

A Mixed Grouped Response Ordered Logit Count Model Framework

Shamsunnahar Yasmin

Postdoctoral Associate

Department of Civil, Environmental & Construction Engineering

University of Central Florida

Tel: 407-823-4815, Fax: 407-823-3315

Email: shamsunnahar.yasmin@ucf.edu

Naveen Eluru*

Associate Professor

Department of Civil, Environmental & Construction Engineering

University of Central Florida

Tel: 407-823-4815, Fax: 407-823-3315

Email: naveen.eluru@ucf.edu

*Corresponding author

Abstract

The study proposes and estimates a new econometric framework for analysing crash count events labeled as the Mixed Grouped Response Ordered Logit Count model. The proposed framework relates the crash count propensity to the observed counts directly while also accommodating for heteroscedasticity and unobserved heterogeneity. The proposed model is demonstrated by using Traffic Analysis Zone level bicycle crash count data for the Island of Montreal. The model framework employs a comprehensive set of exogenous variables – accessibility measures, exposure measures, built environment, road network characteristics, sociodemographic and socioeconomic characteristics. Further, we also compare the performance of the proposed model to the most commonly used negative binomial model and the generalized ordered logit count model by generating a comprehensive set of measures to evaluate model performance and data fit. The alternative modeling approaches considered for the comparison exercise include: (1) negative binomial model without parameterized overdispersion, (2) negative binomial model with parameterized overdispersion, and (3) mixed negative binomial model with parameterized overdispersion, (4) generalized ordered logit count model and (5) mixed generalized ordered logit count model, (6) grouped response ordered logit count model without parameterized variance, (7) grouped response ordered logit count model with parameterized variance and (8) mixed grouped response ordered logit count model with parameterized variance. The comparison exercise clearly highlights that the proposed mixed grouped response ordered logit count model with parameterized variance relative to the mixed negative binomial model with parameterized overdispersion offers either equivalent or superior data fit across various measures in the current study context. The fit measures for comparing the predictive performance also indicate that the proposed grouped response model offers better predictions both at the aggregate and disaggregate levels. Overall, the results from this comparison exercise points out that the grouped response ordered logit count model is a promising alternate econometric framework for examining crash count events.

Key words: Count modeling; grouped ordered response; crash frequency; heteroscedasticity; unobserved heterogeneity.

1 BACKGROUND

Road traffic crashes occur due to a complex interaction among a multitude of factors. An important tool for identifying these factors is the application of econometric/statistical models. The main objective of employing these models is to reveal the quantitative relationship between the dependent variable of interest and the various covariates considered. Traditionally, econometric/statistical analysis in road safety research has evolved along two major streams: crash frequency analysis (see Lord and Mannering, 2010; Yasmin and Eluru, 2016 for a review of crash frequency studies) and crash severity analysis (see Savolainen et al., 2011; Yasmin and Eluru, 2013 for a detailed review of crash severity studies). The proposed research effort contributes to literature on crash frequency analysis. Specifically, we propose and estimate an econometric framework – Mixed Grouped Response Ordered Logit Count model - for analysing crash count events.

For crash frequency analysis, researchers have employed a wide range of models (such as: linear regression, ordered logit, ordered probit, Poisson, negative binomial and variants of these different frameworks) in addressing several methodological and empirical issues associated with count data. However, the application of traditional Poisson and negative binomial regression models remains predominant in examining crash count events in safety literature. With respect to methodological issues, more recently, Bhat and colleagues (Castro et al., 2012; Castro et al., 2013; Narayanamoorthy et al., 2013; Bhat et al., 2014) addressed the drawbacks of these traditional count models in examining multivariate count data. They argued that the estimation of correlated count data by using either Poisson or negative binomial model is likely to be cumbersome and restrictive in the consideration of the impact of unobserved factors. To alleviate the restrictive nature of these frameworks, Bhat and colleagues have proposed and implemented a generalized ordered logit count model framework which addresses many challenges of traditional count models for correlated count data.

The current research effort contributes to the safety literature methodologically and empirically by developing an alternative structure to the work by Bhat and colleagues. In our study, we propose the grouped response ordered logit count model that obviates the need to estimate thresholds by relating the propensity directly to the observed counts. This relaxation ensures that the number of model parameters are not dependent on the number of count categories. Additionally, the proposed model allows for parameterization of variance of the unobserved component (restricted to a variance of 1 in the generalized ordered logit count framework). The model fit of the proposed framework is compared with the model fit of negative binomial and generalized ordered logit count models and their variants. Furthermore, given the recent recognition of the criticality of unobserved heterogeneity in safety models (Mannering et al., 2016), we incorporate unobserved heterogeneity in these three model frameworks. To elaborate, we estimate a mixed negative binomial model¹, mixed generalized ordered logit count model and mixed grouped response ordered logit count model.

The comparison exercise is augmented by evaluating predictive performance measures of the estimated models. It is worthwhile to mention here that the proposed mixed grouped response ordered logit count econometric framework is generic and applicable for examining count events (or grouped ordered variables) for any domain. In current study context, the application of mixed grouped response ordered logit count model is demonstrated by using zonal level bicycle-motor

¹ The mixed negative binomial model is also often referred to as the random components negative binomial model.

vehicle crash count data from the Island of Montreal, Canada, employing a comprehensive set of exogenous variables.

1.1 Earlier Work and Current Study Context

Traditionally, crash frequency models are developed based on non-integer valued events aggregated at a certain planning level (such as census tracts, traffic analysis zones) for a given time interval. Researchers have employed a wide array of econometric approaches including linear regression, count regression and discrete outcome models in quantifying the impact of exogenous factors on count events. Several studies have employed ordinary least squares estimation method in developing crash count models (see Wier et al., 2009 for such modeling approach). Using ordinary least squares regression to analyze crash count events that are usually discrete non-negative integers and non-normal in their distribution may result in biased estimates. The traditional Poisson model is used for examining count events as it can accommodate for the integer properties of count data directly (Hausman et al., 1994). However, one of the restrictive assumptions of Poisson model is the equivalence of the mean and variance. In crash data, it is common to observe that variance of the data often exceeds the mean (Hauer et al., 2001). Towards this end, the negative binomial model that accounts for such overdispersion has frequently been used in safety literature (Noland and Quddus, 2004). To be sure, several other methodological approaches have been employed for examining crash count events including generalized Poisson regression, negative multinomial regression, random effect negative binomial, geographically weighted Poisson regression, geographically weighted negative binomial, Bayesian Poisson lognormal, quasi induced exposure method, Bayesian spatial regression model, zero-inflated and hurdle models (see Yasmin and Eluru, 2016 and Lord and Mannering, 2010 for a detailed review). However, the negative binomial model and its variants (such as mixed negative binomial model) still serve as the workhorse for crash frequency models.

Bhat and colleagues (Castro et al., 2012; Castro et al., 2013; Narayanamoorthy et al., 2014; Bhat et al., 2014) have recently proposed and implemented a generalized ordered logit count model. In these studies, the authors show that the generalized ordered logit count model employed with certain assumptions can replicate Poisson model. In addition to the equivalency, the generalized ordered logit count model approach is beneficial for multiple reasons. First, the generalized ordered logit count approach develops an elegant mathematical structure to accommodate for count specific parameter impacts. In a traditional count model (Poisson or negative binomial), the impact of exogenous variables is restricted to be the same across all count values. The generalized ordered logit count model offers additional flexibility by allowing for the estimation of count specific effects through the parametrization in thresholds. Second, the error structure in the generalized ordered logit count model variant is part of the propensity of a linear in parameter specification system (as opposed to within a non-linear specification in Poisson or negative binomial models). The parameter specification structure lends itself to flexible multivariate modeling approaches as demonstrated by Bhat and colleagues in multiple efforts. However, the parameterization of the thresholds in the form of a Poisson process increases the computational complexity involved in the model estimation and identification relative to the traditional negative binomial model. Given the additional flexibility offered by generalized ordered logit count model the additional complexity is warranted.

