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

Kai Wang, Ph.D. (Corresponding Author)
Connecticut Transportation Safety Research Center
Connecticut Transportation Institute
University of Connecticut
Email: kai.wang@uconn.edu

Tanmoy Bhowmik
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: tanmoy78@Knights.ucf.edu

Shanshan Zhao, Ph.D.
Connecticut Transportation Safety Research Center
Connecticut Transportation Institute
University of Connecticut
Email: shanshan.h.zhao@uconn.edu

Naveen Eluru, Ph.D.
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Email: naveen.eluru@ucf.edu

Eric Jackson, Ph.D.
Connecticut Transportation Safety Research Center
Connecticut Transportation Institute
University of Connecticut
Email: eric.d.jackson@uconn.edu

Submitted for publication to the Journal of Safety Research

August 2020

ABSTRACT

Introduction: Predicting crash counts by severity plays a dominant role in identifying roadway sites that experience overrepresented crashes, or an increase in the potential for crashes with higher severity levels. Valid and reliable methodologies for predicting highway accidents by severity are necessary in assessing contributing factors to severe highway crashes, and assisting the practitioners in allocating safety improvement resources. ***Methods****:* This paper uses urban & suburban intersection data in Connecticut, along with two sophisticated modeling approaches, *i.e*. a Multivariate Poisson-Lognormal (MVPLN) model and a Joint Negative Binomial - Generalized Ordered Probit Fractional Split (NB-GOPFS) model to assess the methodological rationality and accuracy by accommodating for the unobserved factors in predicting crash counts by severity level. Furthermore, crash prediction models based on vehicle damage level are estimated using the same two methodologies to supplement the injury severity in estimating crashes by severity when the sample mean of severe injury crashes (*e.g*. fatal crashes) is very low. ***Results****:* The model estimation results highlight the presence of correlations of crash counts among severity levels, as well as the crash counts in total and crash proportions by different severity levels. A comparison of results indicates that injury severity and vehicle damage are highly consistent. ***Conclusions***: Crash severity counts are significantly correlated and should be accommodated in crash prediction models. ***Practical Application***: The findings of this research could help select sound and reliable methodologies for predicting highway accidents by injury severity. When crash data samples have challenges associated with the low observed sampling rates for severe injury crashes, this research also confirmed that vehicle damage can be appropriate as an alternative to injury severity in crash prediction by severity.

KEYWORDS

Crash Prediction by Severity; Crash Severity Surrogate; Unobserved Heterogeneity; Multivariate Model; Joint Fractional Split Model

1 INTRODUCTION

1.1 Motivation

Each year there are over 38,000 motor vehicle crashes related fatalities in the United States of America, and traffic collisions are one of the most significant causes of untimely death (NHTSA, 2016). Traffic safety is a top priority for both Federal and State transportation agencies and there is still a critical need for effective strategies to reduce crashes and improve highway safety.

Crash prediction models are one of the most effective approaches to help identify roadway locations with overrepresented crashes or the potential for crashes in the future. These predictive model results can then be used to implement countermeasures to improve highway safety. Therefore, selecting an appropriate and effective crash prediction model is critical when trying to identify roadway sites to prioritize for safety improvement. The use of inaccurate or invalid modeling approaches and assumptions, might result in biased crash prediction results and thus lead to the inefficient use of safety improvement resources and reduce the effectiveness of the safety management process. Given the limited safety improvement resources available, sites that experience overrepresented high severity crashes, should be our top priority. The development of reliable crash prediction methodologies, based on crash severity, is imperative for helping to identify hazardous roadway locations and crash contributing factors. This allows for the efficient allocation of highway safety improvement strategies to assist in preventing crashes from occurring in the future.

The first edition of the Highway Safety Manual (HSM, 2010) introduces the crash prediction models for total crashes, which are then multiplied by several constant severity proportions to predict the crashes by different severity levels. This approach might not be feasible as crash severity distributions may vary across sites due to potential variations related to roadway geometric, traffic, and environmental characteristics (Wang *et al.,* 2017; Ma and Kockelman, 2006; Ma *et al.*, 2008; Wang *et al.,* 2017; Wang *et al*., 2018, 2019). Therefore, crash prediction models by severity have been widely investigated to improve the prediction performance of crash counts by different severity levels (Tarko *et al*., 2008; Dixon *et al*., 2015; Oh *et al*., 2004; Russo *et al*., 2016; Liu and Sharma, 2018; Abdel-Aty and Radwan, 2000; Lord and Persaud, 2000; Ulfarsson and Shankar, 2002). In general, two types of methodological frameworks of crash prediction models have been implemented by researchers to achieve a better crash severity count prediction. The first alternative is to estimate crash counts by different severity levels directly. The second alternative which is usually referred to as the two-stage model, is first to estimate crash counts in total, followed by estimating crash severity distributions, and then combine the latter with the former for crash count prediction by severity.

Estimating crash prediction models by severity might be challenging due to the small sample size and low sample mean (Anarkooli *et al*., 2019), especially for the fatal and severe injury crashes. This alternative creates an issue for identifying locations with overrepresented severe crashes when the safety improvement resources are limited. To this end, one objective indicator of crash consequences - the extent of vehicle damage based on the destruction/deformation of the vehicle involved in the crash might be used - to represent the crash consequence as a supplement of

injury severity. The rationality of modeling crashes by vehicle damage level is because 1) the sample mean of crashes with severe vehicle damage levels is higher than the crashes with severe injuries, and 2) the vehicle damage is found to be positively correlated with the injury severity in multiple studies (Wang *et al*., 2015; Qin *et al.,* 2013; Wang *et al*., 2019). For this reason, roadway locations experienced overrepresented crashes with severe vehicle damage levels have a very high potential to experience more severe injury crashes in the future.

1.2 Literature Review

With regard to the methodologies that directly estimate crash counts by severity, the Poisson regression model has been initially used to model crashes by each severity level since the crash frequencies are non-negative integers (Lord and Mannering, 2010). The Poisson model has its implicit restriction - the variance of the data is assumed to be equal to the mean. This assumption might not always be valid as the variance of crash data usually is higher than the mean, which is also known as the over-dispersion (Washington *et al.,* 2011). To address this issue, the Univariate Poisson Lognormal regression and Negative Binomial regression models are then used to predict crash counts by severity (Washington *et al.,* 2011; Mannering and Bhat, 2014). However, traditional Univariate Poisson Lognormal and Negative Binomial models assume crash counts by crash severity to be independent. However, this might not be true due to the presence of shared unobserved factors across different severity levels for each observational record. Modeling crash severity counts together without accounting for their correlations might yield biased parameter estimates, and reduce model prediction accuracy (Ma and Kockelman, 2006; Ma *et al.,* 2008; Wang *et al.,* 2017; Wang *et al*., 2018; Mannering and Bhat, 2014; Mannering *et al*., 2016).

To address correlations among crash counts across different severity levels, a large number of methodologies have been implemented to estimate crash counts by severity jointly. These include, but are not limited to Simultaneous Equations Model (Ye *et al.,* 2013); Multivariate Generalized Poisson Model (Chiou and Fu, 2013, 2015; Chiou *et al*., 2014); Joint-Probability Model (Pei *et al.,* 2011); and Artificial Neural Network (Zeng *et al*., 2016). Recently, the Multivariate regression models have been extensively applied to simultaneously estimate crash counts by severity by accounting for the correlations between crashes among different crash severity levels. Multivariate models have been verified to be superior to the Univariate models in terms of the parameter estimation and crash prediction accuracy. For instance, Ma and Kockelman (2008) applied a Multivariate Poisson-Lognormal model to estimate the crash counts by severity, and they found the crash counts highly correlated among different severity levels. Park and Lord (2007) applied a Multivariate Poisson-Lognormal model to jointly estimated the crash frequencies by severity using the California data. The study implied that the crash frequencies are highly correlated among severity levels, and the Multivariate model obtains more accurate parameter estimates. Wang *et al*. (2017) used a Multivariate Lognormal approach to estimate crash count models for rural two-lane undivided highways, and the results were compared to the Univariate models. The study verified that the Multivariate Lognormal model provides unbiased parameter estimates and significantly enhances the prediction accuracy. A similar study was conducted by Wang *et al*. (2018) for freeway crash prediction, and the results highlighted that freeway crashes significantly correlated among different levels of crash severity. Anastasopoulos *et al*. (2012) used both Multivariate Tobit and Multivariate Negative Binomial models to predict crash rate by severity on multilane divided highways in Washington State. The

study found that the prediction accuracy between the two approaches are very close, and both methods outperform the univariate models.

