A New Econometric Approach for Modeling Several Count Variables: A Case Study of Crash Frequency Analysis by Crash Type and Severity

Tanmoy Bhowmik*
Post-doctoral Scholar
Department of Civil, Environmental & Construction Engineering
University of Central Florida, USA
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, USA
Tel: 407-823-4815, Fax: 407-823-3315
Email: naveen.eluru@ucf.edu
ORCiD number: 0000-0003-1221-4113

*Corresponding author

ABSTRACT
There is limited adoption of research modeling crash severity frequency considering different crash types due to the challenge associated with analyzing large number of dependent variables. The proposed research contributes to burgeoning econometric and safety literature by developing a joint modeling approach that can accommodate for several dependent variables within a parsimonious structure. By recasting the analysis levels for dependent variables, the proposed approach allows for flexible consideration of crashes by type and severity within a single framework. Specifically, we employ a Panel Mixed Negative Binomial- Generalized Ordered Probit Fractional Spilt (PMNB-GOPFS) model where the first component (NB) accommodates for crash frequency by crash type and the later component (GOPFS) studies the fraction of severity outcome for different crash types. The proposed model system increases interaction between dependent variables through observed variables thus reducing the dependency on unobserved interactions across dependent variables. Thus, the proposed approach allows for the estimation of parsimonious specifications reducing the need for computationally intensive simulation based estimation. The proposed system is also flexible to accommodate for common unobserved effects including: 1) common unobserved factors simultaneously affecting crash counts of different crash types; 2) common unobserved factors simultaneously affecting crash severity proportions of different crash types; and 3) common unobserved factors that simultaneously impact crash counts and severity proportions by different crash types. The model system performance is illustrated using a simulation study. The empirical analysis was conducted using zonal level crash count data for the year 2016 from Central Florida while considering a comprehensive set of exogenous variables including roadway, built environment, land-use, traffic and sociodemographic characteristics. To illustrate the applicability of our proposed system, we carried out a comparison exercise between our proposed joint PMNB-GOPFS and the traditional multivariate system for predicting crash counts across different crash severities. The resulting goodness of fit measures clearly highlight the superior/equivalent performance of the proposed PMNB-GOPFS model over the traditional RPMNB model with less than half the number of parameters. The proposed framework can predict several dimensions including total crash counts, total crash counts by crash types, crash counts for each severity level and finally, proportions and counts of crashes for each crash type by severity.

Keywords: Crash type; Crash severity; Panel model; Fractional split model; Unobserved heterogeneity; Multivariate model

1 BACKGROUND

Road traffic crash related morbidity and mortality is acknowledged to be a global challenge. Annually, it is reported that more than one and a quarter million people die in road traffic crashes with the number expected to exceed 2 million by 2020 (WHO, 2018). While many developed countries (such as Canada, and Japan) have been able to achieve a reduction in the number of road crash related fatalities, in the United States, the reduction rate is much lower (or worse with an occasional increase as observed in 2014). In reducing the burden of such unavoidable incidents, safety researchers are continually investigating approaches for crash occurrence reduction and crash consequence mitigation. A major analytical tool employed for examining the critical factors influencing crash occurrence include the econometric crash frequency models (Geedipally et al., 2010; Jonathan et al., 2016; Yan et al., 2009). The proposed research contributes to burgeoning econometric and safety literature by developing a joint modeling approach that can accommodate several dependent variables within a parsimonious structure.

The traditional modeling framework for crash frequency analysis is the univariate frequency model such as Poisson, Negative binomial or the Poisson-Lognormal model (see (Bhowmik et al., 2018; Bhowmik, 2020; Lord and Mannering, 2010) for a detailed review of these studies). In these studies, for an observational unit, the modeling variable of interest is typically the total number of crashes. The approach of aggregating all crashes into a single dependent variable can result in aggregation bias and a loss of information available in the dataset. For instance, consider two zones with 5 observed crashes in the analysis period. For zone 1, the 5 crashes include 5 head-on crashes while for zone 2, the 5 crashes include 4 rear-end crashes and 1 vehicle pedestrian crash. While the crash distribution by crash type across the two zones is quite distinct, an approach focusing on total crashes will consider both zones as having identical dependent variables. The aggregation would make it quite cumbersome to accurately estimate the impact of independent variables on total crashes. For example, in zone 1, geometric design inadequacies might be the reason for head-on crashes while in zone 2, the presence of a significant number of signalized urban intersections might be the reason for rear-end and pedestrian crashes. A single *total crash* model will not be able to parse these distinctions accurately. Hence, it is not surprising that in recent years, safety researchers have focused on disaggregating the data by various attributes such as crash typology (such as head-on or rear-end), injury severity (such as crashes by no injury or crashes by severe injury) and crash location (such as intersection versus non-intersection).

The proposed disaggregation of the crash frequency variable increases the complexity of the modeling effort and presents many additional challenges. The number of dependent variables of interest increase based on the attribute levels of interest. For analyzing these multiple dependent variables, multiple univariate models with frequency by attribute levels (such as crashes by crash type) will need to be estimated. While developing multiple univariate crash frequency models will account for the influence of independent variables, these models ignore that the multiple crash frequency variables for a traffic analysis zone (TAZ) are potentially correlated. The different outcome dimensions (such as crash type and severity outcomes) under consideration are possibly correlated as both of these outcomes (within and across) share the same travel environment within a spatial planning unit over a specific given period of time. For example, for zonal level crash frequency analysis, it is possible that characteristics specific to the zone such as driver behavior, geometric design and build quality (possibly of higher or lower quality relative to the other zones) and traffic signal design objectives might influence different crash counts by crash type (such as head-on, rear-end). For instance, a higher presence of drivers using cell phone while driving in a

zone (information usually unavailable to the analyst) may contribute to a higher number of head on crashes. At the same time, due to the greater dissipation of kinetic energy associated with a head-on collision, the likelihood of serious injury crashes will be higher for the same unobserved factor (cell phone distracted driver proportion). This is an example of how one unobserved variable can significantly affect the two dimensions simultaneously (crash type and severity). Ignoring the presence of such correlation may result in biased parameter estimates and potentially incorrect model predictions and/or inefficient policy implications (see Liu and Sharma, 2018; Mannering et al., 2016; Zeng et al., 2018, Wang et al., 2019 for an extensive discussion). Thus, any modeling approach to analyze the multiple crash frequency variables needs to explicitly account for the presence of these common factors that are most often unobserved. The most common approach employed to address the potential unobserved heterogeneity in safety literature is the development of multivariate crash frequency models.

1.1 Earlier Work

A summary of earlier research efforts investigating crash frequencies by crash type and severity level are presented in Table 1 with information on the spatial unit (aggregation level), the region (covered area, for example state or city), crash unit (type of crash considered), number of dimensions examined (of the dependent variable), methodological framework employed, and different categories of exogenous variables considered in the analysis. The following observations can be made from Table 1. First, the most prevalent mechanism to analyze crash count by different levels are multivariate count regression approaches. Second, several spatial units are considered both at macro and micro level for analyzing the crash counts by type and injury severity including segments and intersections (for micro level); and census block and traffic analysis zone (for macro level). Third, the methodological frameworks adopted in these studies include Negative binomial, Poisson regression, Multivariate Poisson-lognormal, Multivariate Negative Binomial, Multinomial Generalized Poisson and Integrated Nested Laplace Approximation. Fourth, with respect to exogenous variables, the overall findings from earlier research effort are consistent. The various factors identified that influence crash severities include - (1) roadway characteristics such as shoulder width, arterial road length; (2) land-use characteristics such as urban land use and land use mix; (3) built environment characteristics such as number of access points (number of restaurant, entertainment center); (4) traffic characteristics such as Average Annual Daily Traffic (AADT) and truck volume; (5) socio-demographic characteristics such as population density and people by different age group; and (6) weather variables such as precipitation rate. Fifth, the highest number of dependent variables considered in multivariate models is 8. Finally, none of the studies[1] examined the crash counts of different crash types and their corresponding severity outcomes in an integrated framework at the planning level.

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

TABLE 1 Summary of Existing Aggregate Level Multivariate Crash Type and Severity Studies

Studies	Spatial Unit	Region	Crash unit	Number of Levels Explored	Methodological Approach	Independent Variables Considered					
						Roadway Infrastructure	*Land-use*	*Built Environment*	*Traffic Characteristics*	*Socio-demographic*	*Weather*
Crash Type Studies											
(Jonsson et al., 2007)	Highway segments (Micro)	State (California)	Motorized crash	4 (same direction, intersecting direction, opposite direction, single vehicle crashes)	Generalized linear model	√	√	--	√	--	--
(Ye et al., 2009)	Intersections (micro)	County (Georgia)	Any Crash	7 (angle, head-on, rear-end, sideswipe: same and opposite direction, pedestrian)	Multivariate Poisson regression model	√	--	--	√	--	--
(El-Basyouny et al., 2014a)	Citywide (Macro)	City (Edmonton)	Motorized crash	7 (FTC*, FOTS**, SSV***, left turn, ILC****, parked vehicle and off-road)	Multivariate Poisson-lognormal model	--	--	--	--	--	√
(El-Basyouny et al., 2014b)	Citywide (Macro)	City (Edmonton)	Motorized crash	7 (FTC*, FOTS**, SSV***, left turn, ILC****, parked vehicle and off-road)	Multivariate Poisson-lognormal model	--	--	--	--	--	√
(Li et al., 2015)	Freeway (micro)	State (Florida)	Motorized crash	3 (rear-end, sideswipe and angle)	Multivariate Poisson-lognormal model	√	--	--	√	--	--
(Mothafer et al., 2016)	Highway segments (Micro)	State (Washington)	Motorized crash	4 (rear-end, sideswipe, fixed object and others)	Multivariate Poisson gamma mixture count model	√	--	--	√	--	--
(Jonathan et al., 2016)	Road segments (Micro)	County (Pennsylvania)	Any Crash	4 (same direction, opposite direction, angular, fixed object)	Multivariate Poisson-lognormal spatial model	√	--	--	√	--	--
(Serhiyenko et al., 2016)	Highway segments (Micro)	State (Connecticut)	Motorized crash	3 (same direction, opposite direction, single vehicle crash)	Multivariate Poisson-lognormal model	√	√	--	√	--	--
(Cheng et al., 2017)	Intersections (micro)	City (California)	Motorized crash	6 (rear-end, head-on, sideswipe, broad side,	Multivariate Poisson-lognormal model	√	--	--	√	--	--

				hit object crash, others)							
(Wang et al., 2017)	Road segments, intersections (Micro)	State (Minnesota, Washington)	Any Crash	4 (same direction, intersecting direction, opposite direction, single vehicle crashes)	Multivariate Poisson-lognormal model	√	--	--	√	--	--
(Alarifi et al., 2018)	Road segments, intersections (Micro)	County (Florida)	Any Crash	6 (same direction, angular, opposite direction, non-motorized, single vehicle and others)	Bayesian multivariate hierarchical spatial joint model	√	--	--	√	--	--
(Bhowmik et al., 2018)	STAZ (Macro)	State (Florida)	Motorized crash	8 (rear-end, angular, sideswipe, head-on, single vehicle, off-road, rollover and others)	Multivariate negative binomial model, multinomial fractional split model	√	√	√	√	--	--
(Hosseinpour et al., 2018)	Road segments (micro)	Nation (Malaysia)	Motorized crash	4 (head-on, rear-end, angle, and sideswipe)	Multivariate Poisson regression model	√	--	--	√	--	--
(Yasmin et al., 2018)	STAZ (Macro)	State (Florida)	Motorized crash	4 (light truck, van, other vehicle and non-motorized)	Copula-based multivariate NB model	√	√	--	√	√	--
(Guo et al., 2019a)	Freeway (micro)	State (Florida)	Motorized crash	3 (rear-end, sideswipe and angle)	A random parameters multivariate Poisson-lognormal model	√	--	--	√	--	--
(Guo et al., 2019b)	Freeway (micro)	State (Florida)	Motorized crash	3 (rear-end, sideswipe and angle)	A random parameters multivariate Tobit model	√	--	--	√	--	--
(Bhowmik et al., 2019a)	TAZ (Macro)	State (Florida)	Any Crash	6 (rear-end, angular, sideswipe, single vehicle, other multi vehicle and non-motorized)	Panel mixed negative binomial model	√	√	√	√	--	--
(Bhowmik et al., 2021)	STAZ (Macro)	State (Florida)	Any Crash	4 (intersection, on-road, off-road and non-motorized)	Copula Based random parameter multivariate NB model	√	√	√	√	--	--
Crash Severity Studies											

(Narayanamoorthy et al., 2013)	Census tract (Macro)	Region (Manhattan)	Non-motorized Crash	4 (possible injury, non-incapacitating injury, incapacitating injury and fatal injury)	Generalized ordered-response model with Composite Maximum Likelihood	√	√	√	--	√	--
(Ye et al., 2013)	Freeway segment (Micro)	State (Washington)	Any Crash	3 (PDO, possible injury, injury/ fatality)	Joint Poisson regression model	√	--	--	√	--	√
(Barua et al., 2014)	Road segment (Micro)	City (Richmond, Vancouver)	Any Crash	2 (no injury and injury/fatal crashes)	Multivariate Poisson lognormal model	√	√	√	√	--	--
(Chiou et al., 2014)	Freeway segment (Micro)	State (Taiwan)	Motorized Crash	3 (PDO, possible injury, injury/ fatality)	Multinomial Generalized Poisson with error components	√	--	√	√	--	√
(Chiou and Fu, 2015)	Freeway segment (Micro)	State (Taiwan)	Motorized Crash	3 (PDO, possible injury, injury/ fatality)	Multinomial generalized Poisson with spatiotemporal error components	√	--	√	√	--	√
(Zhan et al., 2015)	Census tract (Macro) Roadway segment (Micro)	City, State (New York, Washington)	Pedestrian and Motorized Crash	3 (no injury, possible injury and evident injury)	Multivariate Poisson-lognormal model	√	√	√	√	√	√
(Anastasopoulos, 2016)	Highway segments (Micro)	State (Indiana)	Motorized crash	3 (PDO, injury and fatality)	Random parameter multivariate tobit model, Multivariate zero-inflated negative binomial model	√	√	--	--	--	--
(Barua et al., 2016)	Road segment (Micro)	City (Vancouver)	Any Crash	2 (no injury and injury crashes)	Bayesian multivariate random parameters spatial model	√	√	√	√	--	--
(Dong et al., 2016)	Intersection (Micro)	State (Tennessee)	Any Crash	2 (disabling injury and non-disabling injury)	Random parameter bivariate zero-inflated negative binomial model	√	--	--	√	--	--
(Yasmin et al., 2016)	Road segments (Micro)	State (Florida)	Motorized Crash	5 (no injury, minor injury, moderate injury, serious injury, fatal)	Ordered fractional split model	√	√	√	√	√	--