In the current study, we examine the count events by using grouped ordered response modeling approach that retains the benefits of the generalized ordered logit count model.

Specifically, we formulate and estimate grouped response ordered logit count model for examining count events. The grouped response structure allows for flexible specification of the dependent variable while also not being restricted by additional threshold parameters to be estimated. Earlier research efforts including Bhat (1994) and Stewart (1983) have considered observed thresholds within their formulation. In our study, we extend these approaches by allowing for alternative specific effects and unobserved heterogeneity. For examining count events, as opposed to relating the latent propensity to estimable thresholds, the grouped response ordered logit count models relates the propensity to observed counts directly. It obviates the need to estimate thresholds and reduces the model parameter estimation burden. Additionally, the influence of exogenous variables is allowed to vary across alternatives adopting a simple linear in parameter specification (as opposed to the complex non-linear specification in generalized ordered logit count model). Further, to relate the propensity to the observed counts, we also need to relax the standard logistic assumption for the error term and estimate the variance of the error term. As we accommodate for exogenous variables in the variance component, the model will automatically accommodate for heteroscedasticity. This is analogous to estimating the influence of exogenous variables in the overdispersion component of the negative binomial model. Further, due to the presence of unobserved information, the effect of exogenous variables on crash count events might not be the same across all study units. In non-linear models, neglecting the effect of such unobserved heterogeneity can result in inconsistent estimates (Chamberlain, 1980; Bhat, 2001; Mannering et al., 2016). To that extent, we also incorporate the influence of unobserved heterogeneity in the proposed framework by formulating and estimating the mixed grouped response ordered logit count model. The proposed grouped ordered count approach is the first application of this model for count event analysis.

In the current study context, the proposed model is estimated using traffic analysis zone level crash count data for the Island of Montreal employing a comprehensive set of exogenous variables – accessibility measures, exposure measures, built environment, road network characteristics, sociodemographic and socioeconomic characteristics. In order to demonstrate the applicability and performance of the proposed framework, we also empirically compare the performance of grouped response ordered logit count model with negative binomial and generalized ordered logit count model frameworks (and their variants) in the context of bicycle-motor vehicle crash count events. Thus, in the current study approach, we estimate crash count models based on three different frameworks – negative binomial, generalized ordered logit count and grouped response ordered logit count frameworks². The entire set of models considered in our analysis include: (1) negative binomial model without parameterized overdispersion, (2) negative binomial model with parameterized overdispersion, and (3) mixed negative binomial model with parameterized overdispersion, (4) generalized ordered logit count model and (5) mixed generalized ordered logit count model, (6) grouped response ordered logit count model without parameterized variance, (7) grouped response ordered logit count model with parameterized variance and (8) mixed grouped response ordered logit count model with parameterized variance. We also generate a comprehensive set of measures to evaluate model performance and data fit across these frameworks. For simplicity, in the following sections, we refer to generalized ordered logit count and grouped response ordered logit count models as generalized ordered logit and grouped ordered logit models, respectively.

² Recently, Paleti et al. (2016) proposed a Generalized Extreme Value variant of a count model that subsumes the NB model. However, the framework is based on an unordered model structure and inherently ignores the ordering in the count variable. Hence, we have not considered this model structure for our comparison exercise.

The rest of the paper is organized as follows. Section 2 provides details of the econometric model frameworks used in the analysis. In Section 3, the data is described. Model comparison results and estimation results are presented in Section 4. Section 5 concludes the paper.

2 ECONOMETRIC FRAMEWORK

The focus of our study is to examine bicycle-motor vehicle crash count events at a zonal level based on three different frameworks: negative binomial, generalized ordered logit and grouped ordered logit frameworks. In this section, econometric formulations for these models are presented. Within negative binomial framework, we can obtain negative binomial without and with parameterized overdispersion models by imposing one or more restrictions on mixed negative binomial model with parameterized overdispersion. Thus, negative binomial without and with parameterized overdispersion models are the nested versions of mixed negative binomial with parameterized overdispersion model. Similarly, generalized ordered logit is the nested version of mixed generalized ordered logit model. Finally, with respect to grouped ordered logit framework, mixed grouped ordered logit model with parameterized variance is the full model while grouped ordered logit model without and with parameterized variance are the nested versions. Thus, we can argue that mixed negative binomial with parameterized overdispersion, mixed generalized ordered logit and mixed grouped ordered logit model with parameterized variance are the most generalized version of models within negative binomial, generalized ordered logit and grouped ordered logit frameworks, respectively. For the ease of presentation, we describe these generalized mathematical structures from each framework and identify the restrictions necessary to arrive at the other variants. The estimation routines for the models in current study are coded in GAUSS Matrix Programming software (Aptech, 2015).

Let us assume that i ($i = 1, 2, 3, \dots, N$) be the index for traffic analysis zone and j ($j = 1, 2, 3, \dots, J$) be the index to represent crashes occurring over a period of time in a traffic analysis zone i . Using the above notational indices, in this section, we provide the econometric frameworks for different models considered for examining bicycle crash count events in our research.

2.1 Mixed Negative Binomial Model with Parameterized Overdispersion

The equation system for modeling total number of crash counts in the usual negative binomial formulation can be written as:

$$P_i(j) = \frac{\Gamma\left(j + \frac{1}{\alpha_i}\right)}{\Gamma(j+1)\Gamma\left(\frac{1}{\alpha_i}\right)} \left(\frac{1}{1 + \alpha_i\mu_i}\right)^{\frac{1}{\alpha_i}} \left(1 - \frac{1}{1 + \alpha_i\mu_i}\right)^j \quad (1)$$

where, $P_i(j)$ is the probability that traffic analysis zone i has j number of crashes. $\Gamma(\cdot)$ is the gamma function, α_i is negative binomial overdispersion parameter and μ_i is the expected number of crashes occurring in traffic analysis zone i over a given time period. In equation 1, we can express μ_i as a function of explanatory variables by using a log-link function:

$$\mu_i = E(j_i | \mathbf{z}_i) = \exp((\boldsymbol{\delta} + \boldsymbol{\zeta}_i) \mathbf{z}_i + \varepsilon_i) \quad (2)$$

where, $\mathbf{z}_i$ is a vector of explanatory variables associated with traffic analysis zone i . $\boldsymbol{\delta}$ is a vector of coefficients to be estimated. $\boldsymbol{\zeta}_i$ is a vector of unobserved factors affecting crash count propensity for traffic analysis zone i and its associated zonal characteristics assumed to be a realization from standard normal distribution: $\boldsymbol{\zeta}_i^2 \sim N(0, \boldsymbol{\pi}^2)$. ε_i is a gamma distributed error term with mean 1 and variance α_i .