Furthermore, the Multivariate models have also been extended by researchers for particular perspectives. For instance, to account for the issues of excess zero, unobserved heterogeneity and spatial-temporal correlation in crash data, methodologies including but are not limited to Multivariate Random-Parameter Zero-Inflated model (Dong *et al.*, 2014), Multivariate Poisson Lognormal Spatial model (Barua *et al.*, 2014; Aguero-Valverde, 2013), Multivariate Spatial-Temporal Bayesian model (Liu and Sharma, 2018), Multivariate Poisson Lognormal Conditional-Autoregressive model (Wang and Kockelman, 2013; Xie *et al*., 2019) and Multivariate Random Parameter Spatial Poisson Lognormal model (Barua *et al.*, 2016) were then used to estimate the crash counts by severity. Lord and Mannering (2010) provided comprehensive guidance on model selection and assessment in crash count prediction.

Now on to the second alternative approach described above. Qin *et al*. (2013) used a Negative Binomial model and a Multinomial Logit model to predict total truck crashes and crash counts by each severity level. Chiou and Fu (2013, 2015) and Chiou *et al*. (2014) examined the use of a Multinomial model and a Generalized Poisson model to predict crash frequencies by severity. Anarkooli *et al*. (2019) applied a Negative Binomial model and a Generalized Ordered Probit model to estimate crashes by severity on horizontal curves. Geedipally *et al*. (2013) used a Multinomial Logit model to estimate the severity distributions for freeway segments and interchanges. Wang *et al*. (2011) applied a Bayesian spatial model and a mixed logit model to estimate crashes by severity for the major roads in England. Savolainen *et al*. (2011) provided

comprehensive guidance on model selection and assessment of prediction of crash severity distributions.

However, all of these studies modeled the total crash counts and crash severity distributions separately and independently, which might be inappropriate due to the common observed and unobserved factors that affect both crash counts and crash severity distributions. Yasmin and Eluru (2016) introduced a new modeling framework - The Joint Negative Binomial - Ordered Probit Fractional Split (NB-OPFS) model to estimate the total crashes and crash severity distributions simultaneously. In their method, a Negative Binomial component was employed for estimating crashes in total, and an Ordered Probit fractional split component was employed for estimating crash proportions by severity. Unlike previous studies, their modeling framework jointly estimates the Negative Binomial component and the Ordered Probit component, by accounting for the unobserved heterogeneity across and within the crash count and crash severity proportion modeling components. Further, by implementing the Ordered Probit framework, the method also accounts for the ordinal nature of crash severity in the crash proportion estimation. The authors then further extended their methodology (Yasmin and Eluru, 2018) to a Joint Negative Binomial - Generalized Ordered Logit Fractional Split modeling framework to estimate crash counts by severity at a zonal level for Florida State. This method allowed the correlation between total crash counts and crash severity proportions to vary across zones. The study highlighted the superiority of the joint model framework in terms of the prediction accuracy compared to the independent model framework. Bhowmik *et al*. (2019) applied a Panel Mixed Generalized Ordered Probit Fractional Split model to examine the contributing factors to vehicle operating speed. The study found that roadway related characteristics significantly affect the

vehicular speed, and the proposed model framework performs adequately for the speed prediction.

1.3 Problem Statement, Study Objectives and Contributions

Although different methodologies have been applied in predicting crashes by severity, multiple issues are still existent and need to be addressed. For instance:

1. Most of the previous studies focused on implementing one of the two options for crash prediction by severity, *i.e*. either predicting crash severity counts simultaneously or predicting total crash counts and crash severity proportions together. There is a lack of study in assessing and comparing these two options in highway safety research, which can offer insights on the pros and cons of each method, and shed light on method selection under different data conditions and research needs.
2. Although previous research has verified that the low sample mean of severe crashes (*e.g*. fatal crashes) leads to difficulties in crash prediction by severity level, limited research provided effective alternatives to address this issue. The shortage of crash prediction capability for severe crashes creates troubles to practitioners for identifying target roadway locations, when the highway safety improvement resources are limited.

Accordingly, three major objectives and contributions are addressed and made by this study respectively. They are:

1. Assess and identify the most reliable methodology in predicting crashes by severity, using and extending the two advanced statistical methodologies, *i.e*. the Multivariate Poisson-Lognormal (MVPLN) model and the Joint Negative Binomial - Generalized Ordered Probit Fractional Split (NB-GOPFS) model.

2. Identify and interpret the contributing factors to severe crashes.
3. Evaluate the rationality of using the vehicle damage as an alternative to injury severity in crash prediction models by severity level, to provide practitioners with capabilities of effectively allocating safety improvement resources when the low sample mean leads to difficulties in predicting severe crashes.

The remaining parts of this paper are as follows: the second section describes the two methodological frameworks and the model estimation methods, the third section describes the data used in model estimation, and the fourth section discusses the model estimation results. Model comparisons are provided in the fifth section and conclusions are discussed in the final section.

2 METHODOLOGIES

2.1 Framework for Multivariate Poisson - Lognormal (MVPLN) Model

The first method used to estimate crash counts by severity is the MVPLN model. Assume $y_i = \left(Y_{1i}, Y_{2i} \dots, Y_{Ji}\right)'$ for $i = 1, 2, \dots, N$ be a *J*-dimensional vector (*i.e*. *J* crash severity levels) of crash counts across all *N* sites. In the MVPLN model, we assume the crash counts are correlated among all severity levels. The MVPLN model can be derived as (Serhiyenko *et al.*, 2016):

$$Y_{ji}|\lambda_{ji} \sim Poisson(\lambda_{ji}) \quad (1)$$

where λ_{ji} is the mean of Poisson distribution, which is estimated as:

$$ln(\lambda_{ji}) = Offset + \boldsymbol{\beta}_j \boldsymbol{x_{ji}} + \varepsilon_{ji} \quad (2)$$

where $Offset$ is the log exposure for total observation days in the data set for intersection models (*i.e.* in this study, the offset for both sign-controlled and signalized intersections is *log*(365*5)=7.51). $\boldsymbol{x}_{ji}$ is a vector of independent variables and $\boldsymbol{\beta}_j$ is a vector of coefficients to be

estimated. ε_{ji} is a random term. Assume a vector of the random term $\boldsymbol{\varepsilon_i} = \left(\varepsilon_{1i}, \varepsilon_{2i} \ldots, \varepsilon_{Ji}\right)'$ at site *i* follows a *J*-dimensional normal distribution, *i.e.*

$$\boldsymbol{\varepsilon_i} \sim Normal(\mathbf{0}, \boldsymbol{\Sigma}) \tag{3}$$

where **0** is a *J*-dimensional zero vector, **Σ** is a *J* **J* variance-covariance matrix and let's define $\boldsymbol{\Sigma}=(\sigma_{rs})_{1\le r<s\le J}$. Then the mean, variance and covariance of the crash counts by each severity level at site *i* can be derived as (Serhiyenko *et al.*, 2016):

$$Mean = E\left[Y_{ji}\right] = exp(Offset + \boldsymbol{\beta}_j \boldsymbol{x}_{ji}) exp(\tfrac{\sigma_{jj}}{2}) \tag{4}$$

$$Variance = Var\left[Y_{ji}\right] = exp\left(Offset + \boldsymbol{\beta}_j \boldsymbol{x}_{ji}\right) exp\left(\tfrac{\sigma_{jj}}{2}\right) + exp\big(2\big(Offset + \boldsymbol{\beta}_j \boldsymbol{x}_{ji}\big)\big) \left(exp^2\left(\sigma_{jj}\right) - exp(\sigma_{jj})\right) \tag{5}$$

$$Covariance = Cov[Y_{ri}, Y_{si}] = exp(Offset + \boldsymbol{\beta}_j \boldsymbol{x}_{ji}) exp(\tfrac{\sigma_{rr}}{2}) exp(\tfrac{\sigma_{ss}}{2}) (exp(\sigma_{rs}) - 1) \tag{6}$$

The correlations of crash counts between r^{th} and s^{th} crash severity can be accommodated by the covariance term in equation (Wang *et al.*, 2017) through the σ_{rs}, which is the off-diagonal entry of the *J* **J* variance-covariance matrix. A positive σ_{rs} represents a positive correlation of crash counts between r^{th} and s^{th} crash severity, and a negative σ_{rs} represents a negative correlation of crash counts between r^{th} and s^{th} crash severity. The σ_{rs} can be further derived as:

$$\sigma_{rs} = \rho_{rs} \sqrt{\sigma_{ss} * \sigma_{rr}} \tag{7}$$

where σ_{ss} and σ_{rr} are the diagonal entries of the *J* **J* variance-covariance matrix, and ρ_{rs} is a traditional correlation coefficient to be estimated which is between -1 and 1. The probability distribution of the given total crash counts $\boldsymbol{y}_i$ can be written as (Serhiyenko *et al.*, 2016):

$$g\left(\boldsymbol{y}_i \middle| \boldsymbol{\beta}_j \boldsymbol{x}_{ji}, \boldsymbol{\Sigma}\right) = \int \ldots \int f_{Normal,J}(\boldsymbol{\varepsilon}_i | \mathbf{0}, \boldsymbol{\Sigma}) \prod_{j=1}^{J} f_{Poisson}\left(y_{ij} \middle| \varepsilon_{ji}, \boldsymbol{\beta}_j \boldsymbol{x}_{ji}\right) d\boldsymbol{\varepsilon}_i \tag{8}$$

where $f_{Normal,J}$ is a *J*-dimensional normal distribution function, and $f_{Poisson}$ is a Poisson distribution. As noted from previous studies (Wang *et al.*, 2017; Serhiyenko *et al.*, 2016), the

probability distribution function shown in equation 8 has no closed algebraic solution, and hence the Bayesian framework is used to estimate the coefficients in the MVPLN model. First assume every $\boldsymbol{\beta}_j$ in equation 2 follows a prior normal distribution as *Normal* (*0, ϭ²*) and every $\boldsymbol{\Sigma^{-1}}$ in equation 3 follows a prior Wishart distribution as *Wishart* (*c*, ϖ), where *ϭ²*, *c* and ϖ are all hyperparameters for priors. We used the default hyperparameter specifications in R-INLA (2020) for both the Normal prior ($i.e.\beta_j \sim Normal\ (0, 10^3)$) and the Wishart prior with *c* = 7 (*i.e.* 2*J*+1) degrees of freedom and an identify matric as the precision matrix ϖ. The posterior distributions of the coefficients are estimated using the Bayesian inference (Serhiyenko *et al.*, 2016).

