An Econometric Framework for Integrating Aggregate and Disaggregate Level Crash Analysis

Shahrior Pervaz
Graduate Research Assistant
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 1-407-561-0298; Fax: 1-407-823-3315
Email: shahrior.pervaz@knights.ucf.edu
ORCiD number: 0000-0001-7966-7083

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

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

*Corresponding author

ABSTRACT
Traditionally, aggregate crash frequency by severity and disaggregate severity analysis have been conducted independently in the safety literature. The current research effort contributes to the safety literature by bridging the gap between these two different streams of research by using both aggregate and disaggregate level crash data simultaneously. To be specific, the study proposes a framework that integrates aggregate and disaggregate level models. The proposed framework allows for the influence of independent variables at the crash record level to be incorporated within the aggregate level propensity estimation. The empirical analysis is based on the crash data drawn from the city of Orlando, Florida for the year 2019. The disaggregate level analysis uses 20,204 crash records that contain crash specific variables, temporal characteristics, roadway, vehicle and driver factors, road environmental and weather information for each record. For aggregate level model analysis, the study aggregated the crash records by severity class over 300 traffic analysis zones. An exhaustive set of independent variables including roadway and traffic factors, land-use attributes, built environment, and sociodemographic characteristics are considered in this analysis. The empirical analysis is further augmented by employing several goodness of fit and predictive measures. A validation exercise is also performed using a holdout sample to highlight the superior performance of the proposed integrated model relative to the non-integrated crash count by severity model. The proposed model can also accommodate common unobserved spatial correlation among crash records within the same zone. The model results illustrate the benefits of developing an integrated model system for crash frequency and severity.

Keywords: *Aggregate crash count by severity; Disaggregate severity analysis; Integrated model; Unobserved effects.*

1. BACKGROUND

Transportation safety modeling has evolved along two streams. The first stream of research – crash frequency models – examine the factors affecting the occurrence of crashes on transportation facilities. These studies are generally conducted employing aggregate data at the microscopic (segments or intersections) or macroscopic spatial scales (zones) to improve roadway design and operational efficiency. The second stream of studies – crash severity models – examine factors affecting crash consequences (usually severity) at the disaggregate level (such as driver, vehicle or crash record). Across the two streams, several methodological enhancements have occurred to improve crash frequency and severity models. In recent years, crash frequency models are typically multivariate in nature accommodating for crash frequency by collision type and/or severity (see for example Bhowmik et al., 2021a; Yasmin and Eluru, 2018; Afghari et al., 2020). Crash severity models that allow for the influence for observed and unobserved factors are the expected norm in the literature (see for example Kabli et al., 2020; Xiong and Mannering, 2013). Despite the tremendous progress across the two streams of research, transportation safety literature treats these two model systems as independent. However, each crash record in the crash frequency model has a corresponding record in the crash severity model system i.e., crashes employed in frequency models are aggregated from disaggregate records.

Transportation safety studies consistently incorporate aggregate level factors such as traffic volume, roadway factors, neighborhood characteristics, land-use and built environment characteristics in modeling crash frequency, crash severity and/or crash type at the disaggregate level (see for example Ahmed et al., 2023; Pervaz et al., 2022). However, the current state of modeling does not accommodate for any disaggregate level independent variables in modeling crash frequency. To be sure, important characteristics of crashes such as crash type and crash severity have been considered to create crash frequency variables (by type and severity). Yet, the crash generation process is assumed to be independent across the aggregate and disaggregate resolutions. Consider the following situation. The presence of higher pedestrian volumes in a zone might result in an increase in the number of crashes in the zone. At the disaggregate level, the presence of a crash in higher pedestrian volume location might result in less severe crashes occurring at intersections (such as rear-end crashes). However, in current approaches there is no mechanism to incorporate the crash record specific information in the frequency model (even if it were partitioned by crash type or severity). Thus, current approaches to safety modeling do not allow for adequate interaction of variable impacts across crash frequency and crash severity models. Towards addressing this limitation, the objective of the current paper is to develop a unified model system that improves how the influence of observed and unobserved variables at the disaggregate resolution in the severity model affect crash frequency modeling process. This approach will allow us to evaluate the influence of independent variables that are traditionally examined separately within a unified framework.