(Bhat et al., 2017)	Census tract (Macro)	Region (Manhattan)	Pedestrian Crash	4 (possible injury, non-incapacitating injury, incapacitating injury and fatal injury)	Random coefficients multivariate count model	√	√	√	--	√	--
(Boulieri et al., 2017)	Ward (Macro)	England	Any Crash	2 (slight accidents, fatal accidents	Multivariate Bayesian Model	√	--	--	√	--	--
(Chen et al., 2017)	Highway segment (Micro)	State (Indiana)	Motorized Crash	3 (PDO, possible injury, and injury/fatality)	Multivariate Random Parameters Negative Binomial Approach	√	--	--	√	--	--
(Ma et al., 2017)	Highway segment (Micro)	Interstate I70 (Colorado)	Motorized Crash	2 (injury, no injury)	Multivariate Poisson lognormal (normal, spatial and spatio-temporal)	√	--	--	√	--	√
(Wang et al., 2017)	Road segments, intersections (Micro)	State (Minnesota, Washington)	Any Crash	3 (no injury, possible/non-incapacitating injury and fatal/incapacitating injury crashes)	Multivariate Poisson Lognormal model	√	--	--	√	--	--
(Zeng et al., 2017)	Census tract (Macro) Roadway segment (Micro)	City (Hong Kong)	Any Crash	2 (slight injury crash and killed/seriously injured crashes)	Multivariate Poisson-lognormal model	√	--	--	√	--	√
(Liu and Sharma, 2018)	County (macro)	State (Iowa)	Any Crash	3 (Fatal crashes, major injury crashes, and minor injury crashes)	Multivariate spatio-temporal Bayesian model	√	--	--	√	√	√
(Yasmin and Eluru, 2018)	STAZ (Macro)	STAZ (Macro)	Motorized Crash	4 (no injury, minor injury, incapacitating injury and fatal)	Joint NB-ordered fractional split model	√	√	√	√	√	--
(Lee and Khattak, 2019)	Road segments, (Micro)	City (Lincoln)	Motorized Crash	5 (no injury, minor injury, moderate injury, serious injury, fatal)	Network-based local spatial auto-correlation	√	--	--	√	--	--
(Shaon et al., 2019)	Highway segment (Micro)	State (Wisconsin)	Any Crash	2 (No injury, injury)	Multivariate multiple risk source regression model	√	--	--	√	--	--
(Xie et al., 2019)	Census tract (macro)	City (Manhattan)	Any Crash	3 (no injury, serious and fatal)	Multivariate Conditional	√	√	--	√	--	--

					Autoregressive (MVCAR) model						
(Zeng et al., 2019)	TAZs (macro)	Metropolitan area (Hong Kong	Any Crash	3 (no injury, slight and KSI*****)	Bayesian multivariate random-parameters spatio-temporal Tobit regression	√	--	--	√	--	--
(Huang et al., 2019)	TAZs (macro)	County (Florida)	Any Crash	2 (no injury and injury crashes	Bayesian multivariate random parameters spatial model	√	--	--	√	√	--
(Afghari et al., 2020)	Road segment (Micro)	State (Queensland)	Any Crash	3 (minor, serious injury and fatal)	Joint NB-ordered fractional split model	√	--	--	√	--	--

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

1.2 Current Study in Context and Study Contributions

In multivariate count regression approaches described above, the impact of exogenous variables is quantified through the propensity component of count models. In accommodating the influence of unobserved effects, in general, these approaches partition the error components as a common term and an independent term across dependent variables (see (Mannering et al., 2016) for a detailed discussion of various methodologies). The approaches rely either on Maximum Simulated Likelihood (MSL) or Markov Chain Monte Carlo (MCMC) approach in the Bayesian realm for model estimation. MSL and MCMC methods provide substantial flexibility in accommodating for unobserved heterogeneity.

While several research efforts have developed multivariate crash frequency models for a small number of dimensions (such as 5); there is limited adoption of multivariate approaches for count variables in the presence of larger number of dependent variables (say greater than 15). For example, consider the development of crash frequency models by crash type (say N types) and severity level (say K levels). In the currently employed approaches, the number of crash propensity equations to be estimated will be N*K. While the estimation of $N * K$ univariate model systems is repetitive, it is still feasible. However, accommodating for unobserved heterogeneity with a large number of dependent variables is substantially challenging. The probability evaluation with high dimensional integrals is potentially affected by several challenges including - requirements of generating high dimensionality of random numbers, empirical identification issues due to relatively flat objective functions in larger dimensions and longer computational run times. Furthermore, the stability of the variance-covariance matrix is often sensitive to model specification and number of simulation draws (Bhat, 2011).

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

Second, we draw on another recent work that recasts the multivariate distributional problem (for multiple crash frequency dependent variables) as a repeated measure univariate problem (see (Bhowmik et al., 2019a) for detail). For example, crash frequency by crash type is represented as a repeated measure of crash frequency variable recognizing that each repetition represents a different crash type instead of considering it as a multivariate distribution. The recasting process allows for the estimation of a parsimonious model system by allowing for an improved specification testing of variable impacts across different crash types (see Bhowmik et al., 2019a for detail). Using this consideration, the proposed model system enhances the efficiency of estimation through a single crash frequency model and a single crash proportion model, while also allowing for parameter effects to vary across different crash types through crash type specific deviation terms. Building on this study design, the $N * 2 * (L_1 + L_2)$ could potentially be reduced to $2 * (L_1 + L_2)$. Of course, we envision that the exact number of parameters to be estimated will lie somewhere in the range between $2 * (L_1 + L_2)$ and $N * 2 * (L_1 + L_2)$. The reduction in parameters especially for unobserved factors will contribute to substantial improvements in model efficiency and computational times. In our current study, the number of dependent variables analyzed is 24 (N=6 and K=4). Through our innovative adoption of the fractional split model for severity modeling and recasting of the crash frequency model from a multivariate to a repeated univariate structure we potentially reduce the number of parameters to be estimated to $2 * (L_1 + L_2)$ in the best case and $6 * 2 * (L_1 + L_2)$ in an absolute worst case. To summarize, the proposed model system results in the estimation of about $\frac{1}{12}$ of the model parameters in the best case and about $\frac{1}{2}$ in the worst case. From any metric of comparison this is a substantial improvement.

Third, the proposed approach reduces the reliance on unobserved heterogeneity by allowing for direct impact of observed variable interaction across dependent variables. To elaborate, in traditional models, the interaction between the multiple dependent variables is accommodated through unobserved error correlations. Typically, these correlations are captured through simulation based approaches or specialized dependency structures (such as copulas). For example, in our empirical setting with 6 crash types and 4 crash severities, the traditional approach will need to estimate 24 (6*4) separate equations and then test for the presence of $24C_2$ potential correlation terms. The additional estimation and computational complexity explains why earlier research has focused either on crash type or crash severity only. In our approach, we allow for observed variable interaction in modeling severity proportions. Thus, we actually allow for interaction across dependent variables without needing for the introduction of simulation based error terms. Further, we consider unobserved correlations across dependent variables as needed. Given the direct interaction across the dependent variables, we are likely to have fewer parameters to test because of the model structure. This improves model estimation efficiency as maximum simulated models with large dimensions of integrals are less efficient and are prone to potential inaccuracies (see Bhat, 2011).

Finally, the proposed model structure allows for recognizing the potential ordering in the severity dependent variables. In the traditional model setting, with multivariate models for multiple dependent variables accommodating for such ordering is not possible. Ignoring this potential ordering can possibly manifest itself through the significance of multiple unobserved variables representing correlation across dependent variables (that are actually ordered). Further, as illustrated in existing literature (see Eluru and Yasmin, 2015; Fountas and Anastasopoulos, 2017; Xin et al., 2017; Bhowmik et al., 2019b; Kabli et al., 2020; Wang et al., 2021 for detail), adopting a generalized ordered framework that relaxes the restrictive assumptions of the ordered

outcome model (also referred to as parallel lines assumption) by allowing the threshold parameters to vary in response to observational attributes would be more representative. Thus, in our proposed model structure, we also develop a generalized ordered model structure that also accommodates for the potential parallel line assumption documented in ordered literature.

The model is estimated using zonal level crash count, crash type and severity data for both motorized and non-motorized crashes. The crash data is extracted for the year 2016 from Central Florida region of the USA. The dimension of the dependent variables analysed is 24 $[(6 * 4)$ from 6 crash types (rear-end, angular, sideswipe, head-on, single vehicle and non-motorist crash) and 4 severity levels (severe (fatal and incapacitating as one category), non-incapacitating, possible injury and property damage).

In summary, the current study contributes to safety literature both methodologically and empirically by proposing a joint econometric approach for examining the count events as well as the severity outcome for different crash types. Methodologically, we build an integrated framework that embeds the fractional split model structure within the recasting framework to develop a joint model with high dimensionality. To be specific, we employ a Joint Panel mixed Negative Binomial- Generalized Ordered Probit Fractional Spilt (PM-NB-GOPFS) model where the first component (NB) will accommodate for crash frequency by crash type and the second component (GOPFS) will study the fraction of severity outcome for different crash types. Empirically, the proposed approach allows for flexible consideration of crashes by crash types and severity levels within a single framework. Further, the proposed model results offer insights on important variables affecting crash frequency and severity for different crash types at a zonal (macro) level.

The rest of the paper is organized as follows: The next section presents the methodological framework adopted in the analysis and presents a simulation exercise to illustrate the strengths of the proposed model. The third section provides a detailed description of the dataset. Model findings are discussed in the fourth section followed by the prediction performance evaluation of the of the proposed model in section five. Finally, the concluding remarks are summarized in the last section.

2 METHODOLOGY

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

2.1 Count Model Structure

The focus of our study is to recast the multivariate NB count model as a panel mixed univariate NB modeling framework. For this purpose, we consider the six types of crashes as repeated measures (same TAZ is repeated 6 times) of crash frequency in a univariate NB formulation while recognizing that each repetition represents a different crash type. The econometric framework of the proposed approach is presented in this section. Let's assume i $(i = 1,2,3, \dots, N;\ N = 3{,}815)$ be an index to represent observation unit (TAZs) and r $(r = 1,2, \dots, R;\ R = 6)$ be an index for different crash type and k $(k = 1,2,3, \dots, K; K = 4)$ be the index to represent injury severity categories at observation unit i. Then the probability equation of the NB formulation can be rewritten as follow:

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

where, c_{ir} be the index for crash counts occurring over a period of time in observation unit i and crash type r. $P(c_{ir})$ is the probability that unit i has c_{ir} number of crashes for crash type r. λ' is NB over dispersion parameter and v_{ir} is the expected number of crashes occurring in i over a given time period for crash type r. In equation 1, we can express v_{ir} as a function of explanatory variables using a log-link function as follows:

$$v_{ir} = exp((\tau + \Phi_{ir} + \varrho_i + \eta_{irk})x_{ir} + \varepsilon_{ir}) \quad (2)$$

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

2.2 Severity Model Structure

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

$$y^*_{ir} = (\alpha_r + \gamma_{irk} + \delta_{ir} + \eta_{irk})z_{ir} + \xi_{irk} \quad (3)$$

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

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

$$\psi_{rk} = fn(s_{irk}) \quad (4)$$

where, s_{irk} is a set of exogenous variables (including a constant) associated with k th threshold. Further, to ensure the accepted ordering of observed crash severity proportion ($-\infty < \psi_{\mathrm{r1}} < \psi_{\mathrm{r2}} < \ldots\ldots\ldots < \psi_{\mathrm{rK-1}} < +\infty$), we employ the following parametric form as employed by Eluru et al.(Eluru et al., 2008):

$$\psi_{rk} = \psi_{r,k-1} + exp((\beta_{rk} + \theta_{irk} + \varsigma_{ir} + \eta_{irk})s_{irk}) \quad (5)$$

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

To estimate the model presented in equation 3, we assume that:

$$E(y_{irk}|Z_{irk}) = H_{irk}(\alpha_r, \psi_{rk}, \delta_{ir}, \theta_{irk}), 0 \le H_{irk} \le 1, \sum_{rk=1}^{rK} H_{irk} = 1 \quad (6)$$

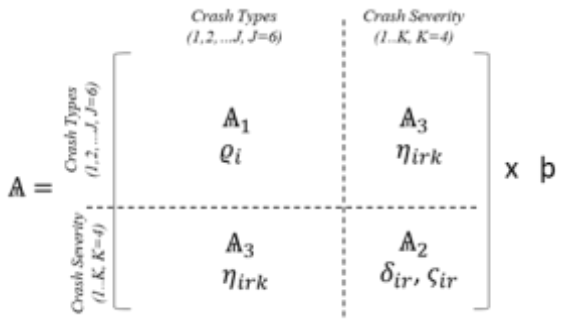
where H_{irk} in our model takes the generalized ordered probit probability form for the severity category k specific to crash type r. Given these relationships across different parameters, the resulting probability for the GOPFS model takes the following form:

$$P_{irk} = G\,[(\psi_{rk} - \{(\alpha + \gamma_{irk} + \delta_{ir} + \eta_{irk})z_{ir}\}] \quad - \quad G\,[(\psi_{r,k-1} - \{(\alpha + \gamma_{irk} + \delta_{ir} + \eta_{irk})z_{ir}\}] \quad (7)$$

where, $\mathrm{G}(\cdot)$ is the standard normal cumulative distribution function (Eluru et al., 2013; Papke, 1996). The proposed model ensures that the proportion for each severity category is between 0 and 1 (including the limits).

2.3 **Correlation Structure**

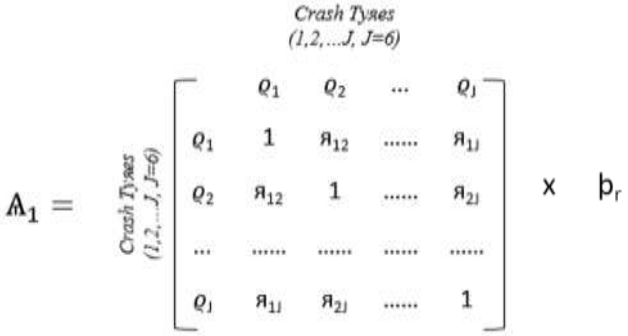
In the current research effort, several unobserved factors are considered. At the observation level (TAZ), we consider influence of common unobserved factors across crash frequency (Φ_{ir}) and crash severity (γ_{irk}, θ_{irk}). In addition to this, a number of correlation terms are tested including: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



(8)

Equation 8 provides the overall structure of the correlation matrix. The order of the correlation matrix is provided by the total number of crash type and crash severity levels (N+K). The reader would note that unobserved factors affecting crash type and severity might have a positive (negative) association i.e. unobserved factors increasing crashes by a type also increase (decrease) severity proportion for that type. Thus, it is not hard to recognize a large possibility of combinations of positive and negative correlations. The matrix Þ is introduced to recognize that the common unobserved factors between crash type, severity and type and severity combination can either be positive or negative. Þ is an appropriately concatenated matrix composed of three matrices $(\text{þ}_r, \text{þ}_k, \text{þ}_{rk})$. The Þ matrix allows us to generate the various possible combinations of positive and negative associations between these correlations.