In current study the overdispersion parameter is parameterized as follows:

$$\alpha_i = \exp(\ell + \boldsymbol{\gamma} \mathbf{s}_i) \quad (3)$$

where, ℓ is a constant, $\mathbf{s}_i$ is a set of exogenous variables associated with zone i and $\boldsymbol{\gamma}$ is the corresponding vector of parameters to be estimated. The parameterization allows for the overdispersion to be different across zones. Thus, conditional on $\boldsymbol{\zeta}_i$, the likelihood function (NBL_i) for the mixed negative binomial model with parameterized overdispersion can be expressed as:

$$NBL_i = \int_{\boldsymbol{\zeta}_i} P_i(j | \boldsymbol{\zeta}_i) g(\boldsymbol{\zeta}_i) d\boldsymbol{\zeta}_i \quad (4)$$

where $g(\boldsymbol{\zeta}_i)$ is the probability density function of $\boldsymbol{\zeta}_i$. Finally, the unconditional log-likelihood function ($NBLL$) takes the following form:

$$NBLL = \sum_i \ln(NBL_i) \quad (5)$$

The parameters to be estimated in mixed negative binomial with parameterized overdispersion model are: $\boldsymbol{\delta}$, $\boldsymbol{\pi}$, ℓ and $\boldsymbol{\gamma}$. To estimate the proposed mixed negative binomial with parameterized overdispersion model, we apply Quasi-Monte Carlo simulation techniques based on the scrambled Halton sequences for the maximum simulated likelihood approach (see Bhat, 2001; Eluru et al., 2008; Yasmin and Eluru, 2013 for examples of Quasi-Monte Carlo approaches in literature). The negative binomial model without and with parameterized overdispersion models are nested versions of mixed negative binomial with parameterized overdispersion model where the density function $g(\boldsymbol{\zeta}_i)$ degenerates to 1 with $\boldsymbol{\pi} = 0$. Hence, the parameters for negative binomial model without and with parameterized overdispersion are estimated using maximum likelihood approaches without numerical integration of the negative binomial function in equation 4. The parameters to be estimated in these models are: $\boldsymbol{\delta}$, ℓ and $\boldsymbol{\gamma}$ for negative binomial model with parameterized overdispersion; and $\boldsymbol{\delta}$ and ℓ for negative binomial model without parameterized overdispersion ($\boldsymbol{\gamma} = \mathbf{0}$).

2.2 Mixed Generalized Ordered Logit Count Model

The mixed generalized ordered logit model assumes that count events are ordered in nature. In the mixed generalized ordered logit model, similar to traditional ordered outcome model, the discrete crash count events (c_i) are assumed to be mapping an underlying continuous latent variable (c_i^*) as follows:

$$c_i^* = ((\boldsymbol{\beta} + \boldsymbol{\vartheta}_i)\mathbf{x}_i + \xi_i), c_i = j \text{ if } \psi_{i,j-1} < c_i^* < \psi_{i,j} \quad (6)$$

where, $\mathbf{x}_i$ is a vector of attributes (not including a constant) that influences the propensity associated with traffic analysis zone i . $\boldsymbol{\beta}$ is the corresponding vector of mean effects. $\boldsymbol{\vartheta}_i$ is a vector of unobserved factors affecting crash count propensity for traffic analysis zone i and its associated zonal characteristics assumed to be a realization from standard normal distribution: $\boldsymbol{\vartheta}_i^2 \sim N(0, \sigma^2)$. ξ_i is an idiosyncratic error term assumed to be identically and independently standard logistic distributed across zone i . The latent propensity c_i^* is mapped to the actual crash count events c_i by ψ thresholds ($\psi_0 = -\infty$ and $\psi_J = \infty$) as presented in equation 6. In current empirical context, we set $J = 12$, in accommodating long right tail distribution of crash count events considered (following Castro et al., 2012)³.

To maintain ordering condition and to allow the thresholds to vary across zones, the thresholds $\psi_{i,j}$ are parameterized as follows (following Castro et al., 2012):

$$\psi_{i,j} = \phi^{-1} \left(e^{\boldsymbol{\vartheta}_i \sum_{j=0}^J \frac{\boldsymbol{\vartheta}_i^j}{j!}} \right) + \boldsymbol{\kappa}_j, \text{ where } \boldsymbol{\vartheta}_i = e^{\boldsymbol{\mathcal{U}}_i} \text{ also } \boldsymbol{\kappa}_0 = 0 \quad (7)$$

where, $\boldsymbol{\mathcal{T}}_i$ is a set of exogenous variables and $\boldsymbol{\mathcal{U}}$ is the associated vector of parameters to be estimated (including constant). ϕ^{-1} is the inverse function of the univariate cumulative standard normal. Given the above set up, the probability that zone i is likely to be associated with crash count event j may be written as:

$$R_i(j) = \Lambda(\psi_{i,j} - (\boldsymbol{\beta} + \boldsymbol{\vartheta}_i)\mathbf{x}_i) - \Lambda(\psi_{i,j-1} - (\boldsymbol{\beta} + \boldsymbol{\vartheta}_i)\mathbf{x}_i) \quad (8)$$

where, $\Lambda(\cdot)$ represents the standard logistic distribution function. Thus, conditional on $\boldsymbol{\vartheta}_i$, the likelihood function for the mixed generalized ordered logit model (OL_i) can be expressed as:

$$OL_i = \int_{\boldsymbol{\vartheta}_i} R_i(j|\boldsymbol{\vartheta}_i) \mathbb{C}(\boldsymbol{\vartheta}_i) d\boldsymbol{\vartheta}_i \quad (9)$$

where $\mathbb{C}(\boldsymbol{\vartheta}_i)$ is the probability density function of $\boldsymbol{\vartheta}_i$. Finally, the unconditional log-likelihood function (OLL) takes the following form:

³ We set $J = 12$ based on extensive testing using alternative count values. The number of records with count value >12 amount to 0.36% of the sample.

$$OLL = \sum_i \ln(OL_i) \quad (10)$$

The parameters to be estimated in mixed generalized ordered logit model are: β , σ , $\mathbf{x}_j$ and $\mathcal{U}$. To estimate the mixed generalized ordered logit model, we apply Quasi-Monte Carlo simulation techniques as described earlier. However, generalized ordered logit model is nested version of mixed generalized ordered logit model where the density function $\mathbb{C}(\boldsymbol{\vartheta}_i)$ degenerates to 1 with $\sigma = 0$. Hence, the parameters of generalized ordered logit model are estimated using maximum likelihood approaches without the numerical integration of the function in equation 9. The parameters to be estimated in generalized ordered logit model are: β , $\mathbf{x}_j$ and $\mathcal{U}$.

2.3 Mixed Grouped Response Ordered Logit Count Model with Parameterized Variance

Similar to generalized ordered logit system, the grouped ordered logit model also assumes that the count events are ordered in nature. In grouped ordered logit model, the equation system for modeling crash count events may be written as:

$$y_i^* = (\boldsymbol{\rho} + \mathbf{v}_i)\mathbf{k}_i + \boldsymbol{\eta}_i, \quad y_i = j, \text{ if } \tau_{j-1} < y_i^* < \tau_j \quad (11)$$

where, y_i^* is the latent (continuous) propensity for zone i . The latent propensity y_i^* is mapped to the actual crash count event categories y_i by τ thresholds ($\tau_0 = -\infty$ and $\tau_J = +\infty$) as presented in equation 11. Like generalized ordered logit model, we also set $J = 12$ in accommodating long right tail effect in estimating grouped ordered logit model. τ_j take the values as: $-\infty, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$, and $+\infty$ in current study context. $\mathbf{k}_i$ is a vector of attributes that influence the bicycle crash risk propensity (including the constant) and $\boldsymbol{\rho}$ is the corresponding vector of mean coefficients. $\mathbf{v}_i$ is a vector of unobserved factors influencing crash count propensity for zone i and its associated zonal characteristics assumed to be a realization from standard normal distribution: $\mathbf{v}_i^2 \sim N(0, \boldsymbol{\Omega}^2)$.