Towards this end, the current study develops an integrated model framework that jointly estimates crash frequency and severity models at their corresponding resolution. The framework involves incorporating the overall impact of independent variables for a crash (crash propensity) from the disaggregate model within the zone level crash frequency propensity estimation as an additional independent variable. The approach can take two forms. In the first structure, the disaggregate level model propensity can be summed up for all crashes at the zone level as a composite score and treated as an exogenous variable i.e., the severity model parameters are fixed and the parameter for the composite score variable is estimated in the frequency propensity equation. Alternatively, disaggregate level model propensity can be treated as endogenous and be

estimated simultaneously with the propensity of the crash frequency model. In this approach, the estimates of the disaggregate model will be allowed to vary while modeling crash frequency. The second approach is computationally more involved while allowing for feedback between aggregate and disaggregate level models. The reader would note that the proposed approach requires the development of crash data in a consistent and integrated manner at the two resolutions. Specifically, the approach should begin at the disaggregate level and then be aggregated to generate crash frequency at the aggregate level. The proposed integrated model system is estimated using data drawn from the City of Orlando for the year 2019 with exhaustive crash and zone-specific variables. The model results illustrate the benefits of developing an integrated model system for crash frequency and severity.

2. EARLIER RESEARCH AND CURRENT STUDY IN CONTEXT

Crash frequency and severity domains have been extensively explored in safety literature. An exhaustive review of literature from these two domains is beyond the scope of the current paper. For recent reviews of methodology, relevant to frequency domain see (Lord and Mannering, 2010; Mannering et al., 2016; Yasmin and Eluru, 2018) and severity domain see (Mannering et al., 2016; Savolainen et al., 2011; Yasmin and Eluru, 2013)[1]. In the current study, we focus on methods that consider interlinking two decision processes. Two approaches are traditionally employed to achieve the interlinking of multiple dependent variables at different resolutions. The first and the more commonly employed approach to achieve the interlinking processes employs unobserved factor-based methods that recognize the repeated presence of the finer resolution data in the coarser resolution record. These models are typically labelled as hierarchical models (for example see Alarifi et al., 2018; Huang et al., 2016; Huang and Abdel-Aty, 2010). The second and the approach proposed in the current study directly interlinks the two decision processes via composite variables derived from observed variables[2]. The approach is relatively new in safety literature and has been adopted in a limited number of studies (Cai et al., 2019; Pervaz et al., 2022). These model systems recognize that crashes at the micro level facilities contribute to total macroscopic level crash counts. To allow for this influence, Pervaz et al. (2022) adds one component per micro level facility type (such as intersection or segment) in the form of an additional variable in the propensity of the macroscopic model system (Pervaz et al., 2022). The component for a facility type is evaluated as the sum of crash propensity for all facilities of that type in the spatial unit. The approach described develops an integrated approach across two aggregate resolutions – macro (zone) and micro (intersections and segments). The proposed approach draws on this idea to interlink aggregate crash frequency and disaggregate severity models.

While such integrated models are relatively new in the safety field, there have been multiple examples of modeling approaches that employ composite variables from finer resolution outcomes in the travel behavior modeling field. The approaches employed include activity travel choice and vehicle ownership choice models. Activity travel choices (such as mode, destination, activity type, activity duration) are likely to be simultaneously considered and hence are assumed to give rise to a sequential deeply nested logit model (Ben-Akiva and Lerman, 1985). As it is not

[1] The reader would note that crash frequency domain studies can also include studies that examine crash rates using censored regression models such as tobit models (see Ahmed et al., 2022; Anastasopoulos et al., 2012a, 2012b; Zeng et al., 2017).

[2] The reader would note that approaches that accommodate for crash frequency by severity or type are an improvement over aggregate crash frequency models (see Bhowmik et al., 2021a; Yasmin and Eluru, 2018; Afghari et al., 2020). However, these models are still aggregate in resolution and do not account for crash record level factors.

computationally feasible to simultaneously estimate deeply nested logit models with multiple decision variables, log-sum from one level is carried up to the next higher level, resulting in a sequential estimation approach (see Eluru et al., 2010 for details). Similarly, vehicle ownership models for vehicle type (such as sedan, SUV) and usage (mileage) analysis consider composite variables (or log-sum variables) generated from finer resolution vehicle make (such as Honda, Toyota) and model (Civic, Camry) attributes within the vehicle type alternative propensity equations (Bhat et al., 2009). It is important to recognize that while these approaches are based on decisions across multiple dependent variables the decisions represent the behavior of a single decision unit.