The Markov Chain Monte Carlo (MCMC) (Ma and Kockelman, 2006; Ma *et al.*, 2008) simulation, which uses the Gibbs sampler and Metropolis-Hasting (M-H) approach, is usually applied to carry out the Bayesian inference on model estimation. However, studies have noticed that the MCMC simulation approach is extremely computationally challenging and time-consuming, especially for a large data sample (Mannering and Bhat, 2014). To address this issue and simplify the model estimation procedure, we applied the Integrated Nested Laplace Approximation (INLA) approach proposed by Rue *et al.* (2009) to carry out the Bayesian inference of the MVPLN model estimation in this study. The INLA approach doesn't rely on the MCMC and it numerically approximates the posterior distributions of parameters. It has been verified to be able to significantly reduce the running time compared to the MCMC approach by multiple studies (Serhiyenko *et al*., 2016; Wang *et al*., 2017; Wang *et al*., 2018). The R-INLA (2020) package was used to run the MVPLN models. The detailed discussions of the INLA

approach and model estimation procedures are referenced in several previous studies (Serhiyenko *et al.*, 2016; Rue *et al*., 2009).

2.2 Framework for Joint Negative Binomial - Generalized Ordered Probit Fractional Split (NB-GOPFS) Model

In the NB-GOPFS model, the total crash counts and crash proportions by each severity are jointly estimated, by accounting for the correlations between total crashes and crash severity proportions. Therefore, two model components are included in the NB-GOPFS method, where a count model (*i.e*. a Negative Binomial framework is used in this study) is used to estimate the total crash counts, and a fractional split model (*i.e*. a Generalized Ordered Probit Fractional Split framework) is used to estimate the crash proportions by each severity level. Similar to the MVPLN model framework, assume *i* (*i = 1, 2, 3 … N*) to be the index for the roadway site, and *j* (*j = 1, 2, 3 … J*) to be the index for the injury severity category. The total crash counts y_i at site *i* can be estimated using a NB framework, which is derived as:

$$Prob[y_i|\mu_i] = p(y_i) = \frac{\Gamma[(\sigma)+y_i]}{\Gamma(\sigma)y_i!}\left[\frac{\sigma}{(\sigma)+\mu_i}\right]^{\sigma}\left[\frac{\mu_i}{(\sigma)+\mu_i}\right]^{y_i} \quad (9)$$

where *Γ* is a gamma function; σ is the inverse overdispersion parameter in the NB model, and μ_i is the expected crash counts at site *i*, which can be written as:

$$ln(\mu_i) = Offset + (\boldsymbol{\beta} + \boldsymbol{\zeta_i})\boldsymbol{x}_i + \varepsilon_i + \eta_i \quad (10)$$

where $\boldsymbol{x}_i$ is a vector of independent variables associated with site *i* and $\boldsymbol{\beta}$ (not including a constant) is a vector of coefficients to be estimated. $\boldsymbol{\zeta_i}$ (which follows a standard normal distribution: $\boldsymbol{\zeta_i} \sim N(\mathbf{0}, \boldsymbol{\pi}^{\mathbf{2}})$) is a vector of estimated coefficients which accounts for the unobserved heterogeneity in crash count estimation at site *i*. $exp(\varepsilon_i)$ is a random term which follows a gamma distribution with mean 1 and variance σ. η_i is a random factor which

accommodates the correlations between total crash counts and crash severity proportions at site *i*, due to the common unobserved factors.

Considering the ordinal nature of crash severity, the estimation of proportions by each crash severity level is carried out by a Generalized Ordered Probit Fractional Split (GOPFS) framework. Let's define the p_{ji} be the actual proportion of crash severity *j* at site *i*, which is assumed to be associated with a latent variable p_i^*. The latent variable can be specified as (Yasmin and Eluru, 2018):

$$p_i^* = (\boldsymbol{\gamma} + \boldsymbol{\rho_i})\boldsymbol{z}_i + \delta_i + \eta_i, \quad p_{ji} = j \;\; if\; \epsilon_{i,j-1} < p_i^* < \epsilon_{i,j} \tag{11}$$

This latent propensity p_i^* is mapped to the actual severity proportion categories p_{ji} by the ϵ thresholds (ϵ_0 =-∞ and ϵ_j= ∞). $\boldsymbol{z}_i$ is a vector of attributes that influences the propensity associated with crash severity proportions. $\boldsymbol{\gamma}$ is a corresponding vector of mean effects, and $\boldsymbol{\rho_i}$ is a vector of unobserved factors on severity proportion propensity for site *i* and its associated characteristics assumed to be a realization from standard normal distribution: $\boldsymbol{\rho_i} \sim N(0, \boldsymbol{k}^2)$. δ_i is an idiosyncratic random error term assumed to be identically and independently standard normal distributed across observational unit *i*. η_i is a random factor which accommodates the correlations between total crash counts and crash severity proportions at site *i*, due to the common unobserved factors.

The GOPFS model relaxes the constant thresholds across observations 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 to account for the heterogeneity. Thus, the thresholds are expressed as:

$$\epsilon_{i,j} = fn(s_{ij}) \quad (12)$$

where, s_{ij} is a set of exogenous variables (including a constant) associated with j^{th} threshold. Further, to ensure the accepted ordering of observed severity proportions $(-\infty < \epsilon_{i,1} < \epsilon_{i,2} < \ldots\ldots\ldots < \epsilon_{i,j-1} < +\infty)$, we use the following parametric form as employed by Eluru *et al*. (2008):

$$\epsilon_{i,j} = \epsilon_{i,j-1} + exp((\tau_j + \theta_{ji})\, s_{ij} + \eta_i) \quad (13)$$

where, τ_j is a vector of parameters to be estimated. θ_{ji} is another vector of unobserved factors moderating the influence of attributes in s_{ij} on the severity proportions for analysis unit *i* and injury severity category *j*. It is noted from equation 11 that p_{ji} is the actual proportion of crash severity *j*, which is different from the traditional generalized ordered Probit model framework where the dependent variable is an indicator of crash severity level. In order to estimate the generalized order Probit framework with a continuous dependent variable, let's assume (Yasmin and Eluru, 2018)

$$E(p_{ji}|\boldsymbol{x}_i) = H_{ji}(\gamma, \epsilon), \quad 0 \le H_{ji} \le 1,\ \sum_{j=1}^{J} H_{ji} = 1 \quad (14)$$

H_{ji} accounts for the ordered Probit probability (P_{ji}) form for the crash severity level *j*, and it is defined as:

$$P_{ji} = \phi\{\epsilon_{i,j} - [(\boldsymbol{\gamma} + \boldsymbol{\rho_i})\boldsymbol{z}_i + \delta_i + \eta_i]\} - \phi\{\epsilon_{i,j-1} - [(\boldsymbol{\gamma} + \boldsymbol{\rho_i})\boldsymbol{z}_i + \delta_i + \eta_i]\} \quad (15)$$

where $\phi(\bullet)$ is the cumulative standard normal distribution. It is noted from previous research (Yasmin and Eluru, 2018) that the correlations between total crash counts and crash proportions by severity may vary across sites. Therefore, we parameterize the correlation parameter in this study as follows:

$$\eta_i = \alpha \boldsymbol{c}_i \quad (16)$$

where, $\boldsymbol{c}_i$ is a vector of exogenous variables, α is a vector of unknown parameters to be estimated (including a constant).