In current study the variance vector is parameterized as follows:

$$\lambda_i = \exp(\mathcal{g} + \boldsymbol{\theta}\mathbf{w}_i) \quad (12)$$

where, $\mathcal{g}$ is a constant, $\mathbf{w}_i$ is a set of exogenous variables associated with traffic analysis zone i and $\boldsymbol{\theta}$ is the corresponding vector of parameters to be estimated. The parameterization allows for the variance to be different across traffic analysis zones accommodating for heteroscedasticity. Thus, the probability expression for crash count event is given by:

$$U_i(j) = \varphi\left(\frac{\tau_j - (\boldsymbol{\rho} + \mathbf{v}_i)\mathbf{k}_i - \boldsymbol{\varrho}_j\mathbf{t}_{ij}}{\lambda_i}\right) - \varphi\left(\frac{\tau_{j-1} - (\boldsymbol{\rho} + \mathbf{v}_i)\mathbf{k}_i - \boldsymbol{\varrho}_{j-1}\mathbf{t}_{ij}}{\lambda_i}\right) \quad (13)$$

where, $\varphi(\cdot)$ represents the standard logistic distribution function. Further $\mathbf{t}_{ij}$ is a vector of attributes specific to zone i and threshold j and $\boldsymbol{\varrho}_j$ is the vector of corresponding count-specific coefficients. The $\boldsymbol{\varrho}_j$ elements allow for count specific variable impacts in our formulation.

Thus, conditional on $\mathbf{v}_i$, the likelihood function (GL_i) for the mixed grouped ordered logit model with parameterized variance can be expressed as:

$$GL_i = \int_{\mathbf{v}_i} U_i(j|\mathbf{v}_i) \wp(\mathbf{v}_i) d\mathbf{v}_i \quad (14)$$

where $\wp(\mathbf{v}_i)$ is the probability density function of $\mathbf{v}_i$. Finally, the unconditional log-likelihood function (GLL) takes the following form:

$$GLL = \sum_i \text{Ln}(GL_i) \quad (15)$$

The parameters to be estimated in mixed grouped ordered logit model with parameterized variance are: $\boldsymbol{\rho}$, $\boldsymbol{\varrho}_j$, $\boldsymbol{\Omega}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$. To estimate the proposed model, we apply Quasi-Monte Carlo simulation techniques described above. However, grouped ordered logit models without and with parameterized variance are nested versions of mixed grouped ordered logit model with parameterized variance where the density function $\wp(\mathbf{v}_i)$ degenerates to 1 with $\boldsymbol{\Omega} = 0$. Hence, the parameters for grouped ordered logit models without and with parameterized variance are estimated using maximum likelihood approaches without the numerical integration of the function in equation 14. The parameters to be estimated in these models are: $\boldsymbol{\rho}$, $\boldsymbol{\varrho}_j$, $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ for grouped ordered logit model with parameterized variance; and $\boldsymbol{\rho}$, $\boldsymbol{\sigma}_j$ and $\boldsymbol{\varphi}$ for grouped ordered logit model without parameterized variance ($\boldsymbol{\theta}=\mathbf{0}$).

3 DATA

3.1 Study Area

Our study area includes the Island of Montreal associated with 837 traffic analysis zones which covers only approximately 12% land area of the Greater Montreal (the second most populous metropolitan area in Canada). It is an old city and is characterized by a heterogeneous built environment with a dense old city near the original port. The transportation system of the city is characterized by a highly developed public transit system that includes a heavy-rail metro, commuter trains, public bike sharing system (BIXI) and an extensive bus network; with nearly 30% of Montrealers using public transit, walking or cycling for commuting (Statistics Canada, 2011). It is considered as one of the best cycling cities in the world (Walker, 2014). Montreal has the highest supply of bicycle infrastructure among all Canadian cities. There are more than 72 km of separated cycling lanes across the Island (Vijayakumar and Burda, 2015). To cope up with increasing popularity of cycling, the city is investing substantially in enhancing existing bicycle infrastructure. However, the bicycle crash rate of Montreal is one of the highest in Canada (7 crashes per 100,000 cycling trips are reported in Vijayakumar and Burda, 2015). Therefore, it is

important to identify the factors contributing to bicycle-motor vehicle crashes to make cycling safer and a more attractive and sustainable mode of transportation.

3.2 Data Description and Summary Statistics

This study is focused on bicycle-motor vehicle crash data at the zonal level. Data for our empirical analysis are sourced from the newspaper data archives of Montreal Gazette for the year 2006 through 2010. The datasets were obtained by the newspapers from the official crash databases maintained by the Société d'assurance automobile du Québec (SAAQ) for Montreal. The geocoded crash data are aggregated at the level of traffic analysis zone for each year. For the five years, Montreal has a record of 3,066 bicycle crashes with an average of 0.73 crashes (ranging from 0 to 28 crashes) per year per traffic analysis zone. In addition to the crash databases, the explanatory attributes considered in the empirical study are also aggregated at the traffic analysis zone level. For the empirical analysis, we selected variables that can be grouped into six broad categories: accessibility measures, exposure measures, built environment, road network characteristics, sociodemographic characteristics and socioeconomic characteristics. These data are extracted from the Geographic information system (GIS) data archive of Transportation Research at McGill (TRAM) of McGill University, Canada⁴.

Accessibility measures considered include number of bus stops, bus route length, commuter rail stations, commuter rail line length and metro line length. Exposure measures considered include bike lane length, bikes and automobiles shared road/lane length, bike path on sidewalk designated bike lane length, multi-use recreation bike path length and number of vehicles. Built environment considered include number of bars, distance from central business district (CBD), university area and land use mix. Road network characteristics considered include number of intersections, one-way road, length of highway and length of local road. Sociodemographic and socioeconomic characteristics considered include dependence (defined as proportion of youth and elderly relative to working adults) and median zonal income, respectively.

Table 1 offers a summary of the sample characteristics of the exogenous factors in the estimation dataset. Table 1 represents the definition of variables considered for final model estimation along with the zonal minimum, maximum and average values of continuous variables and percentages of indicator variables. The final specification of the model development was based on removing the statistically insignificant variables in a systematic process based on statistical confidence (90% confidence level). The specification process was also guided by prior research and parsimony considerations. In estimating the models, several functional forms and variable specifications are explored. The functional form that provided the best result is used for the final model specifications and, in Table 1, the variable definitions are presented based on these final functional forms of variables.

⁴ Exposure measures, sociodemographic characteristics, socioeconomic characteristics data are mainly compiled by TRAM from 2011 National Household Survey of Canada. Accessibility measures and land use including road network data are generally compiled from 2010 Société de transport de Montréal (STM) and 2011 general land use file available from TRAM.

4 EMPIRICAL ANALYSIS

4.1 Model Selection and Overall Measures of Fit

In the research effort, we estimated eight different models: 1) negative binomial model without parameterized overdispersion, 2) negative binomial model with parameterized overdispersion, 3) mixed negative binomial model with parameterized overdispersion, 4) generalized ordered logit model, 5) mixed generalized ordered logit model, 6) grouped ordered logit model without parameterized variance, 7) grouped ordered logit model with parameterized variance and 8) mixed grouped ordered logit model with parameterized variance. After extensively testing for different values in setting the upper limit for different generalized ordered and grouped ordered logit models, we set $J = 12$ for accommodating the long upper tail of bicycle crash count events under consideration. The number of records with counts greater than 12 amount to 0.36% of the full sample. Further, for different grouped ordered logit models, we estimated four different alternative specific components for 0, 1, 2 and 3 count events. Among these four alternative specific components, we found that equations for 0 and 1 categories were also moderated by the effects of different exogenous variables. It is worthwhile to mention here that it is possible to estimate count specific effects for more count events with the proposed grouped ordered logit system. However, adding more count-specific components did not improve the data fit further in the current study context and hence are not included in our final grouped ordered logit model specifications. Few significant parameters in the alternative specific components of the grouped ordered logit model indicate that for this count dataset the effects of a majority of the exogenous variables are monotonic across all count alternatives. Prior to discussing the estimation results, we compare the performance of these models in this section.