Drawing inspiration from the aforementioned studies, the current research proposes an integrated model framework that allows for the influence of disaggregate level variables as a composite variable within the aggregate level propensity estimation. The approach would involve summing up the crash propensity of each disaggregate level severity record within the aggregate resolution and adding the generated value as a new variable in the aggregate model. The propensity will incorporate disaggregate level variables including crash characteristics (such as crash types, first harmful events), driver characteristics (such as driving under influence related, distraction related), and environmental characteristics (such as clear, rainy). In our study, the aggregate model is employed to examine crash frequency by severity and the disaggregate model is employed to examine crash severity. A negative binomial-ordered probit fractional split (NB-OPFS) framework is employed to examine crash frequency by severity. Specifically, the negative binomial component models the total number of crashes and the ordered probit fractional split component determines the proportion of each severity at a zone. The crash severity variable is examined using the ordered probit model. The integrated approach can take two potential forms with these models. In the first structure, the ordered probit model propensity across the crashes in the zone is computed as a composite score and treated as an exogenous variable i.e., the severity model parameters are fixed. In this approach, an additional parameter for the composite variable is estimated in each of the NB-OPFS model components i.e., the composite score is included in the count and proportion model components. Alternatively, composite score can be treated as endogenous and be estimated simultaneously within the NB-OPFS model. In this approach, the estimates of the disaggregate model will be allowed to vary while modeling crash frequency. The second approach is computationally more involved as it allows for feedback between aggregate and disaggregate level models. The model fit measures such as Bayesian Information Criterion (BIC) can be employed to guide the model selection process.

The proposed model system is estimated using data drawn from the City of Orlando for the year 2019. The study considers a total of 300 traffic analysis zones for the aggregate level crash count by severity model. The disaggregate level model contains total 20,204 crash records from these zones. These records contain crash specific variables, temporal characteristics, roadway, vehicle and driver factors, road environmental, and weather information of each crash record. For aggregate level model analysis, an exhaustive set of independent variables including roadway and traffic factors, land-use attributes, built environment, and sociodemographic characteristics are considered.

3. METHODOLOGY

In this study, we employed negative binomial-ordered probit fractional split (NB-OPFS) model and an integrated modeling framework to analyze crash frequency by severity. However, for the sake of space, we will restrict ourselves to presenting the integrated framework only. Further,

within the integrated framework, there are two components: the disaggregate level model (ordered probit) and the aggregate level model (NB-OPFS). For the ease of presentation, we will discuss the methodology by each component.

3.1 Disaggregate Level Model Structure (Ordered Probit Model)

In the traditional ordered response model, the discrete injury severity levels (v_j) are assumed to be associated with an underlying continuous latent variable (v_j^*). This latent variable is typically specified as the following linear function:

$$v_j^* = X_j\Theta + \varepsilon_j, \text{ for } j = 1,2, \dots \dots \dots, \text{n} \quad (1)$$

where, j ($j = 1,2, \dots \dots \dots, n$) represents the crash record. X_j is a vector of exogenous variables (excluding a constant). Θ is a vector of unknown parameters to be estimated. ε_j is the random disturbance term assumed to be standard normal distribution. Let us assume k ($k = 1,2,3, \dots, k$) be the index to represent injury severity categories. In this study, k take the values of 'no-injury' ($k = 1$), 'possible injury' ($k = 2$), 'non-incapacitating injury' ($k = 3$) and 'fatal and incapacitating injury' ($k = 4$). t_k represents the thresholds associated with these severity levels. These unknown t_ks are assumed to partition the propensity into $k - 1$ intervals. The unobservable latent variable v_j^* is related to the observable ordinal variable v_j by the t_k with a response mechanism of the following form:

$$v_j = k, \ if \ t_{k-1} < v_j^* < t_k, \text{ for } k = 1,2, \dots \dots \dots, k \quad (2)$$

In order to ensure the well-defined intervals and natural ordering of observed severity, the thresholds are assumed to be ascending in order, such that $t_0 < t_1 < \dots \dots \dots < t_k$ where $t_0 = -\infty$ and $t_k = +\infty$. Given these relationships across the different parameters, the resulting probability expressions for record j and alternative k for the ordered probit take the following form:

$$\pi_{jk} = Pr(v_j = k|X_j) = \Upsilon(t_k - X_j\Theta) - \Upsilon(t_{k-1} - X_j\Theta) \quad (3)$$

where, $\Upsilon(.)$ represents the standard normal distribution function.