To jointly estimate the NB probability function (see equation 9) for total crash counts and the GOPFS probability function (see equation 15), let's define a structure Ω for all vectors (*i.e.* $\boldsymbol{\zeta}, \boldsymbol{\rho}, \boldsymbol{\theta}\ and\ \alpha$) that account for unobserved heterogeneity, either in NB or GOPFS model framework, and $\Omega \sim N(0, (\boldsymbol{\pi^2}, \boldsymbol{k}^2, \boldsymbol{m^2}, \boldsymbol{n^2}))$. The likelihood function of the Joint NB-GOPFS model can be written as:

$$L_i = \int_{\Omega}\ p(y_i) \times \prod_{j=1}^{J} {P_{ji}}^{\omega_i p_{ji}} d\Omega \tag{17}$$

where ω_i is a dummy indicator where $\omega_i = 1$ represents site *i* has at least one crash, otherwise $\omega_i = 0$. The log-likelihood function can then be written as:

$$LL = \sum_i Ln(L_i) \tag{18}$$

Overall, the parameters to be estimated in the Joint NB-GOPFS model are $\boldsymbol{\beta, \gamma, \sigma, \epsilon, \pi, k, m}$ and $\boldsymbol{n}$. The Quasi-Monte Carlo simulation approach based on the scrambled Halton sequence is applied to estimate the log-likelihood function, using the GAUSS Matrix Programming Software (Aptech, 2019). The detailed discussions of the Joint NB-GOPFS approach and model estimation procedures are referenced in several previous studies (Yasmin *et al*., 2016; Yasmin and Eluru, 2013, 2018; Bhowmik *et al*., 2019; Bhat, 2001; Eluru *et al*., 2008).

3 DATA PREPARATION

To estimate and compare the MVPLN and Joint NB-GOPFS models for crash prediction by severity, urban & suburban intersections were collected from the State of Connecticut and five-year crash data (2014-2018) were collected from the Connecticut Transportation Crash Data Repository (CTCDR) (2019) and assigned to the specific intersections.

A total of 895 intersections are sign-controlled and 1,178 are signalized. To obtain sufficient observations in each crash severity level, crash severity counts were aggregated into three categories (Wang *et al*., 2017; Wang *et al*., 2018):

1) K+A which combines fatal (K) and incapacitating injury (A) crashes;

2) B+C which combines non-incapacitating injury (B) and possible injury crashes (C), and

3) PDO which includes the property damage only (PDO) crashes.

As mentioned earlier, vehicle damage is used as another crash consequence indicator to supplement the crash injury severity in this study to further address the low sample mean issue in crash prediction models, especially for estimating models for crashes with severe injuries such as K and A crashes. According to the Model Minimum Uniform Crash Criteria (MMUCC) guideline (2017), vehicle damage was categorized into five levels and vehicle damage counts were aggregated into three categories in this study (Wang *et al*., 2015; Qin *et al*., 2013; Wang *et al*., 2019):

1) Severe Damage Crashes which contain all crashes with disabling (salvageable or total loss) damage;

2) Moderate Damage Crashes which contain all crashes with broken or missing parts damage, and

3) Minor Damage Crashes which combine crashes with minor/cosmetic damage and crashes with no damage.

To validate the assumption of using vehicle damage to supplement the injury severity to address the low sample mean issue in crash prediction models, Table 1 presents the scatter plots and

Pearson correlation coefficients between the injury severity and vehicle damage for both sign-controlled and signalized intersections. The Pearson correlation coefficients illustrated that the injury severity and vehicle damage are highly correlated for both sign-controlled and signalized intersections. Furthermore, intersection traffic and geometric data were collected based on the urban & suburban arterials chapter in the Highway Safety Manual (HSM) (2010) with regard to both sign-controlled and signalized intersections. Table 1 summarizes the descriptive characteristics of the intersection and crash data used in this study.

TABLE 1 Descriptive Characteristics of Urban & Suburban Intersection and Crash Data

Variables	**Sign-Controlled Intersections (895 Intersections)**		**Signalized Intersections (1,178 Intersections)**	
Crash Data				
K+A Crash	Min. = 0; Max. = 2; Mean =0.1; Std. Dev = 0.3		Min. = 0; Max. = 5; Mean = 0.2; Std. Dev = 0.5	
B+C Crash	Min. = 0; Max. = 19; Mean = 1.6; Std. Dev = 2.3		Min. = 0; Max. = 54; Mean = 6.2; Std. Dev = 6.3	
PDO Crash	Min. = 0; Max. = 61; Mean = 5.1; Std. Dev = 6.4		Min. = 0; Max. = 146; Mean = 20.3; Std. Dev = 18.4	
Severe Damage Crash	Min. = 0; Max. = 22; Mean = 2.4; Std. Dev = 2.7		Min. = 0; Max. = 60; Mean = 5.9; Std. Dev = 5.1	
Moderate Damage Crash	Min. = 0; Max. = 23; Mean = 1.4; Std. Dev = 2.0		Min. = 0; Max. = 40; Mean = 5.0; Std. Dev = 5.3	
Minor Damage Crash	Min. = 0; Max. = 54; Mean = 3.3; Std. Dev = 4.5		Min. = 0; Max. = 129; Mean = 15.5; Std. Dev = 15.7	
Intersection Data	**Frequency**	**Percentage**	**Frequency**	**Percentage**
3-Leg Intersection	687	76.8%	423	35.9%
4-Leg Intersection	208	23.2%	755	64.1%
Partial-Way Sign-Controlled Intersection	806	90.1%	NA	NA
All-Way Sign-Controlled Intersection	89	9.9%	NA	NA
Median Presence at Intersection Approaches	83	9.3%	285	24.2%
Illumination Presence	599	66.9%	877	74.4%
Driveway Presence	363	40.6%	478	40.6%
Exclusive Left-Turn Lane Presence	51	5.7%	766	65.0%
Exclusive Right-Turn Lane Presence	40	4.5%	586	49.7%
Protected Left-Turn Signal Phasing Presence	NA	NA	875	74.3%
No Right-Turn-On-Red	NA	NA	654	55.5%
Major Road AADT	Min. = 550; Max. = 32,800; Mean = 9,180; Std. Dev = 4,842		Min. = 2300; Max. = 68200; Mean = 15,769; Std. Dev = 6,676	
Minor Road AADT	Min. = 20; Max. = 13,900; Mean = 2,534; Std. Dev = 2,212		Min. = 300; Max. = 43,300; Mean = 7,482; Std. Dev = 5,466	
Intersection Skew Angle	Min. = 0; Max. = 89; Mean = 22; Std. Dev = 20		Min. = 0; Max. = 90; Mean = 22; Std. Dev = 23	