To elaborate on the structure, we discuss the three main components of the matrix. The top left part represents the correlation matrix for the crash type only:



(9)

As described in Equation 9, these terms represent the correlation between crash types. For example, the correlation parameter $я_{12}$ in equation 9 captures the common unobserved factors affecting the crash counts of crash type 1 and crash type 2 (which is rear-end and angular for the current study context) simultaneously while $я_{2J}$ represents the potential correlation between crash type 2 and crash type J. þ_r matrix will be +1 if the association is positive, -1 if association is negative and 0 if no association is considered.

Equation 10 represents the lower right part of the correlation matrix in equation 8 that accommodates for the common unobserved heterogeneity across the crash severity proportions.

$$\mathbb{A}_2 = \begin{array}{c} \\ \text{Crash Severity (1..K, K=4)} \end{array} \begin{array}{c} \text{Crash Severity (1..K, K=4)} \\ \begin{bmatrix} & \delta_1 & \dots & \delta_K \\ \delta_1 & 1 & \dots\dots & \mu_{1K} \\ \dots & \dots\dots & \dots\dots & \dots\dots \\ \delta_K & \mu_{1K} & \dots\dots & 1 \end{bmatrix} \end{array} \text{x} \quad \text{þ}_k \tag{10}$$

To elaborate, the correlation parameter $\mu_{1K''}$ captures the presence of common unobserved factors between the crash proportion of severity category 1 and K (which is no and severe injury for the current analysis). $\text{þ}_k$follows similar notation as earlier.

Equation 11, representing the bottom left or top right parts in the correlation matrix from equation 8 captures the potential correlation between a crash type and its' corresponding severity proportion.

$$\mathbb{A}_3 = \begin{array}{c} \\ \text{Crash Types and severities} \end{array} \begin{array}{c} \text{Crash Types and severities} \\ \begin{bmatrix} & \varrho_1 & \varrho_2 & \dots & \varrho_J \\ \delta_1 & \acute{r}_{11} & \acute{r}_{21} & \dots\dots & \acute{r}_{1J} \\ \dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ \delta_K & \acute{r}_{1K} & \acute{r}_{2K} & \dots\dots & \acute{r}_{JK} \end{bmatrix} \end{array} \text{x} \quad \text{þ}_{rk} \tag{11}$$

Specifically, the correlation parameter $\acute{r}_{11}$ captures the presence of potential correlation between the crash counts of crash type 1 and crash proportion of severity category 1. It is important to note that the correlation structure presented is applicable to each independent variable examined in the model (including constants). þ_{rk} follows similar notation as earlier. This indicates that potentially (N+K)* (N+K)/2 elements can be estimated for each variable. While theoretically this is possible, it is important to conduct the estimation judiciously to avoid identification issues. For ease of following þ matrix, an example realization is provided in the Appendix A. As is apparent, it is possible to test a large number of combinations of the þ matrix. However, as opposed to running all possibilities, the estimation is judiciously conducted for expected relationships. The model structure that offers the superior data fit is considered as the final model.

2.4 **Joint (NB-GOPFS) Model Estimation**

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

$$L_i = \int_{\Omega} \prod_{r=1}^{R} \left[\left(P(c_{ir}) \right) \times \prod_{k=1}^{K} (P_{irk})^{\varpi_{ir} d_{irk}} \right] d\Omega \quad (12)$$

where, ϖ_{ir} is a dummy with $\varpi_{ir} = 1$ if TAZ i has at least one crash specific to crash type r over the study period and 0 otherwise. d_{irk} is the proportion of crashes in severity category k for each crash types. Further, we apply simulation techniques to approximate the integrals in the likelihood function and maximize the logarithm of the resulting simulated likelihood function across zones with respect to . The simulation technique approximates the likelihood function in Equation (12) 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 c^{th} draw (c = 1, 2, …, C), then the observational likelihood function is approximated as:

$$DL_i = \frac{1}{C} \sum_{c=1}^{C} (DL_i^c) \quad (13)$$

In our research, we tested the model specification with several realization levels (such as 50, 100, ..200). We found that model parameters were stable around 100. For additional stability, we selected the number of draws as 200 (C value). Finally, the log-likelihood function is:

$$LL = \sum_{i} Ln(L_i) \quad (14)$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 8. The parameters to be estimated in the model are: $\Phi, \varrho, \gamma\ \ \delta, \alpha, \tau, \beta, \psi, \pi$, σ and ν. To estimate the proposed model, we apply Quasi-Monte Carlo simulation techniques based on the scrambled Halton sequence to approximate this integral in the likelihood function and maximize the logarithm of the resulting simulated likelihood function across individuals (see (Bhat, 2001; Eluru et al., 2008) for examples of Quasi-Monte Carlo approaches in literature).The model estimation routine is coded in GAUSS Matrix Programming software (Aptech).

2.5 Simulation Study

We conduct a simulation study to illustrate how the proposed PMNB-GOPFS can be employed for 1) replicating the model results based on an assumed data generation process (DGP) and 2) evaluate the performance of model relative to the traditional random parameters multivariate negative binomial (RPMNB) model.

2.5.1 Parameter Retrieval of PMNB-GOPFS

The simulation and estimation of the proposed exercise is conducted considering 2 crash types and 3 crash severities. Therefore, our proposed joint system will result in 2 count and 3 ordered dependent variables. Among each of these components, we consider three independent variables

drawn from a univariate standard normal random distribution. The simulation exercise was conducted for different sample sizes (including 1,000; 2,000; 3,000 and 5,000 records). Across each sample size, 50 samples each were generated. The results across these various sample sizes are consistent. Hence, to conserve on space, we present the results of the 5,000 observation sample. The performance evaluation is conducted based on the parameter retrieval capability along two dimensions: (1) parameter bias and (2) asymptotic standard error. These two measures examine if the parameter values are recovered while ensuring the parameters are statistically significant. The results of our simulation exercise are presented in Table 2. The columns of the Table include True parameter, Mean Estimate, Absolute Percentage Bias and Asymptotic Standard Error. The mean estimate is obtained as the mean parameter value from the 50 samples. The absolute percentage bias is computed as $\left|\left[\frac{True\ Parameter - Mean\ Estimate}{True\ Parameter}\right] * 100\right|$. The asymptotic standard error is computed as the mean of the parameter standard error across the samples. The values presented in the Table 2 clearly illustrate that the proposed model system retrieves the parameters with small standard errors.

2.5.2 Comparison with the RPMNB model

The simulation exercise is further augmented by evaluating the PMNB-GOPFS model performance relative to the RPMNB model system. The reader would note that the comparison between these model systems results in two challenges. First, the two model systems arise out of different data generation processes (DGPs). Hence, any comparison ought to consider the underlying DGP. Second, the two model systems do not have the same econometric and/or likelihood structure and cannot be compared in the estimation space. The model performance comparison has to be considered with prediction measures.

In our simulation, we consider simulation and estimation using the two DGPs. First, we simulate data using the PMNB-GOPFS DGP following the same process described in Section 2.5.1. Using the data generated, both models are estimated. Subsequently, we generate data using the RPMNB DGP (i.e. assume the crashes occur following the count frequency process) and estimate the two models. In our comparison, we compute the root mean square error (RMSE) value at disaggregate level (see (Bhowmik et al., 2019a, 2018) for detail) and compare their performances for all the crash types and severities. The box plots in Figure 3 represents the RMSE values specific to each severity group for each crash type predicted from the two models under the two DGPs. The Figure clearly highlights the overall superior predictive performance offered by the proposed PMNB-GOPFS over the traditional RPMNB model as indicated by the lower/similar RMSE values, irrespective of the data generation process. The findings further reinforce the applicability of our proposed framework for crash safety literature.

Table 2: Evaluation of the PMNB-GOPFS Model Ability to Recover True Parameter

		Crash Type 1				**Crash Type 2**			
	Coefficient	*TRUE Parameter*	*Mean Parameter Estimates*	*Absolute Percentage Bias (%)*	*Asymptotic Standard Error*	*TRUE Parameter*	*Mean Parameter Estimates*	*Absolute Percentage Bias (%)*	*Asymptotic Standard Error*
PMNB (Crash Count)	τ_0	1.000	0.967	3.287	0.004	-0.750	-0.783	4.345	0.014
	τ_1	0.800	0.816	2.025	0.002	1.800	1.893	5.167	0.012
	τ_2	-0.350	-0.357	1.947	0.002	-0.350	-0.376	7.571	0.004
	τ_3	1.500	1.520	1.307	0.003	0.500	0.541	8.202	0.005
	λ	0.450	0.440	2.239	0.010	1.500	1.488	0.788	0.027
GOPFS (Crash Severity Proportions)	ψ_1	1.000	0.998	0.187	0.001	0.250	0.234	6.245	0.002
	ψ_2	2.750	2.708	1.527	0.001	2.500	2.394	4.229	0.004
	α_1	-1.500	-1.456	2.954	0.001	-1.500	-1.507	0.458	0.003
	α_2	-1.750	-1.712	2.185	0.001	0.250	0.252	0.861	0.001
	α_3	-0.150	-0.144	4.188	0.001	-2.150	-2.099	2.363	0.002

Note: τ_0, τ_1, τ_2 and τ_3 represent the vector of coefficients for independent variables and λ – over-dispersion parameter; $\alpha_1, \alpha_2, \alpha_3$represent the coefficients for independent variables and ψ_1, ψ_2-are thresholds in the GOPFS model

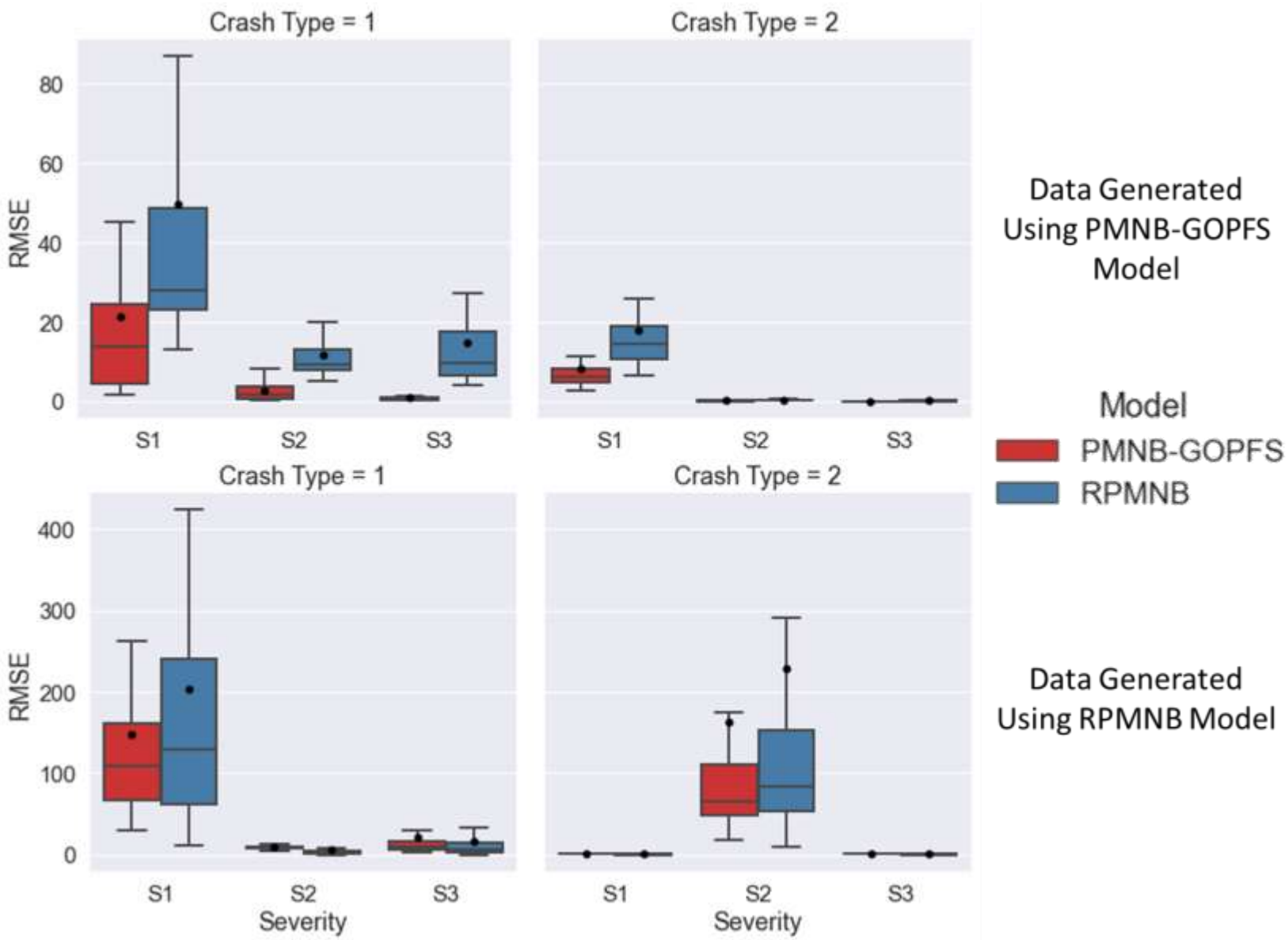


Figure 3 Performance Comparison of PMNB-GOPFS and RPMNB Model for Simulated data

3 DATA PREPARATION

Our study area includes the Central Florida Region encompassing a total of 11 counties in the state of Florida with 4,747 zones. The crash records compiling the information of crash types and the corresponding severity outcomes were acquired from Florida Department of Transportation (FDOT), Crash Analysis Reporting System (CARS) and Signal Four Analytics (S4A) databases. The analysis is conducted using the 2016 crash records considering six different types of crashes. At first, the 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. Then for each crash types, crashes are further classified by injury severity levels such as fatal (K), incapacitating (A), non-incapacitating (B), possible injury (C), and property damage only (O) crashes. Based on crash records, fatal and incapacitating injuries are combined as one category and defined as severe injury. Finally, the crash records are aggregated at a zonal level and the corresponding severity proportions by crash type are as follows: (1) proportion of no injury (property damage only) crashes, (2) proportion of minor injury crashes, (3) proportion of non-incapacitating injury crashes, and (4) proportion of severe injury crashes.