Considering the spatial arrangement of the crash records within the same zone, i.e., the adjacency heterogeneity (dependency), the equation for disaggregate level propensity can be updated as,

$$v_j^{*\prime} = X_j\Theta + \boldsymbol{\theta}_{ij} + \varepsilon_j, \text{ for } i = 1,2, \dots \dots \dots, \text{N} \quad (4)$$

where, i ($i = 1,2, \dots, N$) is the index for traffic analysis zone. $v_j^{*\prime}$ is the latent propensity capturing spatial dependency and $\boldsymbol{\theta}_{ij}$ is a vector of unobserved effects specific to the zone for crash records highlighting the spatial arrangement within the same zone. This $\boldsymbol{\theta}_{ij}$ will be same across the crash records if they correspond to same zone and thus, the adjacency heterogeneity (dependency) will be captured through the proposed system. The reader would note that the spatial unobserved heterogeneity can vary across the crash records. Therefore, in the current study, we parameterized the correlation parameter $\boldsymbol{\theta}_i$ as a function of observed attributes as follows:

$$\boldsymbol{\theta}_{ij} = \boldsymbol{\gamma}_{ij}\boldsymbol{s}_{ij} \quad (5)$$

where, $\boldsymbol{s}_{ij}$ is a vector of exogenous variables at the zonal level i (including a constant) employed for crash record j, $\boldsymbol{\gamma}_{ij}$ is a vector of parameters to be estimated. The model estimation process employs log-likelihood (LL) generated using the formula in equation 3 with the updated propensity from equation 4.

3.2 Aggregate Level Model Structure (NB-OPFS Model)

3.2.1 Count framework

In our study, the count framework estimates total number of crashes using the negative binomial model.

Once the disaggregate level propensities are estimated, we adopt two alternative approaches to estimate the aggregate (zonal) level propensities as presented in equation 6 and 7 respectively.

$$\mu_i = E(c_i|\boldsymbol{z}_i) = exp\left((\boldsymbol{\delta} + \boldsymbol{\zeta}_i)\boldsymbol{z}_i + \rho_c * \ln\left(\sum_{p=1}^{j_i}\left(exp\left(v_j^{*\prime}\right)\right)\right) + \varepsilon_i + \eta_i\right) \tag{6}$$

$$\mu_i = E(c_i|\boldsymbol{z}_i) = exp\left((\boldsymbol{\delta} + \boldsymbol{\zeta}_i)\boldsymbol{z}_i + \rho_c * \ln\left(\sum_{p=1}^{j_i}\left(exp\left(X_j\Theta + \boldsymbol{\theta}_{ij} + \varepsilon_j\right)\right)\right) + \varepsilon_i + \eta_i\right) \tag{7}$$

where, $\boldsymbol{z}_i$ is a vector of explanatory variables associated with zone i. $\boldsymbol{\delta}$ is a vector of coefficients to be estimated. $\boldsymbol{\zeta}_i$ is a vector of unobserved factors on crash count propensity for zone i and its associated zonal characteristics assumed to be a realization from standard normal distribution: $\boldsymbol{\zeta}_i \sim N(0, \boldsymbol{\pi}^2)$. ρ_c is a scalar associated with the disaggregate level highlighting the share of disaggregate level propensity to be linked with the aggregate level propensity for count component. p is a counter here ranging from 1 to j_i represents the crash record j in zone i. For example, if 5 crashes occurred in the zone i, then we will sum the propensity for these 5 crashes to obtain a value for j_i. The main difference between the two approaches is that the disaggregate level propensity will remain fixed and only the scalar parameter will be estimated for approach 1. In the second approach, we allow the disaggregate level parameters to be jointly influenced by disaggregate and aggregate fit. ε_i is a gamma distributed error term with mean 1 and variance α. η_i captures the influence of common unobserved factors that impact total number of crashes and proportion of crashes by severity for zone i.

For the count model, the equation system for modeling total crash count in the usual negative binomial formulation can be written as:

$$P(c_i) = \frac{\Gamma\left(c_i + \frac{1}{\alpha}\right)}{\Gamma(c_i + 1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(1 - \frac{1}{1 + \alpha\mu_i}\right)^{c_i} \quad (8)$$

where, c_i be the index for crashes occurring over a period of time in zone i. $P(c_i)$ is the probability that zone i has c_i number of crashes. $\Gamma(\cdot)$ is the gamma function, α is negative binomial overdispersion parameter and μ_i is the expected number of crashes occurring in zone i over a given time period (as presented in the equation 6 and equation 7).