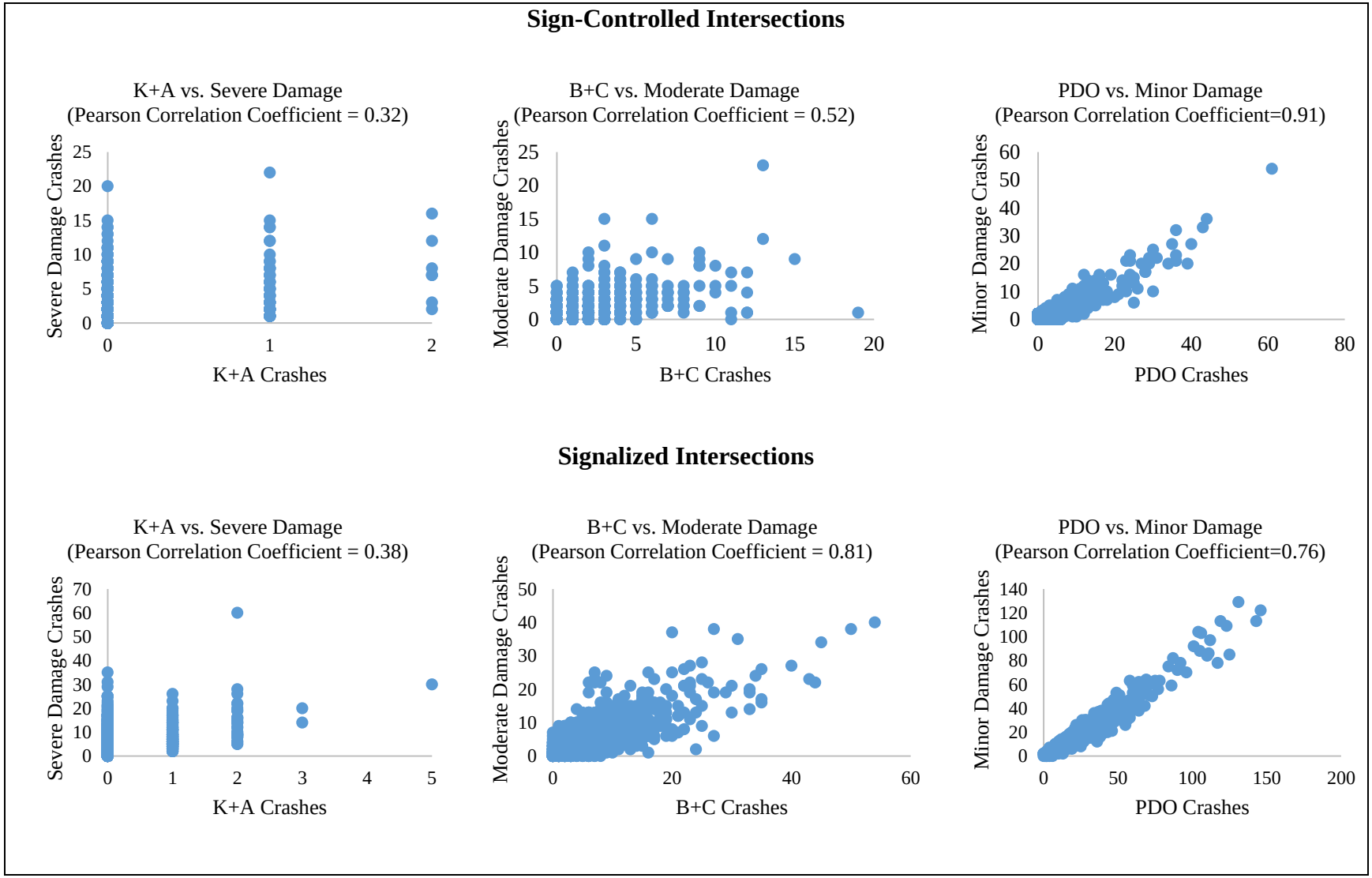


Figure 1 Correlation between Injury Severity and Vehicle Damage

4 MODEL ESTIMATION RESULTS

Table 2 and table 3 show the estimation results for the MVPLN and Joint NB-GOPFS models by different injury severity and vehicle damage levels for urban & suburban sign-controlled and signalized intersections, respectively. In each cell, the first value represents the estimated coefficient, followed by the *p*-value of the coefficient in parenthesis. "---" represents the coefficient is not statistically significant at the 10% significance level, and the results only included variables that are significant at least in one of the models. "NA" represents the variable is not applicable in the specific model.

4.1 Urban & Suburban Sign-Controlled Intersections

Table 2 presents the model estimation results for urban & suburban sign-controlled intersections. The upper part shows the estimated coefficients of the crash prediction models by injury severity level, and the lower part shows the estimated coefficients of the crash prediction models by vehicle damage level.

4.1.1 Model Estimation for Injury Severity Component

With regard to the MVPLN model by injury severity, the crash counts by three different severity levels (*i.e*. K+A, B+C and PDO crashes) are simultaneously estimated using a Poisson-Lognormal framework by accounting for their correlations due to the common unobserved factors. Both the major and minor road AADT are found to be statistically significant and are positively associated with all three levels of crash severity counts. Compared with 3-leg intersections, 4-leg intersections are associated with increased crash counts by all severity levels, which may be due to the fact that there are more conflicting points at 4-leg intersections. As expected, all-way stop-controlled intersections have experienced decreased crashes with severe injuries because right-of-way is separated for all approaches, and the vehicle speed is lower as all

vehicles are ordered to stop first and then go at all-way stop-controlled intersections. If a driveway (such as a driveway for gasoline, parking and commercial store *etc*.) is present at the intersection, crash counts by all three severity levels are expected to decrease. This might be because drivers tend to drive more carefully at these intersections where vehicles may exit from the nearby driveway. Exclusive left-turn lanes are associated with decreased crash counts for all three severity levels, and exclusive right-turn lanes are associated with increased crashes with less severe injuries. These findings are consistent with the study conducted by Wang *et al*. (2017) that exclusive left-turn lanes may reduce specific crash types relating to severe consequences such as head-on crashes, while exclusive right-turn lanes may increase some crash types corresponding to less severe injuries such as read-end crashes at sign-controlled intersections. The correlation coefficients from the MVPLN model highlight that the crash counts are highly correlated among all crash severities, which indicates that accounting for their correlations might yield more accurate estimation results when simultaneously estimating crash counts by severity level.

The Joint NB-GOPFS model has two components, where a NB modeling framework is used to estimate the total crash counts, and a GOPFS modeling framework is used to estimate the crash proportions by each severity level. In the NB modeling framework, a positive coefficient indicates a positive correlation between the independent variable and total crash counts, and vice versa. In the GOPFS modeling framework, a positive coefficient represents that the independent variable is associated with increased proportions of severe injury crashes, and vice versa. The coefficient estimates in the NB modeling framework are consistent with the MVPLN model, in which the major and minor road AADT, 4-leg intersections and exclusive right-turn lanes are

associated with increased total crash counts, while all-way controlled intersections and presence of driveways are associated with decreased total crash counts. The coefficient estimates of GOPFS modeling framework illustrate that more traffics in major road and 4-leg intersections are highly correlated with increased proportions of severe injury crashes, and exclusive left-turn lanes are significantly associated with decreased proportions of severe injury crashes. The threshold parameters in the Joint NB-GOPFS model indicate the demarcation points between severity categories which have no substantial interpretation (Yasmin and Eluru, 2018). One important finding in the Joint NB- GOPFS model is that the total crash counts and the threshold between the proportions of B+C and K+A crashes are positively correlated. This finding implies that sites with higher number of total crashes are more likely to incur higher proportions of B+C crashes (as the threshold will move rightward in the generalized ordered Probit fractional split framework, and the thresholds for B+C and K+A are used to define the crash proportions of B+C crashes), and their correlation is found to be constant across different intersections. This verifies the presence of common unobserved factors affecting both total crash counts and the proportions of crashes by severity and accounting for the unobserved factors when simultaneously estimating total crashes and crash severity proportions may provide more accurate estimation results.

4.1.2 Model Estimation for Vehicle Damage Component

As mentioned earlier, we also estimated crash prediction models by vehicle damage level to supplement the injury severity, which can be used as an alternative to identify locations that may experience severe injury crashes in the future when the current sample mean of severe injury crashes is very low which leads to the difficulty of developing crash prediction models by injury severity. As shown in the results, the MVPLN model coefficient estimates regarding the vehicle damage component are highly consistent with the injury severity component. The correlation

coefficients show that the crash counts by different vehicle damage levels are significantly correlated.

Similarly, the coefficient estimates for the Joint NB-GOPFS model are consistent with those for the injury severity component. The correlation coefficients in the Joint NB-GOPFS demonstrate that the total crash counts and the proportions of crashes by vehicle damage are positively correlated, which implies that sites with a higher number of crashes are more likely to incur higher proportions of severe vehicle damage crashes. The consistent model estimation results between injury severity and vehicle damage components provide support to our initial hypothesis of using vehicle damage as a supplemental indicator of injury severity for estimating crash prediction models by different severity levels.

TABLE 2 Model Estimation Results for Sign-Controlled Intersections

Variables	Injury Severity Component				
	MVPLN Model			Joint NB-GOPFS Model	
	K+A	B+C	PDO	Total Crashes	Severity Proportions
Constant	-17.11 (0.00)	-17.59 (0.00)	-18.24 (0.00)	-9.84 (0.00)	NA
Ln (Major AADT)	0.55 (0.03)	0.68 (0.00)	0.82 (0.00)	0.80 (0.00)	0.10 (0.10)
Ln (Minor AADT)	0.24 (0.05)	0.54 (0.00)	0.60 (0.00)	0.56 (0.00)	---
4-Leg Intersection	0.46 (0.08)	0.77 (0.00)	0.46 (0.00)	0.64 (0.00)	0.11 (0.09)
All-Way Sign-Controlled	-1.21 (0.06)	-0.55 (0.00)	---	-0.31 (0.00)	---
Driveway Presence	-0.45 (0.07)	-0.18 (0.02)	-0.16 (0.00)	-0.16 (0.01)	---
Exclusive Left-Turn Lane Presence	-1.71 (0.05)	-0.41 (0.02)	-0.19 (0.08)	---	-0.23 (0.04)
Exclusive Right-Turn Lane Presence	---	0.59 (0.00)	0.37 (0.00)	0.35 (0.00)	---
Overdispersion	0.65 (0.01)	0.47 (0.00)	0.33 (0.00)	0.24 (0.05)	NA
Threshold 1	NA	NA	NA	NA	1.10 (0.06)
Threshold 2	NA	NA	NA	NA	0.43 (0.00)
Correlation Coefficients	K+A	B+C	PDO	**Correlation Coefficients**	Total Crashes
K+A	1.00	0.79 (0.00)	0.53 (0.00)	Propensity of proportions of severe injury crashes	---
B+C		1.00	0.74 (0.00)	Threshold between B+C and K+A proportions	0.31 (0.09)
PDO			1.00	Threshold between PDO and B+C proportions	---