To compare models across negative binomial, generalized ordered logit and grouped ordered logit model frameworks, we undertake the comparison by employing a two-step approach. First, we evaluate the best fitted model across different frameworks by using the Bayesian Information Criterion (BIC). Second, an in-depth comparison for all count events are undertaken between two superior models identified from first step. The formulas for BIC measure is provided below:

$$BIC = -2LL + M * \ln(Q) \quad (16)$$

where LL is the log-likelihood value at convergence, M is the number of parameters and Q is the number of observations. The results for BIC measure along with number of parameters, log-likelihood at constant and log-likelihood at convergence for all estimated models are presented in Table 2. The model with the lower BIC value is the preferred model. From Table 2, we can see that the best model fit is obtained for mixed negative binomial model with parameterized overdispersion, mixed generalized ordered logit model and mixed grouped ordered logit model with parameterized variance within negative binomial, generalized ordered logit and grouped ordered logit frameworks, respectively. Overall, the lowest BIC value was obtained for mixed grouped ordered logit model with parameterized variance followed by mixed negative binomial model with parameterized overdispersion. The comparison exercise clearly highlights the superiority of the mixed grouped ordered logit model with parameterized variance in terms of data fit compared to mixed negative binomial model with parameterized overdispersion and mixed generalized ordered logit model in the current study context. Further, from the model comparison

it is also clear that the mixed negative binomial model with parameterized overdispersion performs better than the mixed generalized ordered logit model. Henceforth, for the second step of data fit evaluation, we consider mixed grouped ordered logit model with parameterized variance and mixed negative binomial model with parameterized overdispersion.

To further investigate the performance differences between mixed grouped ordered logit model with parameterized variance and mixed negative binomial model with parameterized overdispersion, we undertake several in-depth comparisons. For each count alternative, we compute contributions to log-likelihood (from 0 to ≥ 12). The results for these measures are presented in Table 3. From Table 3, we can see that for the count categories with at least one percent of data (for 0, 1, 2, 3 and 4 count categories), the mixed grouped ordered logit model with parameterized variance outperformed the mixed negative binomial model with parameterized overdispersion in terms of contribution to log-likelihood values. In summary, the mixed grouped ordered logit model with parameterized variance offers equivalent or superior performance to the corresponding count regression system in the context of bicycle crash count modeling and thus is a promising alternate econometric framework for examining crash count events.

4.2 Estimation Results

In presenting the effects of exogenous variables, we will restrict ourselves to the discussion of the mixed negative binomial model with parameterized overdispersion and the mixed grouped ordered logit model with parameterized variance only. Table 4 presents the results of the mixed negative binomial model with parameterized overdispersion results in second and third columns of the table and mixed grouped ordered logit model with parameterized variance results from fourth to thirteen columns. A positive (negative) sign for a variable in the negative binomial model of Table 4 indicates that an increase in the variable is likely to result in more (less) bicycle crashes. In the mixed grouped ordered logit model, a positive (negative) sign for a variable in the propensity indicates that an increase in the variable is likely to result in more (less) bicycle crashes. At the same time, when the count-specific parameter of the mixed grouped ordered logit model is positive (negative), the result implies that the likelihood of that specific count event is bound to increase (decrease); the actual effect on the probability is quite non-linear and can only be judged in conjunction with the influence of the variable on propensity and other count-specific effects. In the following sections, the estimation results are discussed by variable groups for both models.

4.2.1 Accessibility Measures

An increase in the number of bus stop increases the likelihood of bicycle-motor vehicle crashes in both negative binomial and grouped ordered logit components. Our analysis from both models also shows that zones with more bus route are likely to be positively correlated with higher bicycle crash risk. The negative binomial component estimates reveal that commuter rail station has a negative correlation with bicycle crash risk. The commuter rail station result of the grouped ordered logit component indicates that the latent crash risk is lower for zone with higher number of station. The effect of commuter rail line length is found significant only in the negative binomial component and the result indicates a lower likelihood of bicycle crash risk in the presence of higher commuter rail line length.

4.2.2 Exposure Measures

Several exposure measures considered are found to be significant determinants of bicycle crash risk. Among those, we can observe that zones with more kilometers of bikes and automobiles shared road/lane length in the grouped ordered logit component decrease the likelihood of bicycle-motor vehicle crashes. At the same time, length of bike path in the negative binomial component shows a negative correlation with bicycle crash risk. The negative binomial component result for multi-use recreation bike path indicates a higher likelihood of bicycle crash risk. The corresponding results from the grouped ordered logit framework also suggest that bicycle crash risk propensity is higher for zones with more multi-use recreational bike path. The effect of the variable on count-specific equation for 0 count is also significant and indicates a lower likelihood of zero zonal level bicycle crashes. Our study also found that more vehicles within a zone leads to higher probability of bicycle crashes in both negative binomial model and grouped ordered logit models.

4.2.3 Built Environment Characteristics

With regards to built environment, the results for negative binomial and grouped ordered logit frameworks reveal that bicycle crashes are positively associated with higher number of bars in the neighborhood. As expected, the possibility of higher bicycle crashes decreases with increasing distance from CBD to the zone and the effects are significant in both models. Moreover, the coefficient of distance from CBD in the grouped ordered logit framework is moderated by unobserved effects resulting in statistically significant standard deviation parameters. With the estimated parameter, more than 99% of the distribution is less than zero. In Table 4, greater university area in a zone has positive impact on bicycle crash risk for both models. Further, the zones with higher land use mix are more likely to be associated with higher bicycle crash risk in both models. The reader would note that in both systems, the impacts of land use mix are moderated by unobserved effects resulting in statistically significant standard deviation parameters in both models. Effect of land use mix variable in the negative binomial model indicate that almost 90% of zones are likely to have bicycle-motor vehicle crashes while it increases the probability of bicycle-motor vehicle crashes in almost 64% cases in the grouped ordered logit model.

4.2.4 Road Network Characteristics

From Table 4, we can see that number of intersections has significant impact on bicycle crash risk. An increase in total number of intersections in a zone increases the likelihood of bicycle crash risk in the mixed negative binomial model with parameterized overdispersion. The variable has similar effect in mixed grouped ordered logit model with parameterized variance. At the same time, the negative value of the 0-count equation of the variable reflects a decrease in zero crash outcome at zonal level. In terms of one-way road, negative binomial and grouped ordered logit frameworks estimates reveal very similar effects indicating that the odds of bicycle crash risks are higher in zones with higher number of one-way road links. The effect of highway road length is found significant in both models and the results show that in presence of more highway roads in a zone, the possibility of crash risk decreases. The grouped ordered logit model estimates for highway road length also results in a parameter that is normally distributed with a mean -0.351 and standard deviation 1.147, which indicates that higher highway length increases the probability of bicycle crash risk in almost 38% cases. Further, the variable also has a positive effect in 1-count event equation of grouped ordered logit model which implies that zones are likely to have higher number of 1 crash when there are more highway roads.