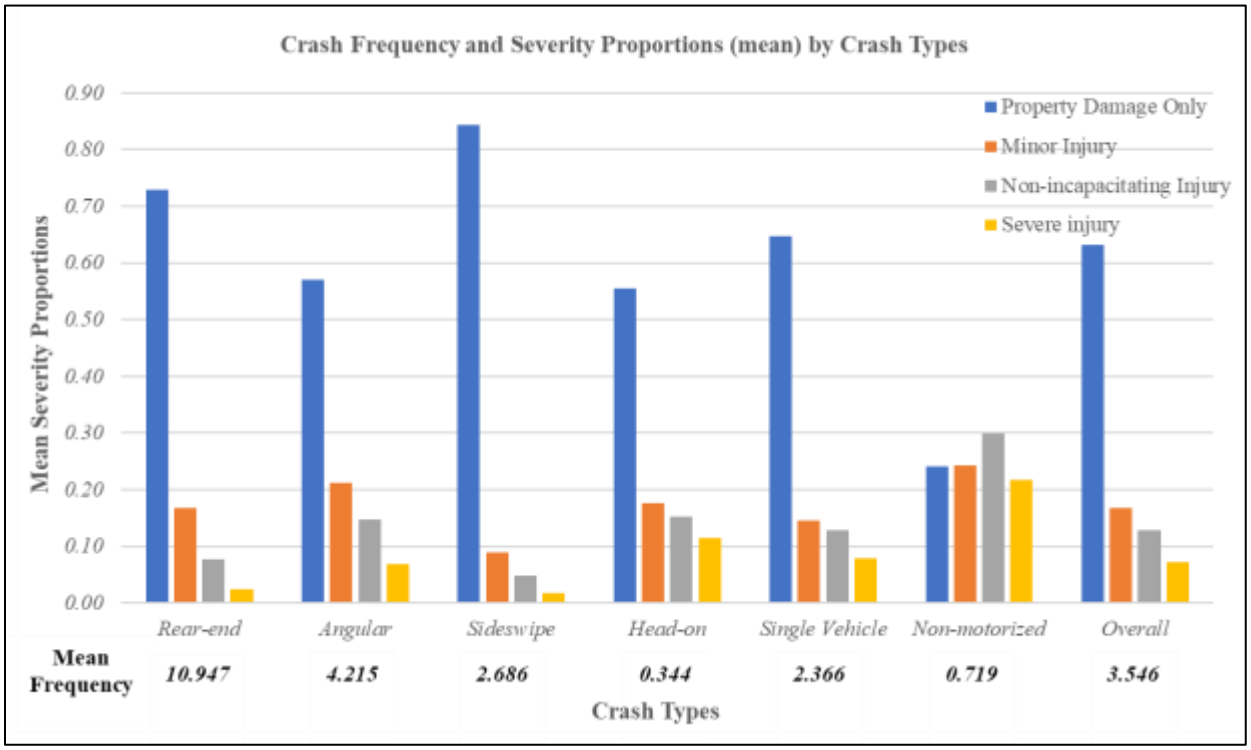


Figure 1 Crash Frequency and Severity Proportions (mean) by Crash Types

A total of 1,14,458 motorized (ranging from 0 to 243) and 3,413 non-motorized crashes (ranging from 0 to 12) were reported in the Central Florida 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. The crash counts and severity outcome proportions for each crash type are presented in Figure 1. From the Figure 1, we can

[2] For our analysis, non-motorized crashes refer to the crashes where at least one non-motorist was involved

observe that number of no injury crashes has the highest proportion followed by proportion of minor injury crashes. Further, in terms of crash types, the Figure 1 shows that non-motorists are more prone to severe crashes whereas the injury outcomes are higher for motorists involved in head-on crashes. On the other hand, in approximately 84% and 72% sideswipe and rear-end crashes, respectively, the outcomes were no injury. The most commonly used approach of modeling severity frequency or proportion without considering crash type would result in an inaccurate aggregation. From the Figure (1), it is evident that severity proportions by crash type vary significantly across crash types.

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. With respect to 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 center, school, hospital, recreational center, restaurant and shopping center 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 zonal level sociodemographic attributes included 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 3 summarizes sample characteristics of the explanatory variables with the appropriate definition considered for final model estimation along with the minimum, maximum and mean values at a zonal level. 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.

TABLE 3 Summary Statistics of Exogenous Variables (Zonal Level)

Variables	Definition	Zonal (N=4,747)			
		Minimum	Maximum	Mean	Std. Deviation
Roadway Characteristic					
Proportion of rural road	(Rural road length/total road length)	0.000	1.000	0.121	0.309
Proportion of urban road	(Urban road length/total road length)	0.000	1.000	0.806	0.381
Proportion of arterial road	(Arterial road length/total road length)	0.000	1.000	00377	0.393
Proportion of local road	(Local road length/total road length)	0.000	1.000	0.053	0.170
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
Proportion of road with separated median	Length of road with separated median/total road length	0.000	1.000	0.510	0.459
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
Road ≥55mph	Proportion of road length greater than 55mph	0.000	1.000	0.088	0.174
Proportion of poor pavement road	Road length with poor pavement condition/ total road length	0.000	1.000	0.035	0.144
Land-use Attributes					
Urban area	Ln (urban area+1) in acre	0.000	9.440	4.921	1.970
Recreational area	Ln (recreational area+1) in acre	0.000	9.814	0.470	1.408
Office area	Ln (office area+1) in acre	0.000	6.440	0.877	1.383
Residential area	Ln (residential area+1) in acre	0.000	8.131	3.811	2.075
Industrial area	Ln (industrial area+1) in acre	0.000	7.067	1.118	1.306
Institutional area	Ln (institutional area+1) in acre	0.000	6.617	1.946	1.589

Land use mix	Land use mix = $\left[\frac{-\sum_k (p_k (lnp_k))}{lnN}\right]$, where k is the category of land-use, p is the proportion of the developed land area for specific land-use, N is the number of land-use categories	0.000	0.946	0.369	0.221
Built Environment Characteristics					
No of business center	Z score: No of business center	-0.138	19.664	0.000	1.000
No of commercial center	Z score: No of commercial center	-0.270	9.521	0.000	1.000
No of educational center	Z score: No of educational center	-0.487	11.610	0.000	1.000
No of recreational center	Z score: No of park and recreational center	-0.475	16.678	0.000	1.000
No of restaurant	Z score: No of restaurant	-0.464	11.021	0.000	1.000
No of shopping center	Z score: No of shopping center	-0.442	19.728	0.000	1.000
Traffic Characteristics					
VMT	Vehicle miles travelled	0.000	15.026	7.914	3.368
Congested traffic	AADT> 85th percentile of AADT	0.000	1.000	0.130	0.336
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
Employee	Total number of commuter/1,000	0.000	8.265	0.401	0.584
Non-motorist commuter	Ln (Non motorized means to work for a TAZ)	0.000	5.261	1.278	1.098
Transit commuter	Ln (transit means to work for a TAZ)	0.000	6.369	0.756	1.114
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 EMPIRICAL ANALYSIS

4.1 Model Specification and Overall Measure of Fit

The number of TAZs in the study area is 4,747. Among these zones, 3,815 TAZs are randomly selected for model estimation and the records from other 932 TAZs are set aside for validation purposes. Thus, the estimation sample has 22,890 (3,815*6) records and the validation sample has 5,592 (932*6) data records. The empirical analysis involved a series of model estimations. First, we estimated separate independent models (NB and GOPFS models) to establish a benchmark for comparison. Second, we proposed a parsimonious model structure using the same independent model system (NB and GOPFS) while restricting the parameters across different crash types considered. To elaborate, observing the model specifications in the independent models (NB and GOPFS), we identify potential parameters that can be restricted to be the same across various crash types and test that restriction (both NB and GOPFS dimension) in our proposed model system (see (Bhowmik et al., 2019a) for more details). Third, within our proposed system, we consider the unobserved heterogeneity in the joint model estimation. In summary, we estimated three different models in the current research effort including: 1) Independent NB-GOPFS model; 2) Panel NB-GOPFS model without unobserved component parameters and 3) Joint Panel NB-GOPFS model with unobserved heterogeneity. The log-likelihood values at convergence for these estimated models are: a) Independent NB-GOPFS (with 131 parameters) is -51,904.45 (b) Panel NB-GOPFS model without unobserved component (with 100 parameters) is -51,912.92.11 and (c) Joint Panel NB-GOPFS model with unobserved heterogeneity (with 105 parameters) is -50,945.82. We also compute the Bayesian Information Criterion (BIC) (lower is better) for these three frameworks to determine the best model. The corresponding BIC values for the three models are as follows: 105,123.93 (independent NB-GOPFS model), 104,650.50 (panel NB-GOPFS model) and 102,757.53 (joint panel NB-GOPFS model). Based on the BIC values, two observations can be made. First, the proposed framework that accounts for penalty for additional parameters provide improved data fit compared to the traditional model (independent NB-GOPFS model). This supports our hypothesis that the impact of some variables may not differ across the crash types and through the proposed structure (recasting), we can have a parsimonious model system with improved parameter efficiency. Second, models considering unobserved heterogeneity outperforms the respective independent models which underscores the importance of accommodating for such unobserved effects in examining crash frequencies and severities at the planning level for different crash types.

4.2 Model Estimation Results

This section offers a detailed discussion of exogenous variable effects on the crash count as well as the severity outcome for different crash types. In discussing the model results, for the sake of brevity, we will restrict ourselves to the discussion of the joint panel model (NB-GOPFS) only (see Appendix A for the results of independent NB-GOPFS model). For the ease of presentation, we first present an intuitive discussion of crash count component (Table 4) followed by the discussion of the severity component (Table 5) for different crash types.

4.2.1 Count Component

The coefficients in Table 4 represent the effect of exogenous variables on the frequency component of each crash type. The reader would note that, the variables in the crash count component of Table 4 with positive (negative) sign indicates that an increase in the variable is

TABLE 4 Joint Panel Mixed NB-GOPFS Model Results (Count Component)

Variables (np)	Rear-End		Angular		Sideswipe		Head-on		Single vehicle		Non-motorized	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Constant (6)	-0.626	-7.73	-1.684	-15.407	-2.687	-21.582	-3.557	-18.081	-0.744	-9.932	-2.580	-23.487
Roadway Characteristics												
Proportion of arterial roads (2)	0.166	4.034	0.166	4.034	--[1]	--	--	--	-0.284	-5.105	0.166	4.034
Number of intersections (1)	--	--	0.347	11.804	--	--	0.347	11.804	--	--	0.347	11.804
Signal intensity (3)	0.416	2.422	--	--	-0.630	-3.277	--	--	-0.447	-1.746	0.416	2.422
Road length over 55mph (5)	0.468	3.679	-1.573	-7.679	0.468	3.679	-1.022	-2.877	0.892	7.676	-1.172	-4.591
Standard deviation	--	--	0.903	3.288	--	--	--	--	--	--	--	--
Variance of Speed (2)	0.040	3.697	0.040	3.697	0.069	4.451	--	--	--	--	--	--
Roads with separated median (2)	0.172	3.798	0.172	3.798	0.172	3.798	-0.156	-1.411	--	--	--	--
Average outside shoulder width (4)	-0.308	-7.120	-0.439	-8.323	-0.563	-9.800	-0.308	-7.120	-0.115	-2.621	--	--
Average sidewalk width (1)	--	--	--	--	--	--	--	--	--	--	-0.215	-3.693
Traffic Characteristic												
VMT (4)	--	--	0.131	8.496	0.259	16.909	0.185	8.852	--	--	0.021	1.678
Truck VMT (2)	0.179	15.852	--	--	--	--	--	--	0.270	26.819	--	--
Land-use attributes												
Urban area (4)	0.164	15.359	0.164	15.359	0.149	9.530	0.111	4.094	--	--	0.114	6.279
Office area (2)	0.148	10.384	--	--	0.148	10.384			--	--	0.127	7.389
Residential area (1)	--	--	--	--	-0.077	-6.915	-0.077	-6.915	--	--	--	--
Built environment characteristic												
No. of restaurants (3)	0.273	10.432	--	--	0.084	3.394	--	--	--	--	0.175	7.958

No. of shopping centers (1)	0.030	1.712	--	--	0.030	1.712	--	--	--	--	--	--
Socio-demographic characteristics												
Non-motorists (3)	0.052	2.892	0.148	7.166	0.168	7.581	--	--	--	--	0.052	2.892
Transit users (1)	0.222	13.287	--	--	--	--	--	--	--	--	0.222	13.287
Over dispersion (6)	0.671	16.130	0.251	6.515	0.284	8.270	1.002	6.245	0.713	19.270	0.235	4.459

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

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

likely to result in more (less) crashes. In the subsequent sections, we provide a discussion of model results for different crash types by variable groups. The reader would note that Table 4 identifies the number of parameters estimated for each variable from a possible set of six (one effect for each crash type).

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

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

vehicular maneuvers and thus reduce the potential of all kinds of motorized crashes. With respect to sidewalk width, a number of earlier research concluded that increased sidewalk width is associated with higher pedestrian activity and as a result, they are more exposed to crashes. In our current study, we found an opposing (negative) effect of average sidewalk width for non-motorized crashes. However, there is a reasonable explanation for the effect identified. First, the reader would note that we consider the non-motorist activity separately in the model framework (will be discussed in the following sections) and second, increased sidewalk width will provide additional safety to the non-motorist from colliding with a motorized vehicle.

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

Land-use Attributes: With respect to land-use attributes, several factors exert significant impact on crash count components across crash types. The coefficient corresponding to urban area indicates that zones with higher urbanized area are likely to have increased crash risk for five of the six crash types (except single vehicle crashes). Similarly, office area in a zone is also found to be positively associated with rear-end, sideswipe and non-motorized crashes. These two variables basically reflect presence of higher vehicular and non-motorist interactions and in turn, higher exposure for both road user groups. Further, the result in Table 4 reveals a reduced propensity for sideswipe and single vehicle crashes with higher residential area.

Built Environment Attributes: In terms of built environment attributes, several variables have been explored out of which only number of restaurants and shopping centers are found to be related with zonal level crash risks. As is evident from Table 4, we can observe that both number of restaurants and shopping centers have positive influence on rear-end and sideswipe crashes, perhaps indicating a higher density of traffic volume for these areas. With respect to non-motorized crashes, number of restaurants is found to be a significant determinant with a positive impact (see (Yasmin et al., 2021) for similar result).

s

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

4.2.2 Severity Component

The coefficients in Table 5 represent the effect of exogenous variables on the injury severity proportion across different crash types. In the propensity, a positive (negative) coefficient corresponds to increased (decreased) proportion for severe injury categories specific to each crash type. When the threshold parameter is positive (negative), the result implies that the threshold is

TABLE 5 Panel Mixed NB-GOPFS Model Results (Severity Component)

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

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

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

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

bound to increase (decrease). The estimation results are discussed by variable groups in the following sections. The reader would note that Table 5 identifies the number of parameters estimated for each variable from a possible set of six (one effect for each crash type).