Variables	Vehicle Damage Component				
	MVPLN Model			Joint NB-GOPFS Model	
	Severe Damage	Moderate Damage	Minor Damage	Total Crashes	Damage Proportions
Constant	-12.64 (0.00)	-16.69 (0.00)	-15.73 (0.00)	-7.10 (0.00)	NA
Ln (Major AADT)	0.35 (0.00)	0.71 (0.00)	0.66 (0.00)	0.65 (0.00)	---
Ln (Minor AADT)	0.35 (0.00)	0.34 (0.00)	0.41 (0.00)	0.39 (0.00)	---
4-Leg Intersection	0.56 (0.00)	0.36 (0.00)	0.16 (0.02)	0.48 (0.00)	0.14 (0.02)
All-Way Sign-Controlled	-0.33 (0.01)	---	---	-0.18 (0.08)	-0.01 (0.09)
Driveway Presence	-0.18 (0.01)	---	---	-0.09 (0.09)	-0.10 (0.07)
Exclusive Right-Turn Lane Presence	0.27 (0.05)	0.75 (0.00)	0.21 (0.08)	0.38 (0.00)	-0.04 (0.04)
Overdispersion	0.33 (0.00)	0.38 (0.00)	0.34 (0.00)	0.04 (0.00)	NA
Threshold 1	NA	NA	NA	NA	-0.96 (0.06)
Threshold 2	NA	NA	NA	NA	-0.69 (0.00)
Correlation Coefficients	Severe Damage	Moderate Damage	Minor Damage	**Correlation Coefficients**	Total Crashes
Severe Damage	1.00	0.73 (0.00)	0.61 (0.00)	Propensity of proportions of severe vehicle damage crashes	0.47 (0.00)

Moderate Damage		1.00	0.78 (0.00)	Threshold between moderate and severe damage proportions	---
Minor Damage			1.00	Threshold between minor and moderate damage proportions	---

Notes: the first value represents the estimated coefficient, followed by the p-value of the coefficient and the following value in parenthesis; "---" represents the variable is not statistically significant at the 10% significance level; "NA" represents the variable is not applicable in the model.

4.2 Urban & Suburban Signalized Intersections

4.2.1 Model Estimation for Injury Severity Component

Table 3 presents the model estimation results for urban & suburban signalized intersections. In terms of the injury severity component, the estimated coefficients for major and minor road AADT and 4-leg intersections are consistent with the urban & suburban sign-controlled intersections, and are associated with increased crash counts for all three levels of crash severity. Presence of driveway is found to be correlated with increased B and C and PDO crashes at signalized intersections. The exclusive right-turn lanes are found to be negatively associated with all severity counts at signalized intersections. One interesting finding is that the protected left-turn signal phasing is correlated with decreased severe injury crashes (K and A crashes), but is correlated with increased B, C and PDO crashes. This might be because the protected left-turn signal phasing can be effective at reducing the head-on crashes, but it might increase the rear-end crashes when the leading vehicle unexpectedly brakes and collided by the following vehicle when the left-turn signal turns to yellow or red. The presence of no right-turn-on-red at signalized intersections is correlated with the increased PDO crashes only, which may be due to the driver's violation of this type of traffic control. The MVPLN model indicates that the crash counts are highly correlated among all crash severity levels at the urban & suburban signalized intersections.

With respect to the Join NB-GOPFS model, the estimation results for total crashes in the NB modeling component are still consistent with the MVPLN model. Three variables are found to be significant for estimating crash proportions by severity level in the GOPFS modeling component. 4-leg intersections are associated with increased proportions of severe injury crashes. If a

depressed median is present on any of the intersection approaches, the proportions of severe injury crashes are expected to be increased. The exclusive left-turn lanes are associated with decreased proportions of severe injury crashes. Different from the urban & suburban sign-controlled intersections, the estimated correlation coefficients from the Joint NB-GOPFS model indicate that the total crash counts and crash severity proportions are independent at urban & suburban signalized intersections.

4.2.2 Model Estimation for Vehicle Damage Component

The model estimation results for the vehicle damage component yield consistent parameters with the injury severity component, and the crash counts are prone to be correlated among crashes with all vehicle damage levels. For the Joint NB-GOPFS model, higher traffic volumes yield decreased proportions of severe damage crashes, which fits the expectation because vehicle speed tends to be lower when the traffic is heavy. Protected left-turn signal phasing is associated with decreased proportions of severe damage crashes. Same as the injury severity component, the total crashes and crash proportions by each vehicle damage level are found to be independent in the Joint NB-GOPFS model.

TABLE 3 Model Estimation Results for Signalized Intersections

Variables	Injury Severity Component				
	MVPLN Model			Joint NB-GOPFS Model	
	K+A	B+C	PDO	Total Crashes	Severity Proportions
Constant	-22.12 (0.00)	-17.56 (0.00)	-17.37 (0.00)	-9.58(0.00)	NA
Ln (Major AADT)	0.89 (0.00)	0.81 (0.00)	0.89 (0.00)	0.87 (0.00)	---
Ln (Minor AADT)	0.60 (0.00)	0.40 (0.00)	0.43 (0.00)	0.48 (0.00)	---
4-Leg Intersection	0.24 (0.10)	0.48 (0.00)	0.40 (0.00)	0.43 (0.00)	0.09 (0.00)
Intersection Approach Median Presence	---	---	---	---	0.07 (0.03)
Driveway Presence	---	-0.20 (0.00)	-0.17 (0.00)	-0.17 (0.01)	---
Exclusive Left-Turn Lane Presence	---	---	---	---	-0.09 (0.01)
Exclusive Right-Turn Lane Presence	-0.32 (0.05)	-0.17 (0.00)	-0.10 (0.01)	-0.08 (0.05)	---
Protected Left-Turn Signal Phasing Presence	-0.35 (0.07)	0.09 (0.06)	0.12 (0.00)	---	---
No Right-Turn-On-Red	---	---	0.07 (0.01)	---	---
Overdispersion	0.54 (0.00)	0.34 (0.00)	0.23 (0.00)	0.29 (0.00)	NA
Threshold 1	NA	NA	NA	NA	0.55 (0.14)
Threshold 2	NA	NA	NA	NA	0.56 (0.00)
Correlation Coefficients	K+A	B+C	PDO	**Correlation Coefficients**	Total Crashes
K+A	1.00	0.77 (0.00)	0.65 (0.00)	Propensity of proportions of severe injury crashes	---
B+C		1.00	0.83 (0.00)	Threshold between B+C and K+A proportions	---
PDO			1.00	Threshold between PDO and B+C proportions	---

Variables	Vehicle Damage Component				
	MVPLN Model			Joint NB-GOPFS Model	
	Severe Damage	Moderate Damage	Minor Damage	Total Crashes	Damage Proportions
Constant	-15.32 (0.00)	-19.37 (0.00)	-18.98 (0.00)	-9.53 (0.00)	NA
Ln (Major AADT)	0.67 (0.00)	0.95 (0.00)	0.93 (0.00)	0.86 (0.00)	-0.08 (0.03)
Ln (Minor AADT)	0.33 (0.00)	0.44 (0.00)	0.53 (0.00)	0.47 (0.00)	-0.08 (0.00)
4-Leg Intersection	0.48 (0.00)	0.42 (0.00)	0.42 (0.00)	0.43 (0.00)	---
Driveway Presence	-0.16 (0.00)	-0.15 (0.00)	-0.20 (0.00)	-0.17 (0.00)	---
Exclusive Left-Turn Lane Presence	---	-0.11 (0.05)	---	---	---
Exclusive Right-Turn Lane Presence	-0.22 (0.00)	-0.15 (0.00)	-0.12 (0.00)	-0.09 (0.03)	---
Protected Left-Turn Signal Phasing Presence	-0.08 (0.09)	0.13 (0.03)	0.14 (0.01)	---	-0.10 (0.00)
Overdispersion	0.22 (0.00)	0.32 (0.00)	0.27 (0.00)	0.28 (0.00)	NA

Threshold 1	NA	NA	NA	NA	-1.41 (0.00)
Threshold 2	NA	NA	NA	NA	-0.68 (0.00)
Correlation Coefficients	Severe Damage	Moderate Damage	Minor Damage	**Correlation Coefficients**	Total Crashes
Severe Damage	1.00	0.69 (0.00)	0.67 (0.00)	Propensity of proportions of severe vehicle damage crashes	---
Moderate Damage		1.00	0.83 (0.00)	Threshold between moderate and severe damage proportions	---
Minor Damage			1.00	Threshold between minor and moderate damage proportions	---

5 MODEL COMPARISONS

In order to evaluate the model prediction capability between the MVPLN and Joint NB-GOPFS models, we randomly selected 80% of the datasets (*i.e*. estimation datasets) to estimate the model coefficients, and used the remaining 20% datasets (*i.e*. validation datasets) to evaluate the model prediction accuracy, based on the criteria of Mean Absolute Error (MAE) calculated as:

$$MAE = \sum_{i=i}^{N} \frac{|Y_{i,pred.} - Y_{i,obs.}|}{N} \quad (19)$$

where $Y_{i,pred.}$ represents the predicted crash counts for intersection *i* corresponding to the specific injury severity or vehicle damage; $Y_{i,obs.}$ represents the observed crash counts for intersection *i* corresponding to the specific injury severity or vehicle damage; and *N* represents the sample size.

