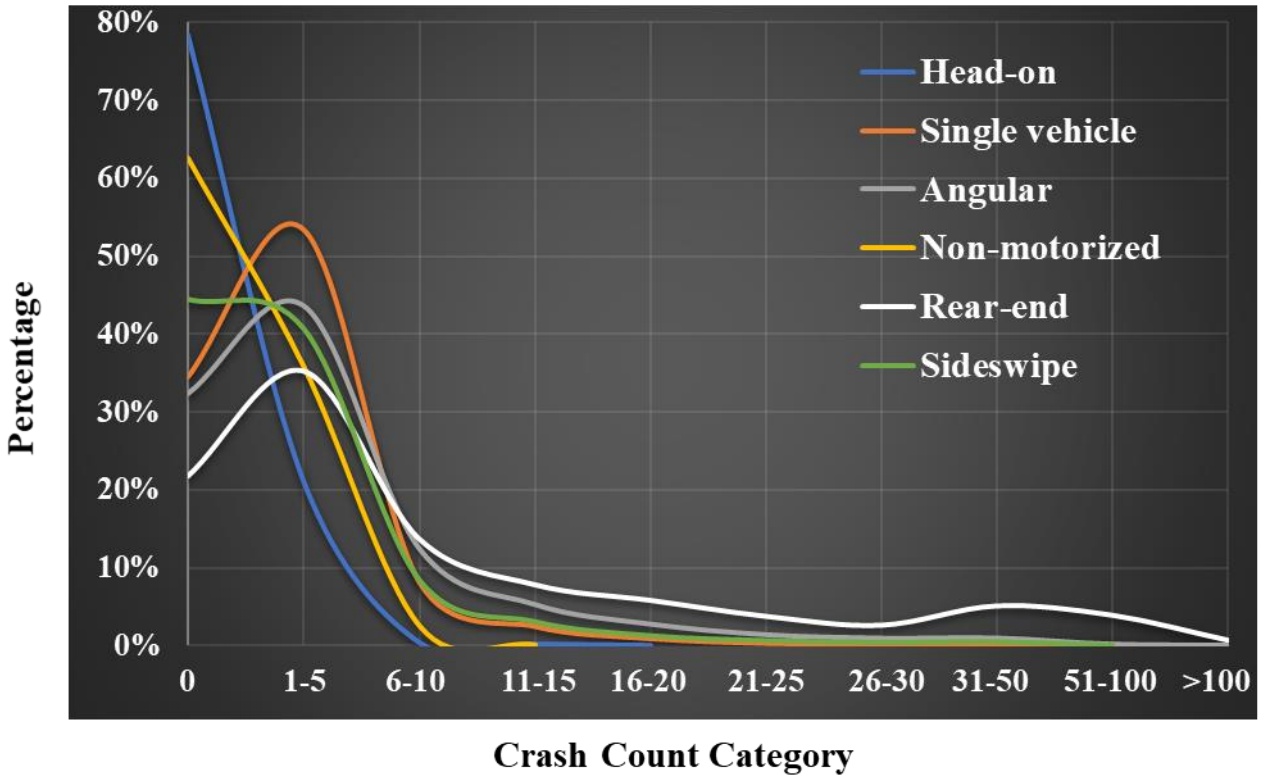
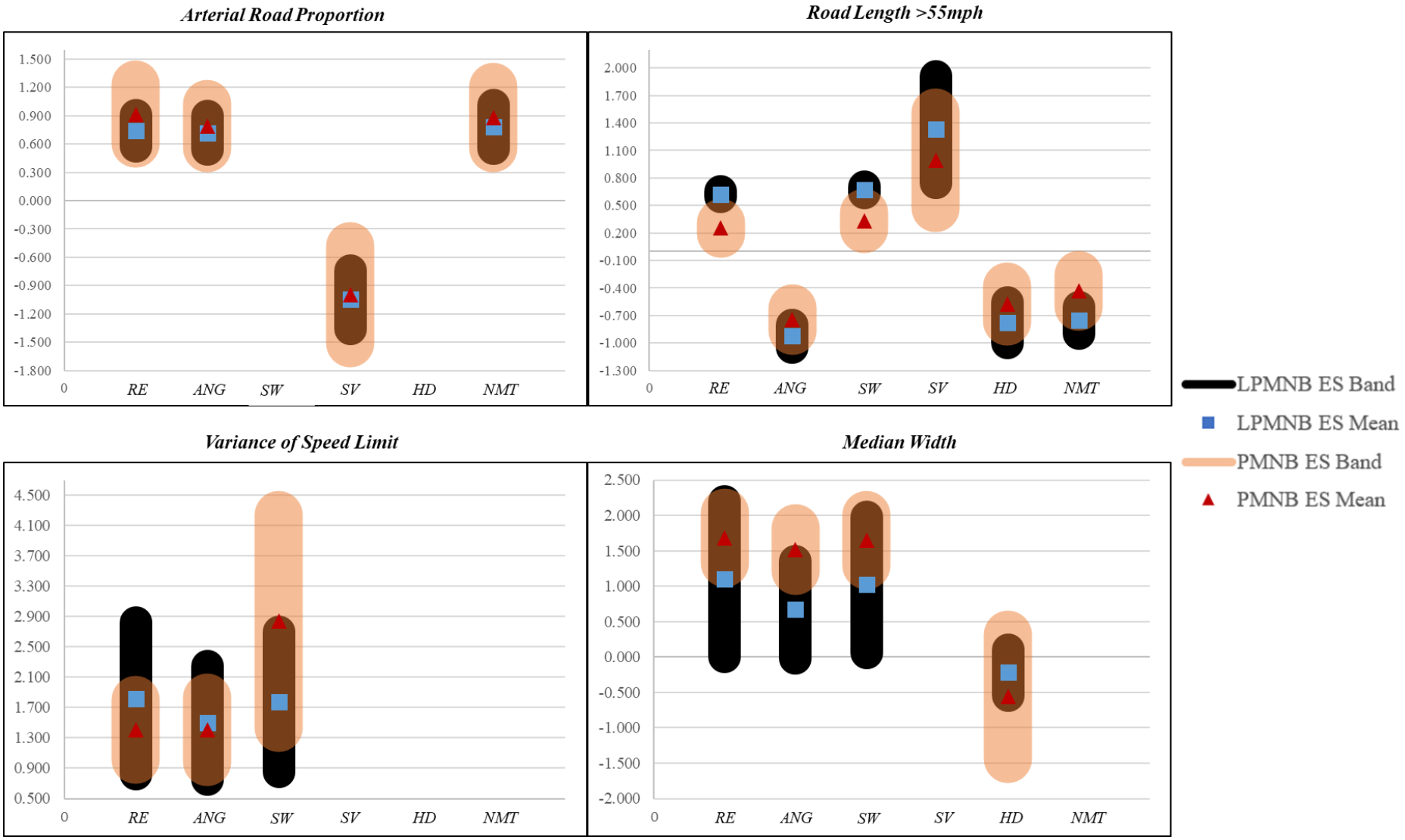
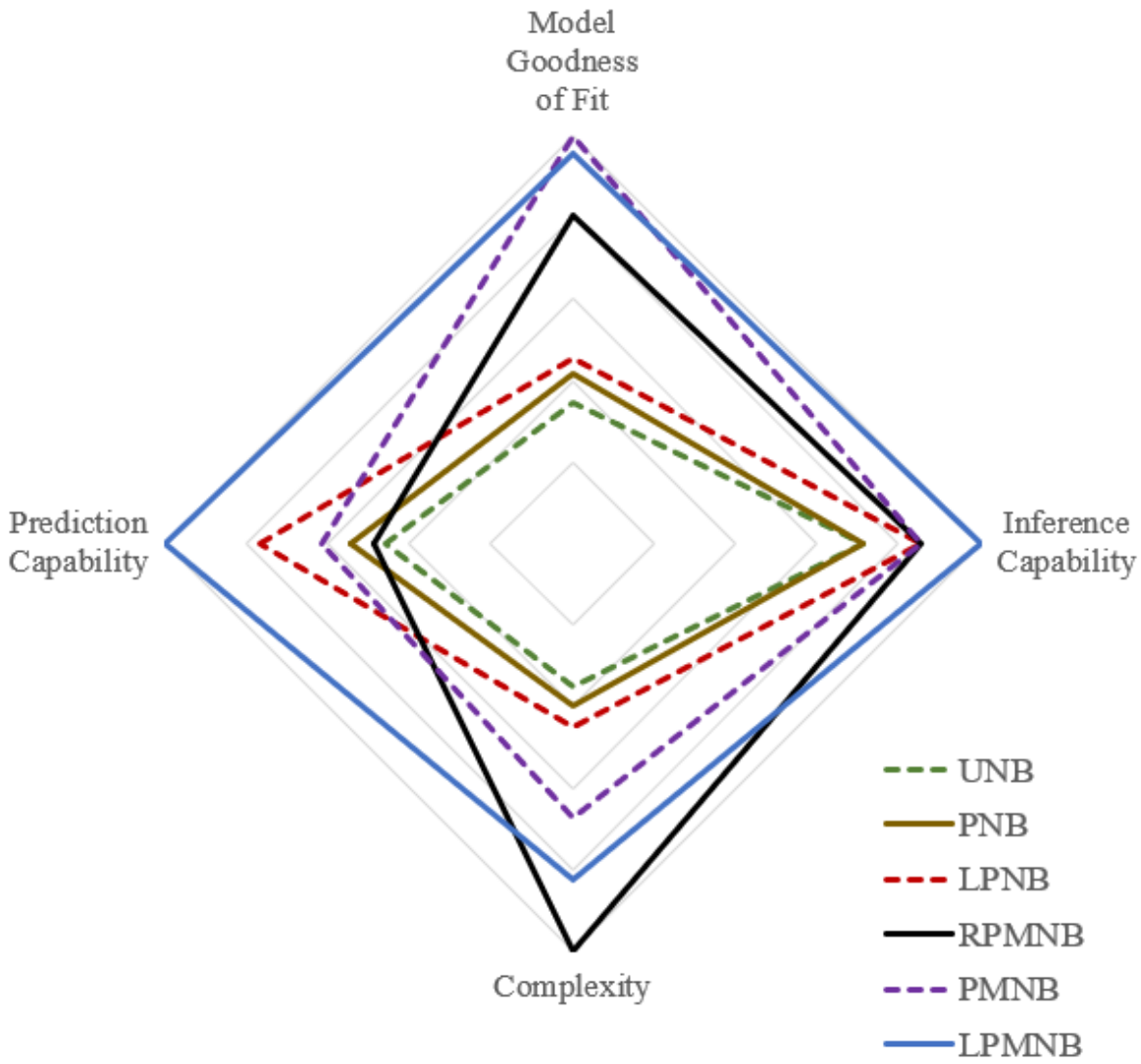


Figure 1: Distribution of Different Crash Types



*ES: Elasticity, RE: Rear-end; ANG: Angular; SW: Sideswipe; SV: Single vehicle; HD: Head-on; NMT: Non-motorized

Figure 2: Elasticity Band of LPMNB and PMNB Model



Note: UNB: Univariate Negative Binomial model; RPMNB: Random parameter multivariate negative binomial model

Figure 3: Modelling Trade-off Across Inference capability, Prediction Capability and Complexity