Roadway Characteristics: The variable specific to arterial road indicates that the likelihood of more severe crashes (proportions) increases with increasing share (length) of arterial road in a zone, particularly for rear-end, angular and single vehicle crashes. Further, we found an effect of arterial road on threshold value for rear-end crashes which provide a sense of how the probability of injury in specific injury categories is affected relative to the case of fixed thresholds. The negative coefficient of the variable on the threshold value highlights the higher proportions of serious injury (non-incapacitating or severe) crashes for rear-end crashes with increased length of arterial roads. Moreover, it can be seen from Table 5 that crashes on local road tends to be less severe for head-on and non-motorized crash types. The reduced likelihood of severe crashes for these two crash types perhaps can be attributed to reduced driving speed on local roads.

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

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

Traffic Characteristics: Traffic congestion and truck VMT are found to have significant impact on crash proportions by severity levels for different crash types. As is evident from Table 5, we can observe that roads are typically safer in a congested traffic environment. In particular, the likelihood of severe crash proportion for rear-end and angular crashes are lower in a congested traffic environment (>85^{th} percentile traffic) compared to the uncongested condition (<=85^{th} percentile traffic).

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

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

Built Environment Attributes: In terms of built environment attributes, several factors are considered including number of commercial, recreation, restaurants and shopping centers. Interestingly, all of these reveal negative associations with the crash severity proportions across different crash types, perhaps indicating that with higher traffic density vehicle operating speed is likely to be lower and thus crash consequences are possibly less severe. For instance, consistent with previous findings (Yasmin et al., 2021), number of commercial centers reduce the higher injury risk propensity for non-motorized crashes. Similarly, in the presence of higher number of recreational centers in a zone, a lower proportion of severe crash outcomes for single vehicle crashes is observed. Further, the GOPFS model results reveals that higher number of restaurants are associated with lower likelihood of severe crash proportions for single vehicle crashes, as indicated by the negative coefficient. The positive coefficient of the variable on the threshold value further reflects the lower probability of severe crash proportions. Finally, the variable corresponding to shopping centers results in lower likelihood of severity outcome, particularly for angular, sideswipe and head-on crashes (same impact). We also found a positive effect of the variable on the threshold which further implies the lower possibility of higher injury risk for sideswipe crashes with increased number of shopping centers in a zone.

Socio-demographic Characteristics: The results for the effect of socio-demographic characteristics indicate that non-motorists are less prone to high injury risk with increased number of commuters in a zone (see (Yasmin et al., 2021) for similar results). The likelihood of being involved in a severe crash is higher for increasing share of motor vehicle commuters, particularly for angular crashes. Previous studies (Pai and Saleh, 2008) also confirm the findings. Further, as found in previous studies (Quddus, 2008), the estimated results suggests that zones with more older people are associated with fewer severe crash proportion for non-motorized crashes. The coefficient specific to proportion of households without vehicle indicates a positive influence on severity outcome for non-motorized crashes indicating a higher propensity of more severe crash proportion for non-motorized crashes (for similar results, see (Quddus, 2008)).

4.2.3 Unobserved Effects

The final set of variables in both Table 6 correspond to the correlation matrix (unobserved heterogeneity) in the joint model. As discussed earlier, in the current research effort, a number of correlation effects are tested including: 1) common unobserved factors affecting crash counts of different crash types simultaneously; 2) common unobserved factors affecting crash severity proportions of different crash types simultaneously and 3) common unobserved factors that simultaneously impact crash counts and severity proportions by different crash types. The reader would note that in our empirical setting with 6 crash types and 4 crash severities, we can test for the presence of $24C_2$ potential correlation terms. However, testing for such high number of correlation results in estimation and computational complexity. In our approach, we have made every effort to reduce the dimensionality of the model parameters to ensure parsimonious specification. This has resulted in restricting correlations across multiple dependent variables

TABLE 6 Panel Mixed NB-GOPFS Model Results (Unobserved Correlation)

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
Within Crash Counts of Different Crash types												
Correlation 1 (1)** ϱ_{ir1}, matrix A_1	0.585	21.818	--	--	--	--	--	--	--	--	0.585	21.818
Correlation 2 (1) ϱ_{ir2}, matrix A_1	--	--	0.957	43.658	0.957	43.658	0.957	43.658	--	--	--	--
Between Crash Counts and Severity Proportions of Different Crash Types												
Correlation 3 (1) *** η_{irk1}, matrix A_3	--	--	--	--	--	--	--	--	--	--	0.072	2.831
Correlation 4 (1) η_{irk2}, matrix A_3	--	--	--	--	--	--	0.479	4.516	--	--	--	--

**Correlation 1 refers to common unobserved factors affecting rear-end and non-motorized crashes and
Correlation 2 refers to common unobserved factors affecting other multi vehicular crashes (angular, sideswipe and head-on) simultaneously

***Correlation 3 refers to common unobserved factors affecting non-motorized crash counts and its corresponding threshold between proportion of no and possible injuries simultaneously

Correlation 4 refers to common correlation affecting head-on crash counts and its corresponding threshold between non-incapacitating and severe crash proportions simultaneously

resulting in the appearance of a small number of correlations. Further, we allow for observed variable interaction in modeling severity proportions through the fractional split framework. The inherent ordering considered in severity module contributes to reducing correlation across crash types as a subset of the dependent variables are explicitly interconnected in their mathematical structure. Further, we consider unobserved correlations across dependent variables as needed. Given the direct interaction across the dependent variables, we are likely to have fewer parameters to test because of the model structure. We believe this is an inherent advantage of our proposed system.

Within crash type itself, the potential number of correlations that can exist is $6C_2$ (15). However, these 15 correlations can materialize in many ways. For example, all 6 dimensions might have the same correlation (unobserved effect remain the same for all dimensions). Similarly, it is possible for 3 dimensions to have a common effect. The exact parameters estimated are arrived at based on exhaustive model estimation. After exhaustive testing, we find two correlations are significant including (1) correlation 1: common unobserved factors affecting rear-end and non-motorized crashes and (2) correlation 2: common unobserved factors affecting other multi vehicular crashes (angular, sideswipe and head-on). Further as indicated earlier, we test for the correlations with both positive and negative sign and based on the result, we find that positive sign offers improved data fit for both correlations (1 and 2). Following this sign, the correlation 1 implies that zones with higher number of rear end crashes are more likely to have higher number of non-motorized crashes. Similar interpretation can also be stated for correlation 2. It should be noted that the reason we allow for same unobserved effect (correlation 2) across the three dependent variables is to reduce the number of separate unobserved terms we estimate. Additionally, we restrict them to be the same to improve model estimation efficiency without any loss in data fit. The estimated two unobserved effects allow for correlation across at least 5 of the 6 crash type count dimensions."

On the other hand, with respect to common factors between two components (count and proportions), we found two correlation terms significant including: (1) correlation 3: common unobserved factors affecting non-motorized crash counts and its corresponding threshold between proportion of no and possible injuries; and (2) correlation 4: common correlation affecting head-on crash counts and its corresponding threshold between non-incapacitating and severe crash proportions. Again, the correlation could be either positive or negative and the sign can change for every unobserved factor estimated. In our analysis, we found the negative sign offers better fit for common correlation between total crash counts and threshold between proportion of no and possible injuries for non-motorized crashes. This indicates that a zone with higher number of non-motorized crashes are more likely to incur lower proportions of no injury crashes. On the other hand, a positive common correlation is found between the total number of head-on crashes and the corresponding threshold between proportion of non-incapacitating and severe crashes which implies that zones with higher number of head-on crashes intrinsically are more likely to incur higher proportions for serious crashes. Overall, the results clearly support our hypothesis that common unobserved factors influence the two components (crash counts and severity proportion).

5 PREDICTIVE PERFORMANCE EVALUATION

5.1 Comparison Exercise

In an effort to illustrate the applicability of our proposed system, we carried out a comparison exercise between our proposed joint PMNB-GOPFS and the existing traditional multivariate

system for predicting crash counts across different crash severities. Further, we realize that sample size in estimation could play a critical role in model performances and hence we considered the influence of different sample sizes in model estimation by estimating the two model systems for different sample sizes. To be specific, from the in-sample dataset (3,815 TAZs), we draw 1,000; 2,000 and 3,000 TAZs randomly and estimate both models for all of these estimation samples and compare their performances based on the final specifications from each estimation sample. Further, for feasibility of estimating the traditional model system, we consider three crash types[3] (rear-end, angular and non-motorized) and for each crash type, we develop a random parameters multivariate negative binomial (RPMNB) model for analyzing the crash counts of different severity levels. In all these samples, the reader would note that the proposed PMNB-GOPFS model has substantially fewer number of parameters compared to traditional RPMNB model. The comparison of parameters between the PMNB-GOPFS and RPMNB model are as follows: estimation sample 1,000 - 54 vs 116; estimation sample 2,000- 56 vs 128; estimation sample 3,000- 56 vs 141; and estimation sample 3,815- 61 vs 148. Clearly, the proposed model structure presents a parsimonious model specification irrespective of the sample size.

Using the final specification of both systems (PMNB-GOPFS and RPMNB), we compute the root mean square error (RMSE) value at disaggregate level (see (Bhowmik et al., 2019a, 2018) for detail) and compare their performances for all the crash severities as well as an overall summary (combining the crash severities) across the three crash types. For comparison purposes, 50 data samples with 500 records (TAZs) each are randomly generated from the holdout sample consisting of 932 TAZs. For these samples, we compute the differences in RMSE between PMNB-GOPFS and RPMNB models and based on the differences, we create 5 categories as follows: 1. Strongly outperform (if PMNB-GOPFS has lower RMSE value with a difference > 5); 2: Outperform (if PMNB-GOPFS has lower RMSE value with a difference >1.5 and ≤ 5); 3. No difference (absolute differences of RMSE between the two system is ≤1.5); 4. Underperform (if PMNB-GOPFS has higher RMSE value form with a difference >1.5 and ≤ 5); and 5. Strongly underperform (if PMNB-GOPFS has higher RMSE value with a difference > 5). Figure 2a and 2b presents the performance (RMSE) of the PMNB-GOPFS model relative to the RPMNB model for different estimation samples.

The reader would note that for the ease of presentation, we considered 5 crash levels including 4 severities and one total crash (overall) across each crash types for the comparison exercise. The overall crash level is simply the summation of all crash severities specific to crash types and is considered to evaluate the crash type (count) errors. In summary, there is a total 3,000 measures computed (5 crash levels *3 crash types*50 validation samples*4 sample sizes) out of which RPMNB model did not provide any improved performance across any measures as indicated by the 0% in the U and SU categories. In fact, out of 3,000 RMSE measures, 19% of the cases PMNB-GOPFS model exhibits superior performance (with at least an order of magnitude difference) while providing equivalent prediction for the remaining 81% measures relative to the traditional RPMNB model. With respect to crash type errors (overall), the performance of the two systems are quite similar over the different estimation samples for angular and non-motorized crashes whereas for rear-end crashes, our proposed model provided superior performance relative to the traditional RPMNB model (out of 200 measures - SO:182 O:17 ND:1; U:0; SU:0-) irrespective of estimation sample size. On the other hand, in terms of injury severities, our proposed system strongly outperforms the RPMNB model across no-injury and minor injury

[3] We can consider all six crash types. However, running RPMNB model with unobserved effects for all crash types will be time consuming. So, to reduce the computational burden. we selected three crash types.