A smaller *MAE* value indicates a better prediction accuracy. The *MAE* is calculated for both model estimation (*EMAE*) and validation (*VMAE*) datasets. Figure 2 and Figure 3 present the model prediction comparison results. In general, the MVPLN model performs slightly better in predicting severe crashes, while the Joint NB-GOPFS model performs better in predicting less severe crashes. Specifically, in terms of the crash prediction by injury severity for both sign-controlled and signalized intersections, the MVPLN model slightly outperforms the Joint NB-GOPFS model in predicting K and A crashes, while the Joint NB-GOPFS model performs better in predicting B, C and PDO crashes. The Joint NB-GOPFS model has a smaller prediction error than the MVPLN model based on the average *MAE* value across all severity levels. With regard to the crash prediction by vehicle damage, the MVPLN model performs slightly better than the GOPFS model in predicting severe and minor damage crashes and the average crashes across all

damage levels for sign-controlled intersections, but it only outperforms the GOPFS model in predicting severe damage crashes for signalized intersection.

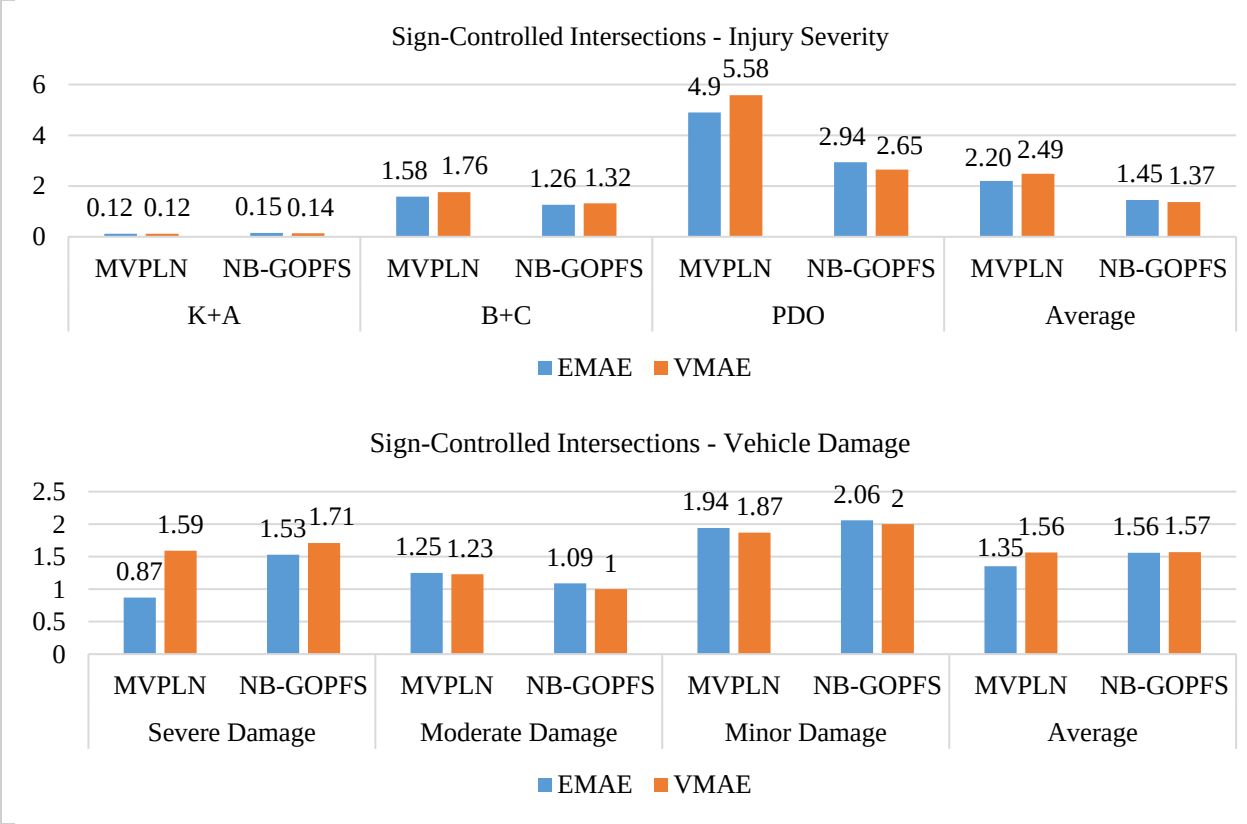
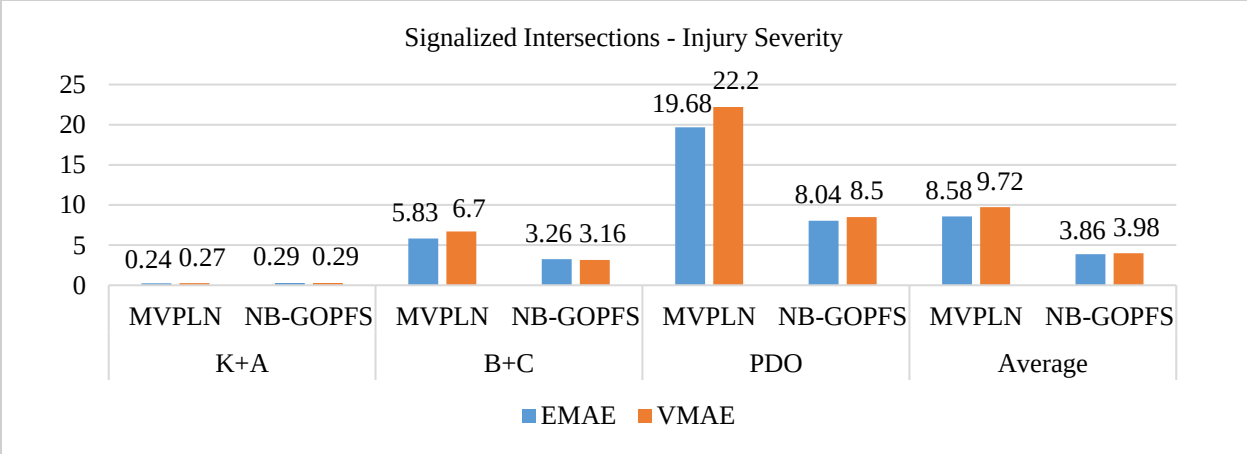


Figure 2 Model Performance Comparisons for Sign-Controlled Intersections



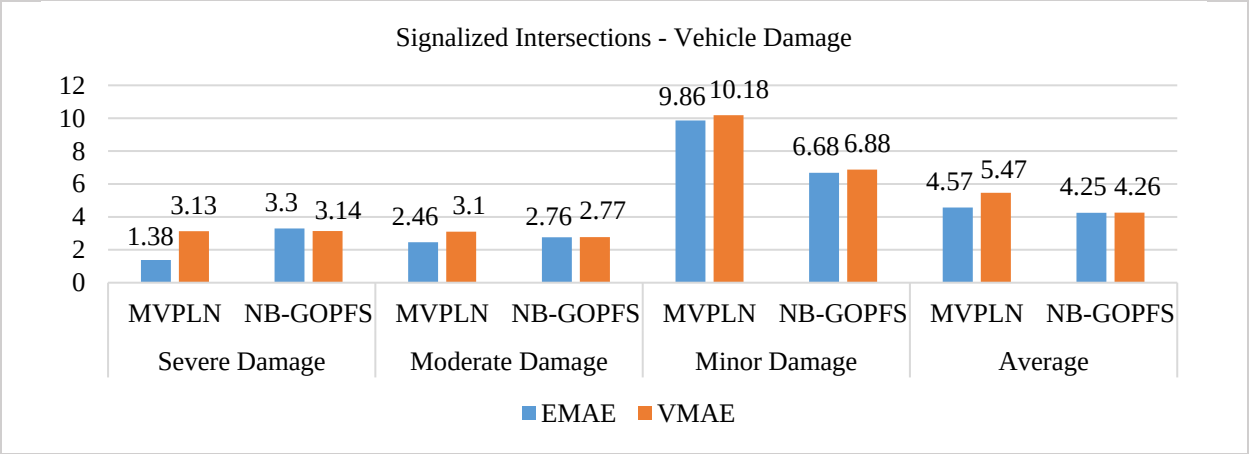


Figure 3 Model Performance Comparisons for Signalized Intersections

6 SUMMARY AND CONCLUSIONS

This paper presents two advanced frameworks in predicting crash counts by each severity level, *i.e*. either directly estimating crash counts by different severity levels, or first estimating crash counts in total, and then estimating crash severity distributions to combine with the total crashes for crash count prediction by severity. Two advanced methodologies are implemented with regard to each of the framework. In terms of the first framework, a MVPLN model is used to simultaneously estimate crash counts by different severity levels, by accounting for their correlations due to the common unobserved factors. In the second framework, a Joint NB-GOPFS model is applied in which a NB modeling component is used to estimate the total crash counts and a GOPFS modeling component is used to estimate the crash proportions by each severity level. The NB and GOPFS modeling components are jointly estimated by accounting for the correlations between total crashes and crash severity proportions due to the common unobserved factors.