4.2.5 Sociodemographic Characteristics

In terms of sociodemographic characteristics, dependence variable (defined as proportion of youth and elderly relative to working adults) reveals a lower probability of bicycle-motor vehicle crash risk for higher values of the variable from both negative binomial and grouped ordered logit frameworks.

4.2.6 Socioeconomic Characteristics

The effects of median zonal income are found to have significant impact on bicycle-motor vehicle crash risk. In terms of zonal income, we can see that medium level median zonal income in both negative binomial and grouped ordered logit frameworks are positively correlated with bicycle crash risk relative to low and high level median income implying that these zones are likely to have higher number of bicycle crash risk relative to high level median income zones.

4.2.7 Overdispersion and Variance Components

As indicated earlier, we hypothesize that the overdispersion profile and the variance profile of the mixed negative binomial model with parameterized overdispersion and mixed grouped ordered logit model with parameterized variance, respectively, are not constant across the entire database. Thus, these components are estimated as functions of observed exogenous variables in current study context. Such parameterization efforts are analogous to the covariance heterogeneity parameterization employed in nested logit models (Bhat, 1997). Ignoring such heterogeneity (when present) will lead to biased and inconsistent estimates (Chamberlain, 1980; Bhat, 1997). From Table 4, we can see that none of the exogenous variables have significant effect in the overdispersion and variance profiles of the mixed negative binomial model with parameterized overdispersion and the mixed grouped ordered logit model with parameterized variance, respectively. The results signify that, in our empirical context, the overdispersion profile of the mixed negative binomial model with parameterized overdispersion and variance profile of the mixed grouped ordered logit model with parameterized variance are constant across the entire database.

4.3 Predictive Performance Evaluation

In an effort to assess the predictive performance of the estimated models, computation of several in-sample goodness-of-fit measures are also carried out. In doing so, performance of the mixed negative binomial model with parameterized overdispersion and the mixed grouped ordered logit model with parameterized variance are compared by predicting bicycle crash frequency across different count events in the estimation sample. The results for both models are presented in the upper row panel of Table 5 along with the observed average crash frequencies in the estimation sample. From Table 5 we can see that both models overpredict and underpredict crash counts for different count events. The underpredictions in average crash frequencies for the mixed negative binomial model with parameterized overdispersion and the mixed grouped ordered logit model with parameterized variance range from (-2.18, -12.59) and (-1.31, -6.36), respectively. The overpredictions in average crash frequencies for the mixed negative binomial model and the mixed grouped ordered logit model range from (0.04, 12.49) and (1.60, 11.71), respectively. In general the grouped ordered logit offers better predictions overall relative to the negative binomial model estimates.

To further evaluate the in-sample predictive performance of the mixed negative binomial model with parameterized overdispersion and the mixed grouped ordered logit model with parameterized variance, we employ three different fit measures: mean prediction bias (MPB), mean absolute deviation (MAD) and mean squared prediction error (MSPE). We compute these measures both at the aggregate and disaggregate levels. For aggregate measures, the measures are computed at each crash-count levels (from 0 to ≥ 12); while at the disaggregate level we compute the measures at the study unit level (zone) and compute the average measures across all units.

MPB represents the magnitude and direction of average bias in model prediction. Negative sign is associated with underprediction and positive sign is associated with overprediction. The model with the lower MPB in terms of value provides better prediction of the observed data and is computed as:

$$MPB = \frac{\sum_{n=0 \text{ or } 1}^N (\hat{y}_n - y_n)}{N} \quad (17)$$

where, $\hat{y}_n$ and y_n are the predicted and observed average number of crashes across different crash count events/study units n . At the count events level $n = 0, 1, 2, \dots, 13$ and at the study unit level $n = 1, 2, \dots, 837$. On the other hand, MAD describes average misprediction of the estimated models. The model with lower MAD value closer to zero provides better average predictions of observed data. MAD is defined as:

$$MAD = \frac{\sum_{n=0 \text{ or } 1}^N |\hat{y}_n - y_n|}{N} \quad (18)$$

Finally, MSPE quantifies the error associated with model predictions and is defined as:

$$MSPE = \frac{\sum_{n=0 \text{ or } 1}^N (\hat{y}_n - y_n)^2}{N} \quad (19)$$

The smaller the MSPE, the better the model predicts the observed data.

The results for both models at the aggregate and disaggregate levels are presented in the lower row panel of Table 5 for these three in-sample predictive measures of fit for mixed negative binomial model with parameterized overdispersion and mixed grouped ordered logit model with parameterized variance. The resulting fit measures for comparing the predictive performance clearly indicate that the mixed grouped ordered logit model with parameterized variance offers superior predictions compared to the mixed negative binomial model with parameterized overdispersion model both at the aggregate and disaggregate levels in current study context. Overall, the results indicate that the mixed grouped ordered logit model with parameterized variance offers either equivalent or superior predictions relative to the mixed negative binomial model with parameterized overdispersion both at aggregate and disaggregate levels. Hence, we can argue that the mixed grouped ordered logit model with parameterized variance can serve as a promising complementary model structure to the mixed negative binomial model with parameterized overdispersion in examining zonal level crash counts.

5 CONCLUSIONS

The study effort contributes to literature on crash frequency analysis. We proposed and estimated an econometric framework – Mixed Grouped Response Ordered Logit Count model – for analysing crash count events. The proposed Grouped Response Ordered Logit Count framework develops an alternative structure to the work by Bhat and colleagues on the Generalized Ordered Logit Count models. The proposed framework relates the crash count propensity to the observed counts directly while also automatically accommodating for heteroscedasticity and unobserved heterogeneity. The proposed model is demonstrated in the study by using Traffic Analysis Zone level crash count data for the Island of Montreal for the year 2006 through 2010 employing a comprehensive set of exogenous variables – accessibility measures, exposure measures, built environment, road network characteristics, sociodemographic and socioeconomic characteristics.

In order to demonstrate the applicability and performance of the proposed framework, we also empirically compared the performance of the grouped ordered logit model with negative binomial and generalized ordered logit frameworks (and their variants) in the context of bicycle-motor vehicle crash count events. Specifically, in the current study approach, we estimated crash count models based on three different frameworks – negative binomial, generalized ordered logit and grouped ordered logit model frameworks. The entire set of models considered in our analysis included (1) negative binomial model without parameterized overdispersion, (2) negative binomial model with parameterized overdispersion, and (3) mixed negative binomial model with parameterized overdispersion, (4) generalized ordered logit count model and (5) mixed generalized ordered logit count model, (6) grouped response ordered logit count model without parameterized variance, (7) grouped response ordered logit count model with parameterized variance and (8) mixed grouped response ordered logit count model with parameterized variance.

We compared the performance of the estimated models by generating a comprehensive set of measures to evaluate model performance and data fit across these frameworks. The comparison exercise based on Bayesian Information Criterion clearly highlighted the superiority of the mixed grouped response ordered logit count model with parameterized variance in terms of data fit compared to the data fit of the other estimated models. Further, from the model comparison it was also clear that the mixed negative binomial model with parameterized overdispersion was the second-best model in terms of data fit compared to other estimated models. Moreover, the mixed grouped ordered logit model with parameterized variance that allows for exogenous variable effects to vary across alternatives offered equivalent performance to the corresponding count regression system in the context of bicycle crash count modeling and thus is a promising alternate econometric framework for examining crash count events. In an effort to assess the predictive performance of the estimated models, computation of several in-sample goodness-of-fit measures were also carried out. The resulting fit measures for comparing the predictive performance also indicated that the mixed grouped response ordered logit count model with parameterized variance offered either equivalent or superior predictions relative to the mixed negative binomial model with parameterized overdispersion both at the aggregate and disaggregate levels. The results from this comparison exercise points out that the proposed grouped ordered logit framework is a promising alternate econometric framework for examining count event.