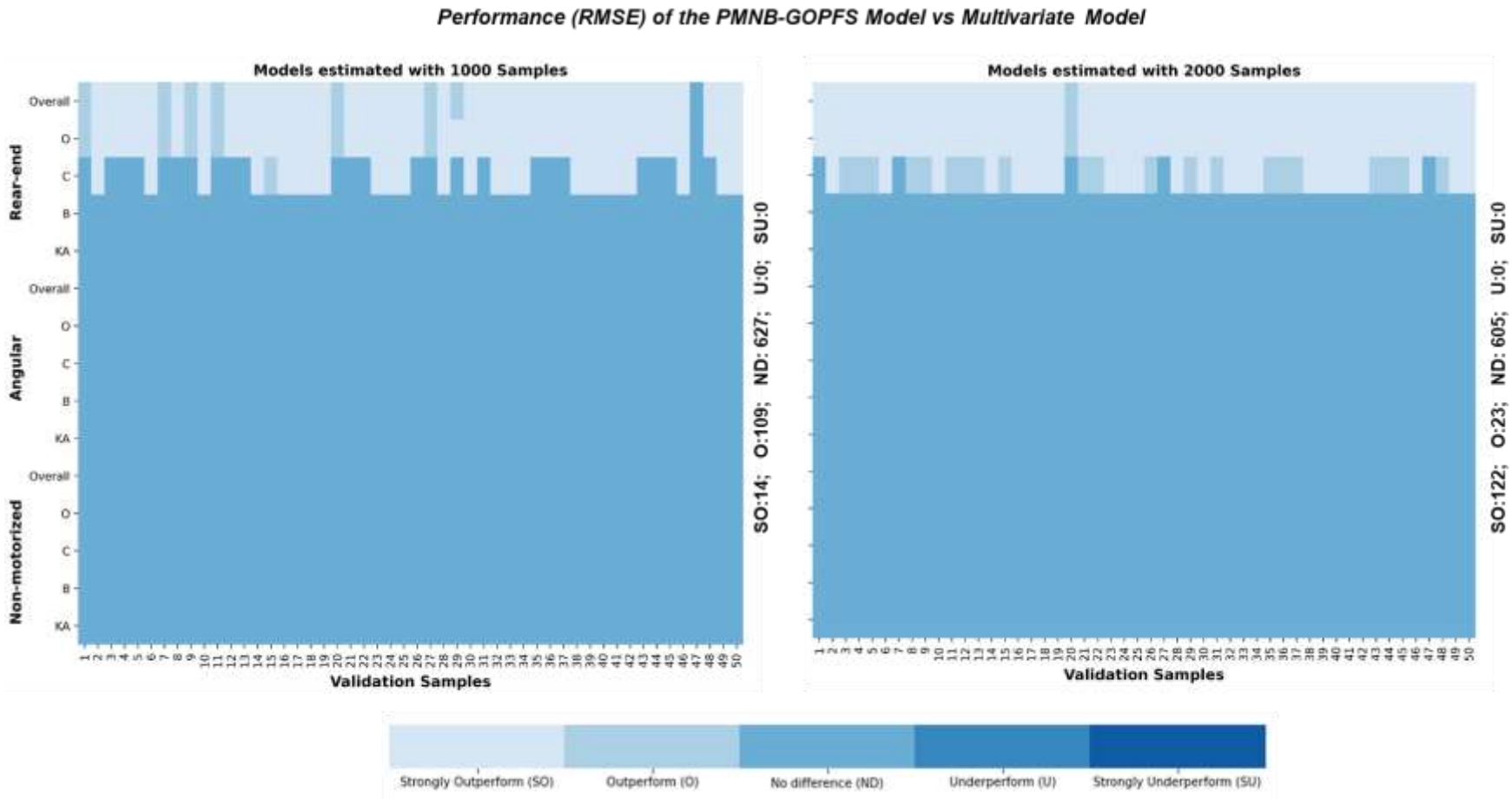


Figure 2a Performance Comparison between PMNB-GOPFS and RPMNB Model

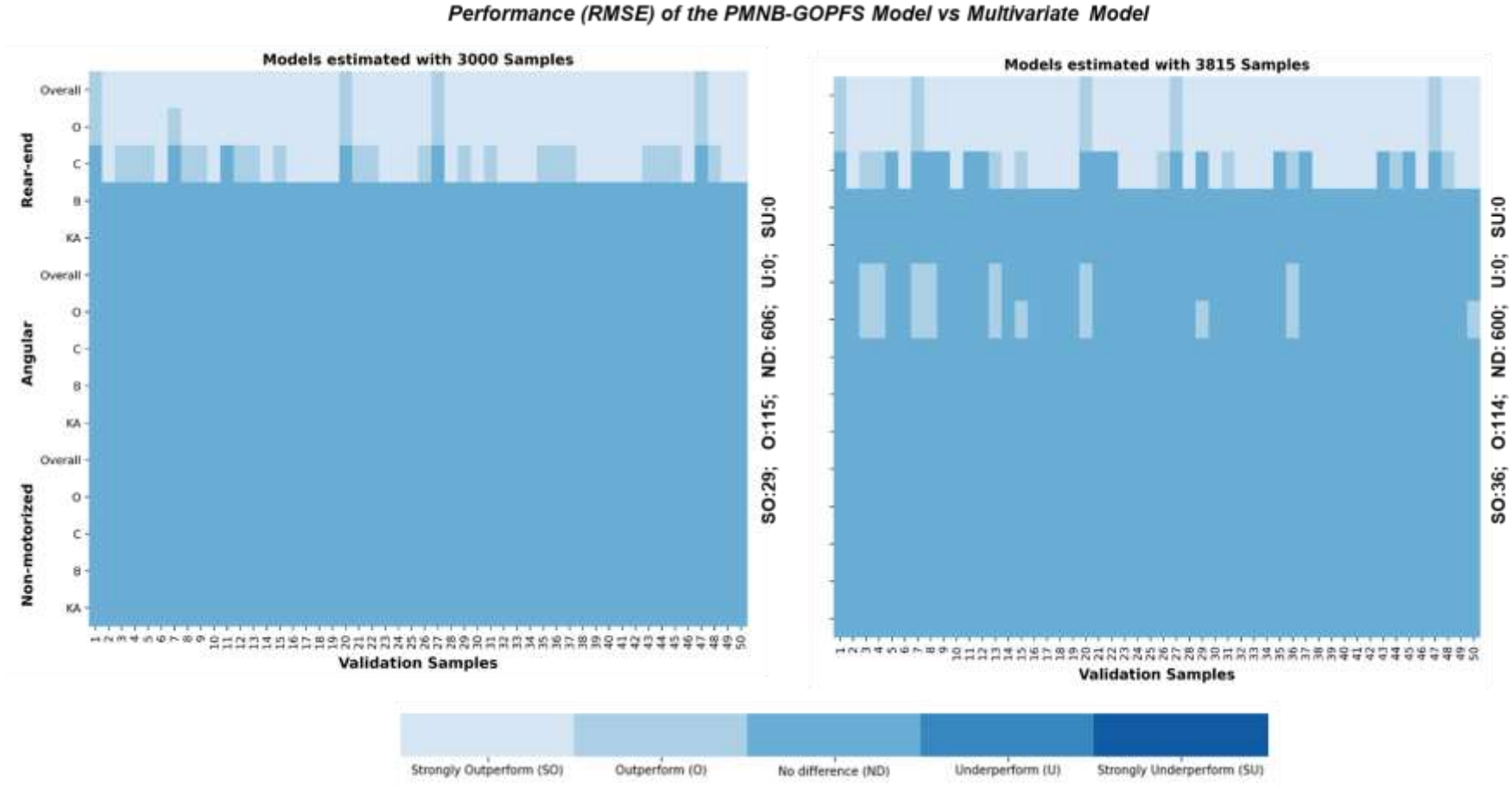


Figure 2b Performance Comparison between PMNB-GOPFS and RPMNB Model

categories for rear-end crashes. However, for angular and non-motorized crashes, we find equivalent performance from the two system regardless of the sample size. The results clearly highlight that with fewer (at most half) number of parameters, our proposed joint system either provides improved or at worst, similar predictive performance compared to the RPMNB model.

5.2 Validation Analysis

A validation exercise was also undertaken using the final model parameter estimates to of the proposed joint (PMNB-GOPFS) model to ensure that the statistical results obtained above are not a manifestation of over fitting to data. In doing so, we employ mean absolute deviation (MAD) and mean absolute percentage error (MAPE) which quantifies the error associated with model prediction and the measure is computed on two datasets including: 1) model estimation sample with 3,815 TAZs and 2) hold out sample (validation sample) with 932 TAZs. One of the major advantages of the proposed framework is that in a single econometric framework, we can predict a number of dimensions including total crash counts, total crash counts by crash types, crash proportions for each severity level, crash counts for each severity level and finally, proportions and counts of crashes for each crash type by severity. In evaluating the predictive performance, we compute the errors (MAD and MAPE) across all the aforementioned dimensions. The prediction results clearly indicate that the joint model for crash counts and severity proportions by crash type performs adequately for both datasets (in-sample and validation sample) under consideration. A summary of the validation procedures and the results are presented in appendix B.

6 CONCLUSIONS

Despite the distinct injury severity profile, there is limited adoption of research modeling severity frequency or proportion considering different crash types. The main challenge is with the number of dependent variables as accommodating unobserved heterogeneity for such large number of dimensions is substantially burdensome. The probability evaluation with high dimensional integrals is potentially affected by several challenges including - requirements of generating high dimensionality of random numbers, empirical identification issues due to relatively flat objective functions in larger dimensions and longer computational run times. In this context, the proposed research contributes to burgeoning econometric and safety literature by developing a joint modeling approach that can accommodate for a large number of dependent variables (considering crash types and severities) within a parsimonious structure. With respect to crash type specific component, instead of considering the crash frequency by crash type as a traditional multivariate distribution, we recasted it as a repeated measures of crash frequency while recognizing that each repetition represents a crash type specific to a zone. At the same time, for the severity component, as opposed to modeling the count events, count proportions by different severity level for a study unit were examined. Finally, we developed a joint model to tie the two components in a single integrated framework while accommodating unobserved heterogeneity across and within the two components (crash frequency and crash severity proportions by types).

In our current research effort, we employed a Panel mixed Negative Binomial- Generalized Ordered Probit Fractional Spilt (PMNB-GOPFS) model where the first component (NB) accommodated for crash frequency by crash type and the later component (GOPFS) studied the fraction of severity outcome for different crash types. The empirical analysis was conducted using the zonal level crash count data for the year 2016 from Central Florida while considering a comprehensive set of exogenous variables including roadway, built environment, land-use, traffic

and sociodemographic characteristics. The empirical analysis involved a series of model estimations including: 1) Independent NB-GOPFS model; 2) Panel NB-GOPFS model without unobserved component parameters; and 3) Joint Panel NB-GOPFS model with unobserved heterogeneity. The comparison exercise, based on the Bayesian Information Criterion (BIC) value highlighted the superiority of the proposed framework that accounts for penalty for additional parameters (model 2 and 3) and within the proposed approach, the model considering unobserved heterogeneity (model 3) outperformed its' counterpart (model 2).

The analysis was further augmented by undertaking a prediction exercise using the final model parameter estimates. One of the major advantages of the proposed framework is that in a single econometric framework, we can predict several dimensions including total crash counts, total crash counts by crash types, crash proportions for each severity level, crash counts for each severity level and finally, proportions and counts of crashes for each crash type by severity. In evaluating the predictive performance, we carried out a comparison exercise between our proposed joint PMNB-GOPFS and the existing traditional multivariate system (RPMNB) for predicting crash counts across different crash severities. Specifically, we compute the RMSE value at disaggregate level and compare their performances across the two models (proposed PMNB-GOPFS and traditional RPMNB). The comparison exercise is conducted using different hold-out sample (50 to be specific) considering different estimation sets to accommodate the influence of different sample sizes in model estimation. The resulting goodness of fit measures clearly highlight the superior/equivalent performance of the proposed PMNB-GOPFS model over the traditional RPMNB model despite having fewer number of parameters. The comparison exercise is further augmented through a simulation exercise (evaluate performance on simulated data) and the results further reinforce the applicability of our proposed PMNB-GOPFS model in crash safety literature.

In summary, the current study contributes to safety literature both methodologically and empirically. Methodologically, we developed a joint framework analysing 24 dependent variables (6*4 from 6 crash types and 4 severities). With this integrated framework, two major enhancements are achieved: 1) increased estimation efficiencies offered by the proposed system (parsimony) and 2) increased interaction across the dependent variables via the observed variables resulting increased model stability with reduced simulation needs. Empirically, by increasing the dimensionality of the dependent variable, the proposed approach allows for flexible consideration of crashes by type and severity within a single framework. Further, the proposed model results offer insights on important variables affecting crash frequency and severity for different crash types. Such macro-level model outcomes can be used to devise safety-conscious decision support tools to facilitate proactive approach in assessing medium and long-term policy-based countermeasures. For instance, transportation planners are required to forecast future crashes given changes in region's characteristics (population increase, addition of new facility (such as road or major facility). The proposed crash prediction models can aid the process. Another application of the planning level models is the consideration of various initiative programs for saving lives. For example, government officials may want to consider initiatives to reduce the number of motor-vehicle related fatalities in a jurisdiction in response to the changes in land-use, and population. Such quantitative exercises will warrant planning level safety models. Thus, it is very useful to develop crash type and severity models at a planning level.

To be sure, the paper is not without limitations. In our study, left-turn and right-turn crashes were considered in the same category due to sample size restrictions despite differences in crash mechanisms of these two categories. In future research, it might be useful to consider these

separately. Moreover, given the inherent aggregation of the dataset, it would also be beneficial to accommodate for the presence of spatial unobserved effects. Further, as the main focus of the paper was on model formulation and estimation of high dimensionality of dependent variables, we focused on one year of data. However, it would be interesting in the future to explore the model development with multiple years to accommodate for temporal effects.

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.

ACKNOWLEDGMENT

The authors would also like to gratefully acknowledge Signal Four Analytics (S4A), Florida Department of Transportation (FDOT) and Department of Revenue (DOR) for providing access to Florida crash data, geospatial data and land-use data.

REFERENCES

Abdel-Aty, M., Wang, X., 2006. Crash Estimation at Signalized Intersections Along Corridors: Significant Factors and Temporal Effects. *Transportation Research Record*, 1953(1), pp. 98–111.

Afghari, A.P., Haque, M.M. and 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 & Prevention, 144, p.105615.

Aguero Valverde, J., Wu, K.F., Donnell, E.T., 2016. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention*, 87, pp. 8–16.

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 & Prevention*, 119, pp. 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, pp. 17–32.

Aptech 2015, Aptech Systems Inc, URL http://www.aptech.com/ (accessed 6.19.18).

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, pp. 28–43.

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

Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35 (7), pp. 677–693.

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

Bhat, C.R., Astroza, S., Lavieri, P.S., 2017. A new spatial and flexible multivariate random-coefficients model for the analysis of pedestrian injury counts by severity level. *Analytic Methods in Accident Research*, 16, pp. 1–22.

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

Bhowmik, T., Yasmin, S. and 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, p.100107.

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, pp. 13–31.

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

Bhowmik, T., Rahman, M., Yasmin, S. and Eluru, N., 2021. Exploring analytical, simulation-based, and hybrid model structures for multivariate crash frequency modeling. *Analytic Methods in Accident Research*, 31, p.100167.

Boulieri, A., Liverani, S., de Hoogh, K., Blangiardo, M., 2017. A space–time multivariate Bayesian model to analyse road traffic accidents by severity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1), pp. 119–139.

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

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 & Prevention*, 99, pp. 330–341.

Chiou, Y.C., Fu, C., 2015. Modeling crash frequency and severity with spatiotemporal dependence. *Analytic Methods in Accident Research*, 5–6, pp. 43–58.

Chiou, Y.C., Fu, C., Hsieh, C.W., 2014. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. *Analytic Methods in Accident Research*, 2, pp. 1–11.

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

El-Basyouny, K., Barua, S. and Islam, M.T., 2014a. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis & Prevention, 73, pp.91-99.

El-Basyouny, K., Barua, S., Islam, M.T., Li, R., 2014b. Assessing the Effect of Weather States on Crash Severity and Type by Use of Full Bayesian Multivariate Safety Models. *Transportation Research Record,* 2432(1), pp. 65–73.

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 & Prevention*,40(3), pp. 1033–1054.

Eluru, N., Chakour, V., Chamberlain, M., Miranda-Moreno, L.F., 2013. Modeling vehicle operating speed on urban roads in Montreal: a panel mixed ordered probit fractional split model. *Accident Analysis & Prevention*, 59, pp. 125–134.

Eluru, N., Yasmin, S., 2015. A note on generalized ordered outcome models. *Analytic Methods in Accident Research*, 8, pp. 1–6.

Fountas, G., Anastasopoulos, P.C., 2017. A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities. *Analytic Methods in Accident Research*, 15, pp. 1–16.

Geedipally, S.R., Patil, S., Lord, D., 2010. Examination of Methods to Estimate Crash Counts by Collision Type. *Transportation Research Record* , 2165(1), pp. 12–20.

Guo, Y., Li, Z., Liu, P. and Wu, Y., 2019a. Exploring risk factors with crashes by collision type at freeway diverge areas: accounting for unobserved heterogeneity. IEEE Access, 7, pp.11809-11819.

Guo, Y., Li, Z., Liu, P. and Wu, Y., 2019b. Modeling correlation and heterogeneity in crash rates by collision types using full Bayesian random parameters multivariate Tobit model. Accident Analysis & Prevention, 128, pp.164-174.

Hosseinpour, M., Sahebi, S., Zamzuri, Z.H., Yahaya, A.S. and Ismail, N., 2018. Predicting crash frequency for multi-vehicle collision types using multivariate Poisson-lognormal spatial model: A comparative analysis. Accident Analysis & Prevention, 118, pp.277-288.