Both sign-controlled and signalized intersections at urban & suburban areas are collected from the State of Connecticut and used for model estimation. The estimated coefficients in the

MVPLN model show that crash counts are highly correlated among all severity levels for both sign-controlled and signalized intersections, which indicates that accounting for their correlations might yield more accurate estimation results when simultaneously estimating crash counts by severity. The estimation results of the Joint NB-GOPFS model show that the total crashes are significantly correlated with the proportions of B and C crashes at sign-controlled intersections, and their correlations should be accommodated when simultaneously estimating total crash counts and crash proportions by severity. The total crash counts are found to be independent with the crash proportions by severity at signalized intersections.

In addition, we further estimated crash prediction models by vehicle damage level to supplement the injury severity, which can be used as an alternative to identify locations that may experience severe injury crashes in the future when the current sample mean of severe injury crashes (such as K and A crashes) is very low which leads to the difficulty of developing crash prediction models by injury severity. The model estimation results for injury severity component and vehicle damage component are highly consistent. This finding verifies our initial assumption that when crash data samples have challenges associated with the low observed sampling rates for severe injury crashes, vehicle damage can be appropriate as an alternative to injury severity in crash prediction by severity. An important finding from the model estimation is that two methodologies may yield different variables that are statistically significant in predicting crashes by severity level. For example, the traffic volumes are shown to be significant in all MVPLN models when crash counts by severity are simultaneously modeling, while the traffic volumes seldom affect the prediction of crash proportions by each severity level in the Joint NB-GOPFS model. This may provide additional insight about variable selection in crash prediction models

by severity level regarding different approaches. In the end, the prediction performance of the two approaches is compared based on the *MAE* values. The comparisons show that the MVPLN slightly outperforms the Joint NB-GOPFS model in terms of predicting severe crashes, while the Joint NB-GOPFS model significantly improves the prediction accuracy of less severe crashes compared to the MVPLN model. This finding contributes to the practical applications of both crash prediction research and safety improvement effort through shedding light on method selection under different data conditions and research needs, in which the MVPLN is recommended when the analysis target is severe crashes while the NB-GOPFS is preferred for less severe crashes.

7 PRACTICAL APPLICATIONS AND FUTURE WORK

The findings of this research can offer additional insight into selecting robust methodological modeling frameworks in estimating crash counts by different severity levels, and provide researchers and practitioners with the capabilities of estimating crash prediction models when the low sample mean leads to difficulties in predicting severe crashes. In this study, we used the intersection data to test the proposed modeling frameworks. Future research can focus on extending the modeling frameworks to roadway segments. Future research can also target on extending the MVPLN model to the generalized MVPLN framework, and further extending the Joint NB-GOPFS modeling framework by accounting for the temporal and spatial heterogeneity.

8 ACKNOWLEDGMENTS

The authors would like to thank the Connecticut Department of Transportation (CTDOT) and the Connecticut Crash Data Repository (CTCDR) for providing the intersection and crash data.

9 DECLARATION OF INTEREST

None.

REFERENCES

Abdel-Aty, M and A. Radwan. Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention 32*. No. 5. 2000, pp. 633-642.

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

Anarkooli, A., B. Persaud, M. Hosseinpour and T. Saleem. Comparison of univariate and two-stage approaches for estimating crash frequency by severity - case study for horizontal curves on two-lane rural roads. *Accident Analysis and Prevention 129*. 2019, pp. 382-389.

Anastasopoulos, P., V. Shankar, J. Haddock and F. Mannering. A Multivariate tobit analysis of highway accident-injury-severity rates. *Accident Analysis and Prevention 45*. 2012, pp. 110-119.

Aptech. Aptech Systems Inc. 2019. https://www.aptech.com/

Barua, S., K. El-Basyouny and M. Islam. A full bayesian multivariate count data model of collision severity with spatial correlation. *Analytic Methods in Accident Research 3-4*. 2014, pp. 28-43.

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

Bhat, C. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological 35*. No. 7. 2001, pp. 677-693.

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

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

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

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

Connecticut Crash Data Repository. *Connecticut Transportation Safety Research Center*. 2019. http://www.ctcrash.uconn.edu/

Dixon, K., C. Monsere, R. Avelar, J. Barnett, P. Escobar, S. Kothuri and Y. Wang. Improved safety performance functions for signalized intersections. *Oregon Department of Transportation*. 2015.

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

Eluru, N., C.R. Bhat and D.A. Hensher. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis and Prevention 40*, No.3. 2008, pp. 1033-1054.

Geedipally, S., J. Bonneson, M. Pratt and D. Lord. Severity distribution functions for freeway segments. *Transportation Research Record 2398*. 2013, pp. 19-27.

Highway Safety Manual (HSM) 1st Edition. *American Association of State Highway and Transportation Officials*. Washington D.C. 2010.

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

Lord, D and B. Persaud. Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Transportation Research Record 1717*. 2000, pp. 102-108.

Lord, D. and F. Mannering. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice 44*. No. 5. 2010, pp. 291-305.

Ma, J. and K. Kockelman. Bayesian multivariate poisson-lognormal regression for models of injury count by severity. *Transportation Research Record 1950*. 2006, pp. 24-34.

Ma, J., K. Kockelman and P. Damien. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. *Accident Analysis and Prevention*. Vol. 40, No. 3. 2008, pp. 964-975.

Mannering, F. and C. Bhat. Analytic methods in accident research: methodological frontier and future directions. *Analytic Methods in Accident Research 1*. 2014, pp. 1-22.

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

Model Minimum Uniform Crash Criteria, Fifth Edition. *U.S. Department of Transportation*, Washington, DC. 2017. https://www.nhtsa.gov/mmucc

National Highway Traffic Safety Administration (NHTSA). Traffic Safety Facts: 2016 Data. 2018. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812580

Oh, J., S. Washington and K. Choi. Development of accident prediction models for rural highway intersections. *Transportation Research Record 1897*. 2004, pp. 18-27.

Park, E. S. and D. Lord. Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record 2019*. 2007, pp. 1-6.

Pei, X., S. Wong and N. Sze. A joint-probability approach to crash prediction models. *Accident Analysis and Prevention 43*. No. 3, 2011. pp. 1160–1166.

Qin, X., A. Sultana, M. Chitturi and D. Noyce. Developing Truck Corridor Crash Severity Index. *Transportation Research Record 2386*. 2013, pp. 103-111.

Qin, X., K. Wang and C. Cutler. Analysis of crash severity based on vehicle damage and occupant injuries. *Transportation Research Record 2386*. 2013, pp. 95-102.

R-INLA. Bayesian Computing with INLA. http://www.r-inla.org/home.

Rue, H., S. Martino and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society 71*. Series B. 2009, pp. 319-392.

Russo, F., M. Busiello A. and Dell. Safety performance functions for crash severity on undivided rural roads. *Accident Analysis & Prevention 93*. 2016, pp. 75-91.

Savolainen, P., F. Mannering, D. Lord and M. Quddus. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis and Prevention 43*. No. 5. 2011, pp. 1666-1676.

Serhiyenko, V., S. Mamun, J. Ivan and N. Ravishanker. Fast bayesian inference for modeling multivariate crash counts. *Analytic Methods in Accident Research 10*, 2016, pp. 44-53.

Tarko, A. P., M. Inerowicz, J. Ramos and W. Li. Tool with road-level crash prediction for transportation safety planning. *Transportation Research Record 2083*. 2008, pp. 16-25.

Ulfarsson, G and V. Shankar. Accident count model based on multiyear cross-sectional roadway data with serial correlation. *Transportation Research Record 1840*. 2002, pp. 193-197.

Wang, C., M. Quddus and S. Ison. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention 43*. No. 6. 2011, pp. 1979-1990.

Wang, K., J. Ivan, A. Burnicki and S. Mamun. Predicting local road crashes using socioeconomic and land cover data. *Journal of Transportation Safety & Security*. Vol. 9, No. 3, 2017, pp. 301-318.

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

Wang, K., S. Zhao and E. Jackson. Multivariate poisson lognormal modeling of weather-related crashes on freeways. *Transportation Research Record 2672*. No. 38. 2018, pp. 184-198.

Wang, K. S. Zhao and E. Jackson. Functional forms of the negative binomial models in safety performance functions for rural two-lane intersections. *Accident Analysis and Prevention*. Vol. 124, 2019, pp. 193-201.

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

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

Washington, S., M. Karlaftis and F. Mannering. Statistical and econometric methods for transportation data analysis, 2nd ed. *Chapman and Hall/CRC*, Boca Raton, FL. 2011.

Wang, Y. and K. Kockelman. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis and Prevention 60*. 2013, pp. 71-81.

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

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

Yasmin, S. and N. Eluru. A joint economic framework for modeling crash counts by severity. *Transportmetrica A: Transport Science 14*. No 3. 2018. pp. 230-255.

Yasmin, S and N. Eluru. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis and Prevention 59*. No. 1. 2013, pp. 506-521.

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

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