The paper is not without limitations. In our research effort, we examined crash data with highest maximum count of 28. However, at an aggregation level, the maximum count in terms of crash count events might be higher. It will be an interesting exercise to evaluate the performance of the proposed model with the traditional count models with higher crash count events. To be

sure, it is important to mention here that the proposed approach is not suggested as a replacement to existing count modeling approaches but as an alternative approach that could potentially augment the available approaches for crash prediction analysis. The model approach proposed is conducive to development of multivariate count models. It would be interesting to embed the proposed model structure within multivariate frameworks and compare the model performance with state of the art multivariate models including multivariate negative binomial or log-normal models, latent class flexible mixture multivariate model, multivariate models with spatial and temporal correlations, and recently proposed fractional split formulations (Mothafer et al., 2016; Heydari et al., 2017; Liu and Sharma, 2017; Liu and Sharma 2018; Yasmin and Eluru, 2018; Bhowmik et al., 2018) Finally, it would also be a useful exercise to compare the performance of the proposed approach in relation to the Generalized Extreme Value based count models proposed recently by Paleti (2016).

ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge Prof. Ahmed El-Geneidy's TRAM website for providing access to the GIS data and would like to acknowledge help provided by Adham Badran, and Michelle Pinto in preparing the datasets. The authors would also like to gratefully acknowledge funding from Florida Department of Transportation. The authors would also like to acknowledge feedback from the editor and two anonymous reviewers.

References

- Aptech 2015, Aptech Systems Inc, accessed from <http://www.aptech.com/> on September 19th 2015.
- Bhat, C. R., 1997. Covariance heterogeneity in nested logit models: econometric structure and application to intercity travel. *Transportation Research Part B*, 31 (1), 11-21.
- Bhat, C. R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B*, 35(7), 677-693.
- Bhat, C. R., Astroza, S., Sidharthan, R., Alam, M. J. B., and Khushefati, W. H., 2014. A joint count-continuous model of travel behavior with selection based on a multinomial probit residential density choice model. *Transportation Research Part B*, 68, 31-51.
- Bhat, C., 1994. Imputing a continuous income variable from grouped and missing income observations. *Economics Letters*, 46 (4), 311-320.
- Bhowmik, T., Yasmin, S., and Eluru, N., 2018. A joint econometric approach for modeling crash counts by collision type. *Analytic Methods in Accident Research* 19, 16-34.
- Castro, M., Paleti, R., and Bhat, C. R. (2013). A spatial generalized ordered response model to examine highway crash injury severity. *Accident Analysis and Prevention*, 52, 188-203.
- Castro, M., Paleti, R., and Bhat, C. R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: application to predicting crash frequency at intersections. *Transportation Research Part B*, 46(1), 253-272.
- Chamberlain, G., 1980. Analysis of covariance with qualitative data. *The Review of Economic Studies*, 47(1), 225-238.

- Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis and Prevention*, 40, 1033-1054.
- Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in empirical bayes estimation. *Accident Analysis and Prevention*, 33(6), 799-808.
- Hausman, J. A., Hall, B. H., and Griliches, Z., 1984. *Econometric models for count data with an application to the patents R&D relationship*. National Bureau of Economic Research Cambridge, Mass., USA.
- Liu, C., and Sharma, A., 2017. Exploring spatio-temporal effects in traffic crash trend analysis. *Analytic Methods in Accident Research*, 16, 104–116.
- Liu, C., and Sharma, A., 2018. Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. *Analytic Methods in Accident Research*, 17, 14–31.
- Lord, D., and Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A*, 44(5), 291-305.
- Mannering, F. L., and Bhat, C. R., 2014. Analytic methods in accident research: methodological frontier and future directions. *Analytic Methods in Accident Research*, 1, 1-22.
- Mannering, F., Shankar, V., and Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, 11, 1-16.
- Mothafer, G.I., Yamamoto, T., Shankar, V.N., 2016. Evaluating crash type covariances and roadway geometric marginal effects using the multivariate Poisson gamma mixture model. *Analytic Methods in Accident Research*, 9, 16–26.
- Narayanamoorthy, S., Paleti, R., and Bhat, C. R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. *Transportation Research Part B*, 55, 245-264.
- Noland, R. B., and Quddus, M. A., 2004. Analysis of pedestrian and bicycle casualties with regional panel data. *Transportation Research Record*, 1897, pp. 28-33.
- Paleti, R., 2016. Generalized extreme value models for count data: application to worker telecommuting frequency choices. *Transportation Research Part B*, 83, 104-120.
- Savolainen, P. T., Mannering, F. L., Lord, D., and Quddus, M. A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident Analysis and Prevention*, 43(5), 1666-1676.
- Statistics Canada, 2011. 2011 Census of population. Ottawa: Statistics Canada.
- Stewart, M.B., 1983. On least squares estimation when the dependent variable is grouped. *Review of Economic Studies*, 737-753.
- Vijayakumar N., and Burda, C., 2015. *Cycle cities: supporting cycling in Canadian cities*. Pembina Foundation for Environmental Research and Education.
- Walker P., 2014. World's best cycling cities, CNN, accessed from <http://www.cnn.com/2014/08/17/travel/best-cycling-cities/> on July 20, 2016.

- Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., and Bhatia, R., 2009. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accident Analysis and Prevention*, 41(1), 137-145.
- Yasmin, S., & Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis and Prevention*, 59, 506-521.
- Yasmin, S., and Eluru, N., 2016. Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. *Accident Analysis and Prevention*, 95, 157-171.
- Yasmin, S., and Eluru, N., 2018. A joint econometric framework for modeling crash counts by severity. *Transportmetrica A*, 14(3), 230-255.

Table 1: Sample Statistics for the Island of Montreal

Variables Name	Definition	Zonal		
		Minimum	Maximum	Average
Accessibility measures				
Bus stops	Ln(Total Société de transport de Montréal (STM) bus stops in TAZ)	0.000	10.697	7.816
Bus route length	Ln(Total Société de transport de Montréal (STM) bus line kilometer in TAZ)	0.000	4.127	1.98
Commuter rail stations	Number of Agence métropolitaine de transport (AMT) commuter rail stations	0.000	1.000	0.030
Commuter rail line length	Length of Agence métropolitaine de transport (AMT) commuter rail line length in kilometer	0.000	4.864	0.117
Metro line length	Ln(Total metro line length kilometer in TAZ)	-6.354	1.401	-0.048
Exposure measures				
Bike lane	Ln(Total length of bike lane in TAZ (kilometer))	-8.580	1.491	-0.136
Bikes and automobiles shared road/lane	Ln(Total length of bikes and automobiles shared road/lane in TAZ (kilometer))	-7.610	1.079	-0.164
Bike path on sidewalk	Ln(Total length of bike path on sidewalk in TAZ (kilometer))	-4.146	0.000	-0.034
Designated bike lane	Ln(Total length of designated bike lane kilometer on road in TAZ (kilometer))	-7.193	1.024	-0.242
Multi-use recreation bike path	Ln(Total length of Multi-use recreation bike path in TAZ (kilometer))	-4.242	2.763	-0.012
Number of vehicles	Ln(Total number of vehicles in TAZ)	0.432	8.991	6.39
Built environment				
Number of bars	Ln(total number of bars in TAZ)	0.000	3.045	0.13
Distance from CBD	Ln(distance from CBD to the TAZ (kilometer))	-2.083	3.516	1.892
University area	Ln(University are in TAZ (square kilometer))	-4.870	13.382	0.572
Land use mix	Land use mix = $\left[\frac{-\sum_k(p_k(\ln p_k))}{\ln N}\right]$, where k is the category of land-use, p is the proportion of the developed land area devoted to a specific land-use, N is the number of land-use categories in a TAZ	0.000	0.999	0.494
Road network characteristics				
Number of intersections	Ln(total number of intersection in TAZ)	0.000	5.746	3.061
One-way road	Ln(Number of one-way road links in TAZ)	0.000	4.920	2.801
Length of highway	Ln(total length of highway kilometer in TAZ)	-6.465	2.757	-0.109
Length of local road	Ln(total length of local road kilometer in TAZ)	-29.698	3.944	1.077
Sociodemographic characteristics				