Hosseinpour, M., Yahaya, A.S., Sadullah, A.F., 2014. Exploring the effects of roadway characteristics on the frequency and severity of head-on crashes: Case studies from Malaysian Federal Roads. *Accident Analysis & Prevention*, 62, pp. 209–222.

Huang, H., Chang, F., Zhou, H. and Lee, J., 2019. Modeling unobserved heterogeneity for zonal crash frequencies: A Bayesian multivariate random-parameters model with mixture components for spatially correlated data. Analytic methods in accident research, 24, p.100105.

Jonsson, T., Ivan, J.N. and Zhang, C., 2007. Crash prediction models for intersections on rural multilane highways: Differences by collision type. Transportation research record, 2019(1), pp.91-98.

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, M. and Khattak, A.J., 2019. Case study of crash severity spatial pattern identification in hot spot analysis. Transportation research record, 2673(9), pp.684-695.

Li, Z., Wang, W., Liu, P., Bai, L. and Du, M., 2015. Analysis of crash risks by collision type at freeway diverge area using multivariate modeling technique. Journal of Transportation Engineering, 141(6), p.04015002.

Li, Z., Wang, W., Liu, P., Bigham, J.M., Ragland, D.R., 2013. Using Geographically Weighted Poisson Regression for county-level crash modeling in California. *Safety Science,* 58, pp. 89–97.

Liu, C., Sharma, A., 2018. Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. *Analytic Methods in Accident Research*, 17, pp. 14–31.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy And Practice*. 44(5), pp. 291–305.

Ma, X., Chen, S., Chen, F., 2017. Multivariate space-time modeling of crash frequencies by injury severity levels. *Analytic Methods in Accident Research*, 15, pp. 29–40.

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, pp. 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, pp. 16–26.

Narayanamoorthy, S., Paleti, R., Bhat, C.R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. *Transportation Research Part B: Methodological*, 55, pp. 245–264.

Pai, C.W., Saleh, W., 2008. Modelling motorcyclist injury severity by various crash types at T-junctions in the UK. *Safety Science*, 46(8), pp. 1234–1247.

Papke, L.E., Wooldridge, J.M., 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11(6), pp. 619–632.

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

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), p.47.

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, pp. 44–53.

Shaon, M.R.R., Qin, X., Afghari, A.P., Washington, S., Haque, M.M., 2019. Incorporating behavioral variables into crash count prediction by severity: A multivariate multiple risk source approach. *Accident Analysis & Prevention*, 129, pp. 277–288.

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 & Prevention*, 99, pp. 6–19.

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

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

World Health Organization, 2018. Global status report on road safety 2018: Summary (No. WHO/NMH/NVI/18.20). World Health Organization.

Xie, K., Ozbay, K. and Yang, H., 2019. A multivariate spatial approach to model crash counts by injury severity. Accident Analysis & Prevention, 122, pp.189-198.

Xin, C., Guo, R., Wang, Z., Lu, Q., Lin, P.S., 2017. The effects of neighborhood characteristics and the built environment on pedestrian injury severity: A random parameters generalized

ordered probability model with heterogeneity in means and variances. *Analytic Methods in Accident Research*, 16, pp. 117–132.

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

Yasmin, S., Eluru, N., Lee, J. and Abdel-Aty, M., 2016. Ordered fractional split approach for aggregate injury severity modeling. Transportation Research Record, 2583(1), pp.119-126.

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

Yasmin, S., Momtaz, S.U., Nashad, T. and Eluru, N., 2018. A multivariate copula-based macro-level crash count model. Transportation research record, 2672(30), pp.64-75.

Yasmin, S., Bhowmik, T., Rahman, M. and Eluru, N., 2021. Enhancing non-motorist safety by simulating trip exposure using a transportation planning approach. *Accident Analysis & Prevention*, 156, p.106128.

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

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(3), pp. 443–452.

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

Zeng, Q., Guo, Q., Wong, S.C., Wen, H., Huang, H. and Pei, X., 2019. Jointly modeling area-level crash rates by severity: a Bayesian multivariate random-parameters spatio-temporal Tobit regression. Transportmetrica A: Transport Science, 15(2), pp.1867-1884.

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

Zeng, Q., Wen, H., Huang, H., Pei, X., Wong, S.C., 2018. Incorporating temporal correlation into a multivariate random parameters Tobit model for modeling crash rate by injury severity. *Transportmetrica A: Transport Science*, 14(3), pp. 177–191.

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

APPENDIX A

Example of Correlation Structure

As discussed in the main manuscript, we test for both positive and negative correlations across different components in our proposed joint framework and finally, we select the model that offer best fit. It is important to note that this sign can vary within the matrix elements. For instance, let us consider the correlation structure across the crash counts only for different crash types which is indicated by the $\mathbb{A}_1$ matrix. The structure of the $\mathbb{A}_1$ matrix is as follows:

$$\mathbb{A}_1 = \underset{\text{Crash Types } (1,2,\ldots J,\ J=6)}{\underset{\text{Crash Types } (1,2,\ldots J,\ J=6)}{\begin{bmatrix} & \varrho_1 & \varrho_2 & \ldots & \varrho_J \\ \varrho_1 & 1 & \aleph_{12} & \ldots\ldots & \aleph_{1J} \\ \varrho_2 & \aleph_{12} & 1 & \ldots\ldots & \aleph_{2J} \\ \ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ \varrho_J & \aleph_{1J} & \aleph_{2J} & \ldots\ldots & 1 \end{bmatrix}}} \times \mathfrak{b}_r \qquad \text{(A.1)}$$

a_1

where the first part (a_1) represents the covariance structure across the crash counts of different crash types and the second component ($\mathfrak{b}_r$) is a vector with a value of +1 if the association is positive, -1 if association is negative and 0 if no association is considered. This $\mathfrak{b}_r$ matrix allows us to generate the various possible combinations of positive and negative associations between these correlations. Now, let us consider the following scenarios:

1. A common positive correlation affecting the rear-end (RE), angular (ANG) and sideswipe (SW) crash frequencies simultaneously C_{R1}: indicates that zones with higher number of rear-end crashes will be more likely to have higher number of angular and sideswipe crashes.
2. A common negative correlation across RE, head-on (HD) and single vehicle (SV) crash counts C_{R2}: indicates that zones with higher number of rear-end crashes will be more likely to incur lower number of head-on and single vehicle crashes.
3. A positive correlation between RE and non-motorized (NMT) crash counts C_{R3}
4. A positive correlation between ANG and SW crash counts C_{A1}
5. A negative correlation between SW and SV crash counts C_{S1}
6. A positive correlation between HD and SV crash counts C_{H1}
7. There are no other correlations

With these scenarios, the first and second part of equation A.1 will be:

$$a_1 = \underset{\text{Crash Types}}{\begin{array}{c} \\ RE \\ ANG \\ SW \\ HD \\ SV \\ NMT \end{array}} \begin{array}{c} \begin{array}{cccccc} RE & ANG & SW & HD & SV & NMT \end{array} \\ \begin{bmatrix} 1 & C_{R1} & C_{R1} & C_{R2} & C_{R2} & C_{R3} \\ C_{R1} & 1 & C_{A1} & 0 & 0 & 0 \\ C_{R1} & C_{A1} & 1 & 0 & C_{S1} & 0 \\ C_{R2} & 0 & 0 & 1 & C_{H1} & 0 \\ C_{R2} & 0 & C_{S1} & C_{H1} & 1 & 0 \\ C_{R3} & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}, \mathfrak{b}_r = \begin{array}{c} \begin{array}{cccccc} RE & ANG & SW & HD & SV & NMT \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ -1 & 0 & -1 & 1 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

The reader would note that the final matrix ($\mathbb{A}_1$) is estimated by conducting an element-by-element multiplication of the two-matrixes mentioned above ($a_1 and\ \mathfrak{b}_r$). For instance, lets' assume t (1,2,…6) and u (1,2…6) be the indexes represent the row and column number of the matrixes ($a_1, \mathfrak{b}_r\ and\ \mathbb{A}_1$) respectively. So, each element of the $\mathbb{A}_1$ matrix will be:

$$\mathbb{A}_{1(\mathrm{tu})} = a_{1(tu)} \times \mathfrak{b}_{r(tu)} \qquad \text{(A.2)}$$

Therefore, the final $\mathbb{A}_1$ matrix will be:

$$\mathbb{A}_1 = \begin{array}{c} \\ \textit{Crash Types} \end{array} \begin{array}{c} \begin{array}{cccccc} & \textit{RE} & \textit{ANG} & \text{SW} & \textit{HD} & \textit{SV} & \textit{NMT} \end{array} \\ \begin{bmatrix} \textit{RE} & 1 & C_{R1} & C_{R1} & -C_{R2} & -C_{R2} & -C_{R3} \\ \textit{ANG} & C_{R1} & 1 & C_{A1} & 0 & 0 & 0 \\ \text{SW} & C_{R1} & C_{A1} & 1 & 0 & -C_{S1} & 0 \\ \textit{HD} & -C_{R2} & 0 & 0 & 1 & C_{H1} & 0 \\ \textit{SV} & -C_{R2} & 0 & -C_{S1} & C_{H1} & 1 & 0 \\ \textit{NMT} & -C_{R3} & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

As we can see, using our proposed joint system, within the crash count component itself, we can test for different combination of correlations with different signs. We allow the similar $\mathfrak{b}_k$ or $\mathfrak{b}_{rk}$ terms within the $\mathbb{A}_2\ and\ \mathbb{A}_3$ correlation matrix to allow for the positive and negative correlations.

Independent Model Results

TABLE A.1 Independent Panel NB model

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

No. of shopping center (1)	0.066	2.040	--	--	0.066	2.040	--	--	--	--	--	--
Socio-demographic characteristics												
Non-motorists (3)	0.067	3.996	0.145	7.109	0.144	6.858	--	--	--	--	0.067	3.996
Transit user (1)	0.222	14.898	--	--	--	--	--	--	--	--	0.222	14.898
Over dispersion (6)	0.992	30.183	1.176	25.054	1.024	21.268	1.995	8.123	0.713	19.272	0.455	8.840
Log-Likelihood (np)	-39954.27 (53)											

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

TABLE A.2 Independent Panel GOPFS (Generalized Ordered Probit Fractional Split Model) 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
Threshold 1	0.564	14.394	0.221	6.458	0.948	9.321	-0.039	-0.344	0.264	5.202	-0.665	-7.978
Threshold 2	-0.395	-12.332	-0.492	-19.891	-0.678	-13.289	-0.688	-9.480	-0.808	-23.019	-0.445	-10.656
Threshold 3	-0.262	-6.063	-0.373	-10.521	-0.439	-6.093	-0.505	-6.480	-0.469	-12.778	-0.062	-1.610
Roadway Characteristics												
Arterial road	0.085	3.644	0.183	4.804	--	--	--	--	0.085	3.644	--	--
Possible and non-incapacitating injury	-0.081	-1.710	--	--	--	--	--	--	--	--	--	--
Local road	--	--	--	--	--	--	-0.334	-2.128	--	--	-0.334	-2.128
Number of intersections	--	--	--	--	--	--	-0.051	-3.564	-0.051	-3.564	--	--
Traffic signal	--	--	-0.040	-4.192	-0.040	-4.192	-0.040	-4.192	--	--	--	--
Average inside shoulder width	--	--	--	--	-0.171	-3.479	--	--	--	--	--	--
Average outside shoulder width	-0.046	-1.702	--	--	--	--	--	--	--	--	--	--
Proportion of road over 55mph speed	0.330	5.101	0.330	5.101	--	--	0.876	3.026	0.330	5.101	0.330	5.101
Non-incapacitating and severe injury	-0.669	-2.965	--	--	--	--	--	--	--	--	-1.338	-3.725
Poor pavement condition	--	--	--	--	0.208	2.821	--	--	--	--	--	--
Traffic Characteristic												
Traffic Intensity (Congested)	-0.074	-3.311	-0.074	-3.311	--	--	--	--	--	--	--	--
Non-incapacitating and severe injury	--	--	0.122	1.965	--	--	--	--	--	--	--	--
Truck VMT	--	--	--	--	0.046	4.640	0.046	4.640	--	--	--	--
Land Use Characteristic												

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

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

APPENDIX B

Validation Analysis

For the validation analysis, we compute MAD at a disaggregate level by generating measures at the study unit level (TAZ) and compute the average measures across all units (total crash, crash type and severity). Other than total crash counts and crash count by crash type, we generate crash counts by severity levels for different crash types using the following equation:

$$E(\boldsymbol{P}_{irk}) = \mu_{ir} * \Lambda(y_{irk} = k) \quad (15)$$

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

$$\text{MAD} = \ mean\ |\hat{y}_i - y_i| \quad (16)$$

where, $\hat{y}_i$ and y_i are the predicted and observed, number of crashes occurring over a period of time in a TAZ i (corresponds to different dimension: total crash, crash type, severity etc). Figure B.1 and B.2 presents the value of MAD for estimation and validation sample, respectively.

On the other hand, we employ MAPE measures at an aggregate level where we estimate the number and proportion of crashes for corresponding dimension and predict the TAZ shares for different count and proportion alternatives and compared it with the observed shares. For example, let us consider the crash counts by crash type where we predict the number of crashes for each crash type at an individual level (observation) and then we estimate how many TAZs have 0,1….250 crashes. Finally, we compute the MAPE as:

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

where, $\hat{y}_n$ and y_n are the predicted and observed, number of TAZs (corresponds to different dimension) for different count alternative n. Figure B.1 and B.2 presents the value of MAPE for estimation and validation sample, respectively.

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

(category 3 and 4). Fourth, the MAPE values corresponds to crash types offer interesting insights. While we observe a lower accuracy for rear-end crashes in both datasets (12.4 % and 11.4 % respectively), the model performs adequately for other crash types with a maximum of 6.4% error rate for angular crash in the estimation sample. Finally, within each crash type, the MAPE values for each severity fractions are quite reasonable without any significant trend highlighting the appropriateness of the proposed model.

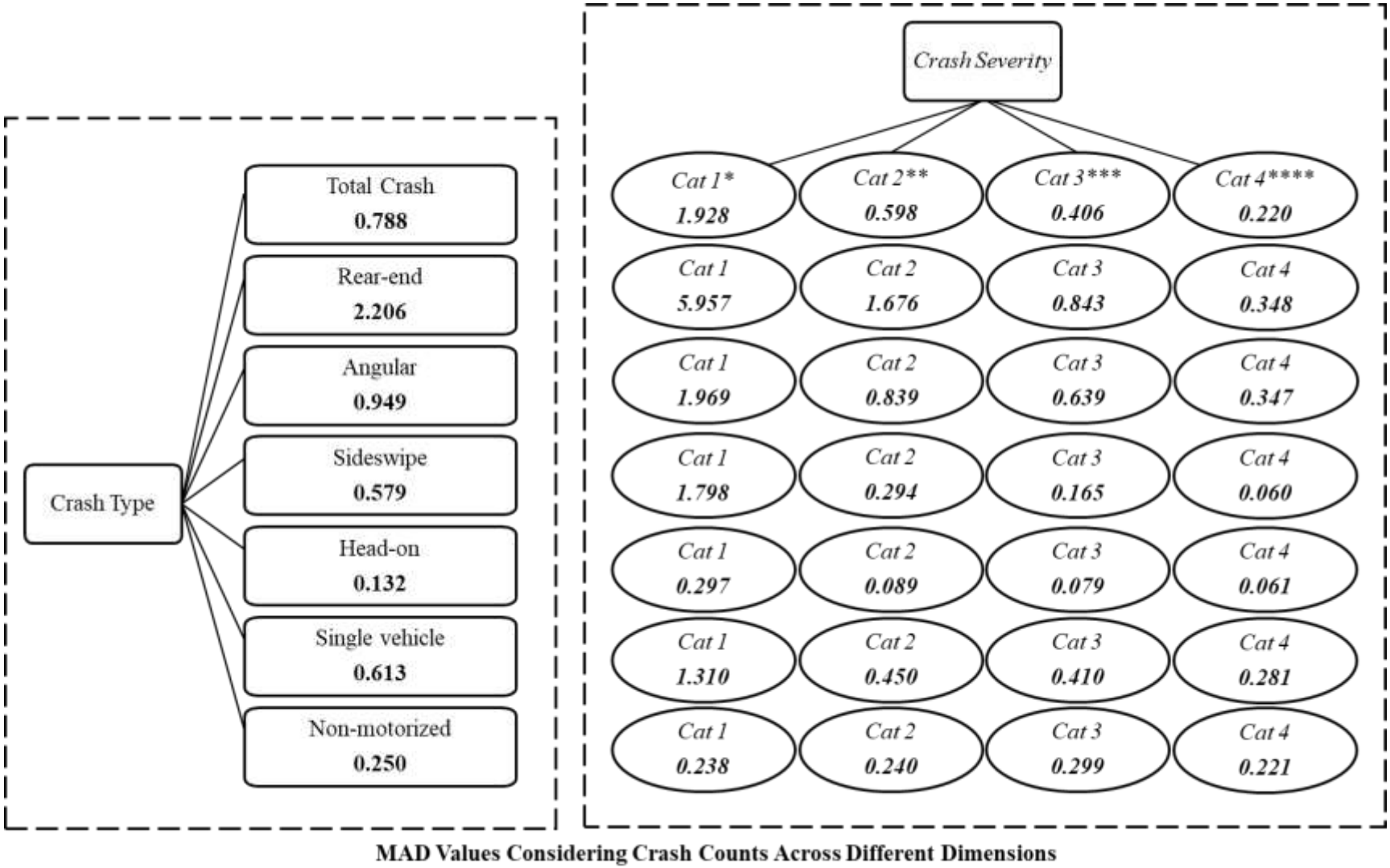


Figure B.1 MAD Tree for Estimation Sample (3,815 TAZs)

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

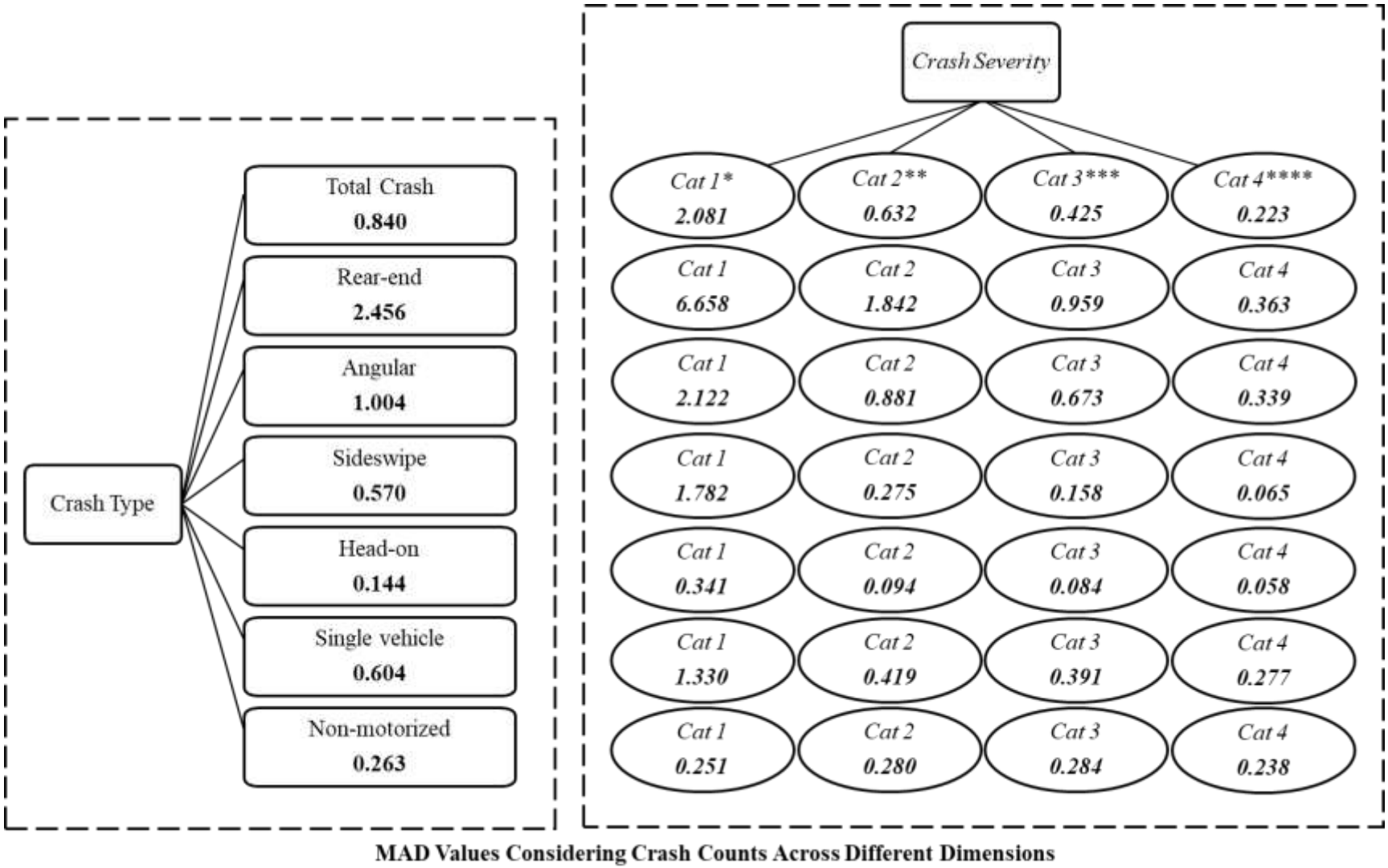


Figure B.2 MAD Tree for Validation Sample (932 TAZs)

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

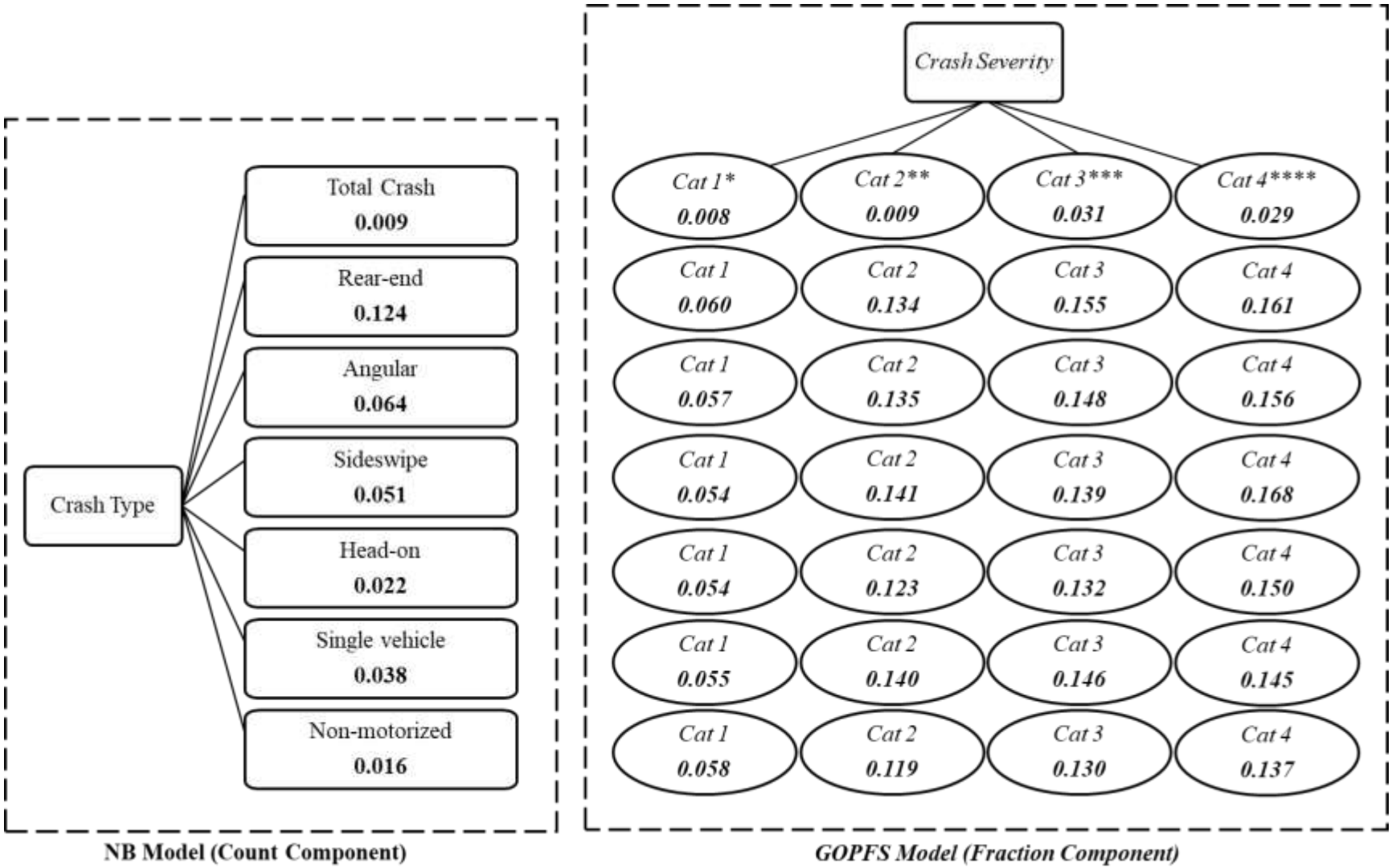


Figure B.3 MAPE Tree for Estimation Sample (3,815 TAZs)

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

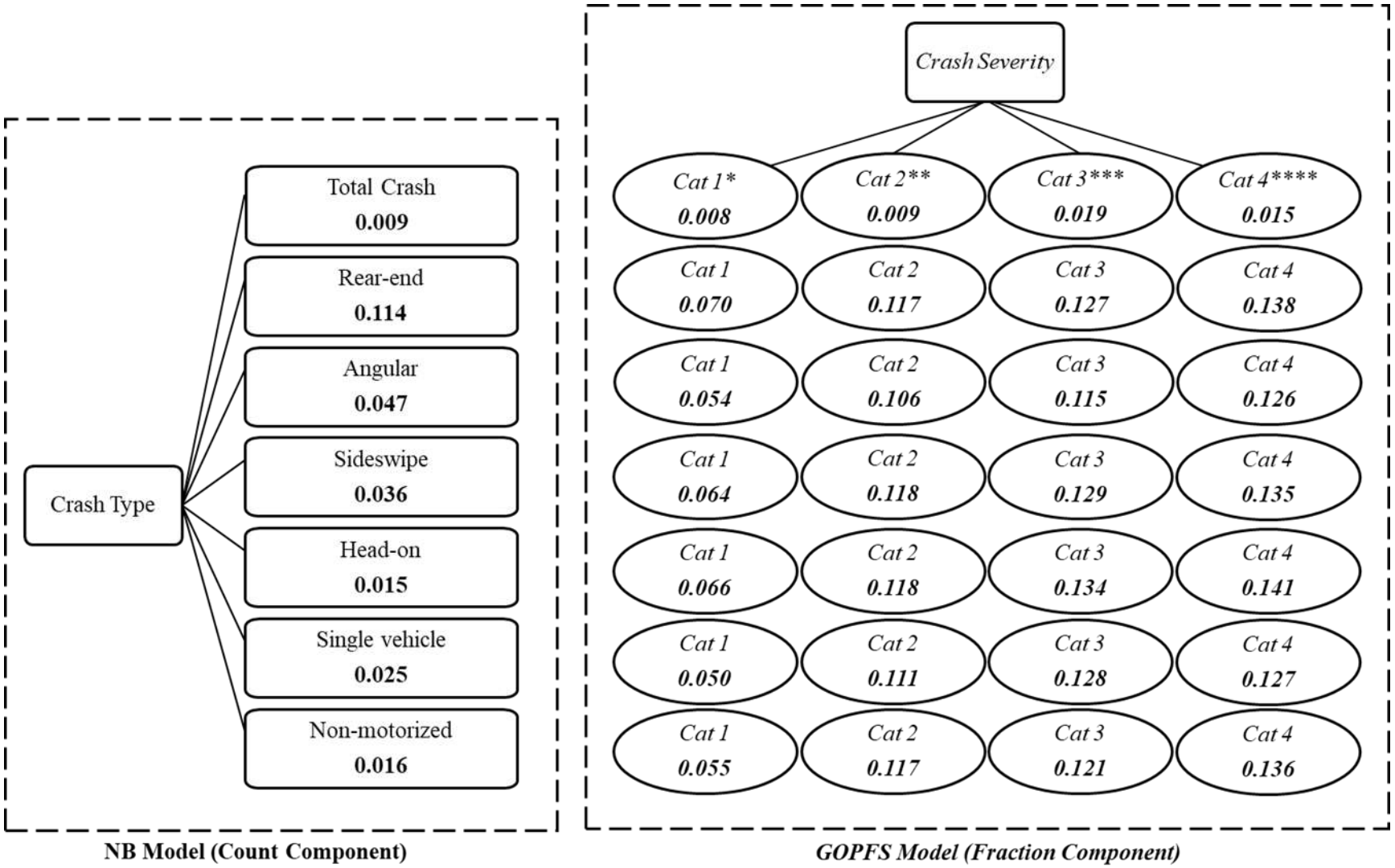


Figure B.4 MAPE Tree for Validation Sample (932 TAZs)

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