Dependence	Ratio of youth (19 years or younger) and elderly (65 years or more) to working age persons	0.000	1.557	0.547
Variables		Sample Share Frequency (Percentage)		
Median TAZ income				
Low (<\$40,000)		1565 (37.395)		
Medium (\$40,000-\$80,000)		460 (10.992)		
High (≥\$80,000)		2160 (51.613)		

Note: TAZ = Traffic analysis zone; CBD = Central Business District

Table 2: Data Fit Measures across Different Models

Frameworks	Models	Number of parameters	Log-likelihood at constant	Log-likelihood at convergence	BIC
NB	NB	18	-4704.526	-4382.825	8915.757
	PNB	25	-4704.526	-4359.956	8928.394
	MNB	20	-4704.526	-4373.838	<u>8914.461</u>
GOLC	GOLC	24	-5787.813	-4512.644	9225.430
	MGOLC	24	-5787.813	-4371.316	8942.775
GROLC	GROLC	26	-4761.484	-4344.490	8905.802
	PGROLC	29	-4761.484	-4333.275	8908.388
	MGROLC	26	-4761.484	-4336.162	<u>8889.145</u>

Note: NB = Negative binomial model without parameterized overdispersion; PNB = Negative binomial model with parameterized overdispersion; MNB = Mixed negative binomial model with parameterized overdispersion; GOLC = Generalized ordered logit model; MGOLC = Mixed generalized ordered logit model; GROLC = Grouped ordered logit model without parameterized variance; PGROLC = Grouped ordered logit model with parameterized variance; MGROLC = Mixed grouped ordered logit model with parameterized variance

Table 3: Measures of Fit across Different Count Events

Crash count	Number of Zones		Log-likelihood	
	Frequencies	Percentage	MNB	MGROLG
0	2785.00	66.547	-1052.285	-1062.553
1	756.00	18.065	-1246.882	-1236.737
2	308.00	7.360	-759.444	-742.887
3	137.00	3.274	-433.276	-433.910
4	81.00	1.935	-299.866	-288.498
5	30.00	0.717	-126.535	-136.496
6	29.00	0.693	-125.707	-130.733
7	17.00	0.406	-86.875	-83.832
8	12.00	0.287	-61.480	-60.492
9	12.00	0.287	-65.118	-64.544
10	1.00	0.024	-4.242	-5.925
11	2.00	0.048	-12.366	-12.450
≥12	15.00	0.358	-99.761	-77.100
Total	4185	100%	---	---

Note: MNB = Mixed negative binomial model with parameterized overdispersion; MGROLG = Mixed grouped ordered logit model with parameterized variance

Table 4: Model Estimates for Montreal Island

Variables Name	MNB		MGROLC									
	Coeff	t-stat	Propensity		0-count event		1-count event		2-count event		3-count event	
			Coeff	t-stat	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Constants	-4.401	-13.355	-22.228	-7.060	8.047	5.430	3.301	4.463	1.668	3.554	0.773	2.815
Accessibility measures												
Bus stops	0.237	3.783	0.810	3.516	--	--	--	--	--	--	--	--
Bus route length	0.074	2.145	0.314	2.612	--	--	--	--	--	--	--	--
Commuter rail stations	-0.612	-2.435	-2.864	-3.482	--	--	--	--	--	--	--	--
Commuter rail line length	-0.217	-2.350	--	--	--	--	--	--	--	--	--	--
Exposure measures												
Bikes and automobiles shared road/lane	--	--	-0.293	-2.042	--	--	--	--	--	--	--	--
Bike path on sidewalk	-0.401	-5.022	--	--	--	--	--	--	--	--	--	--
Multi-use recreation bike path	0.247	2.746	1.062	2.832	-0.570	-1.931	--	--	--	--	--	--
Number of vehicles	0.371	8.390	1.243	5.976	--	--	--	--	--	--	--	--
Built environment												
Number of bars	0.227	3.168	1.174	3.744	--	--	--	--	--	--	--	--
Distance from CBD	-0.597	-9.832	-2.146	-7.079	--	--	--	--	--	--	--	--
SD Distance from CBD	0.211	4.337	--	--	--	--	--	--	--	--	--	--
University area	0.050	3.852	0.132	2.477	--	--	--	--	--	--	--	--
Land use mix	0.858	5.003	1.791	2.832	--	--	--	--	--	--	--	--
SD Land use mix	0.644	3.259	5.089	5.515	--	--	--	--	--	--	--	--
Road network characteristics												
Number of intersections	0.143	2.470	0.992	3.745	-0.585	-3.367	--	--	--	--	--	--
One-way road	0.208	4.584	0.713	3.843	--	--	--	--	--	--	--	--
Length of highway	-0.092	-2.587	-0.351	-2.359	--	--	0.259	2.376	--	--	--	--
SD Length of highway	--	--	1.147	3.160	--	--	--	--	--	--	--	--

Sociodemographic characteristics												
Dependence	-0.582	-3.431	-2.415	-3.527	--	--	--	--	--	--	--	--
Socioeconomic characteristics												
Median TAZ income												
Medium (\$40,000-\$80,000)	0.274	4.404	1.061	3.754	--	--	--	--	--	--	--	--
Overdispersion parameter	0.164	1.239	--	--	--	--	--	--	--	--	--	--
Variance parameter	--	--	2.826	9.522	--	--	--	--	--	--	--	--

Note: MNB = Mixed negative binomial model with parameterized overdispersion; MGROLG = Mixed grouped ordered logit model with parameterized variance; TAZ = Traffic analysis zone; CBD = Central Business District; SD = Standard deviation

Table 5: Predictive Performance Evaluation

PREDICTIVE PERFORMANCE ACROSS COUNT ALTERNATIVES														
Crash Count		0	1	2	3	4	5	6	7	8	9	10	11	>=12
Observed Frequency		2784	756	308	137	81	30	29	17	12	12	1	2	16
Predicted Frequency	MNB	2788.49	754.43	295.51	139.27	73.84	42.59	26.20	16.96	11.46	8.01	5.77	4.27	18.18
	MGROL C	2772.29	754.40	310.13	139.43	83.31	32.33	24.44	18.31	13.60	10.03	7.36	5.37	14.00
AGGREGATE PREDICTIVE MEASURES OF FIT														
Models		MPB				MAD				MSPE				
MNB		0.000				4.763				37.518				
MGROL C		0.000				3.638				20.457				
DISAGGREGATE PREDICTIVE MEASURES OF FIT														
Models		MPB				MAD				MSPE				
MNB		0.020				0.888				2.702				
MGROL C		0.020				0.851				2.009				

Note: MNB = Mixed negative binomial model with parameterized overdispersion; MGROL C = Mixed grouped ordered logit model with parameterized variance; MPB = Mean prediction bias; MAD = Mean absolute deviation; MSPE = Mean squared prediction error