3.2.2 Fractional split framework

The modeling of crash proportions by severity levels is undertaken using the ordered probit fractional split model. In the ordered outcome framework, the actual injury severity proportions (y_{ik}) are assumed to be associated with an underlying continuous latent variable (y_i^*). Following the same approach as presented in the negative binomial component, we adopt two alternative approaches to estimate latent propensity equation as follows:

$$y_i^* = \left((\boldsymbol{\beta} + \boldsymbol{\rho_i})\boldsymbol{x}_i + \rho_f * \ln\left(\sum_{p=1}^{j_i}\left(exp\left(v_j^{*\prime}\right)\right)\right) + \xi_i \pm \eta_i\right), \quad y_{ik} = k \; if \; \tau_{k-1} < y_i^* < \tau_k \quad (9)$$

$$y_i^* = \left((\boldsymbol{\beta} + \boldsymbol{\rho_i})\boldsymbol{x}_i + \rho_f * \ln\left(\sum_{p=1}^{j_i}\left(exp\left(X_j\Theta + \boldsymbol{\theta}_{ij} + \varepsilon_j\right)\right)\right) + \xi_i \pm \eta_i\right), y_{ik} = k \; if \; \tau_{k-1} < y_i^* < \tau_k \quad (10)$$

The latent propensity y_i^* is mapped to the actual severity proportion categories y_{ik} by τ thresholds $(\tau_0 = -\infty \; and \; \tau_K = +\infty)$ as presented in equation 9 and equation 10. $\boldsymbol{x}_i$ is a vector of attributes (not including a constant) that influences the propensity associated with severity proportion categories. $\boldsymbol{\beta}$ is the corresponding vector of mean effects. $\boldsymbol{\rho_i}$ is a vector of unobserved factors on severity proportion propensity for zone i and its associated zonal characteristics assumed to be a realization from standard normal distribution: $\boldsymbol{\rho} \sim N(0, \boldsymbol{\sigma}^2)$. ρ_f is a scalar associated with the disaggregate level highlighting the share of disaggregate level propensity to be linked with the aggregate level propensity for fractional split component. ξ_i is an idiosyncratic error term assumed to be identically and independently standard normally distributed across zone i. η_i term generates the correlation between equations for total number of crashes and crash proportions by severity levels and also allows for considering the influence of various unobserved factors affecting the frequency and proportion variables. The $\pm$ sign in front of η_i indicates that the correlation in unobserved individual factors between total crashes and crash proportions by severity levels may be positive or negative. A positive sign implies that zones with higher number of crashes are intrinsically more likely to incur higher proportions for severe crashes. On the other hand, negative sign implies that zones with higher number of crashes intrinsically incur lower proportions for severe crashes. To determine the appropriate sign one can empirically test the models with both

$'+'$ and $'-'$ signs independently. The model structure that offers the superior data fit is considered as the final model.

It is important to note here that the unobserved heterogeneity between total number of crashes and crash proportions by severity levels can vary across zones. Therefore, in the current study, the correlation parameter η_i is parameterized as a function of observed attributes as follows:

$$\eta_i = \boldsymbol{G}_i \boldsymbol{Q}_i \tag{11}$$

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

To estimate the model presented in equation 9 and equation 10, we assume that:

$$E(y_{ik}|\boldsymbol{x}_i) = H_{ik}(\beta, \tau),\ 0 \le H_{ik} \le 1, \sum_{k=1}^{K} H_{ik} = 1 \tag{12}$$

H_{ik} in our model takes the ordered probit probability (Λ) form for the severity category k.

Given these relationships across different parameters, the resulting probability (Λ) for the ordered probit fractional split model takes the following form:

$$\Lambda(y_{ik} = k) = \varphi\{\tau_k - (y_i^*)\} - \varphi\{\tau_{k-1} - (y_i^*)\} \tag{13}$$

where, $\varphi(\cdot)$ is the standard normal cumulative distribution function.

3.3 Model Estimation

In examining the model structure of total crash count and proportions of crashes by severity levels, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\zeta}, \boldsymbol{\rho}, \boldsymbol{G}$ and $\boldsymbol{\gamma}$ represented by Ω. In this study, it is assumed that these elements are drawn from independent realization from normal population: $\Omega \sim N(0, (\boldsymbol{\pi^2}, \boldsymbol{\sigma^2}, \mathrm{g}^2, \boldsymbol{\vartheta}^2))$. Thus, conditional on Ω, the likelihood function for the integrated probability can be expressed as:

$$L_i = \int_{\Omega} P(c_i) \times \prod_{k=1}^{K} \left(\Lambda(y_{ik} = k)\right)^{\varpi_i d_{ik}} \times \prod_{p=1}^{j_i} \prod_{k=1}^{K} \pi_{jk}\, d\Omega \tag{14}$$

where, ϖ_i is a dummy with $\varpi_i = 1$ if zone i has at least one crash over the study period and 0 otherwise. d_{ik} is the proportion of crashes in severity category k. Finally, the log-likelihood function is:

$$LL = \sum_i Ln(L_i) \tag{15}$$

All the parameters in the model are estimated by maximizing the logarithmic function LL presented in equation 15. The parameters to be estimated in the model are: $\boldsymbol{\delta}$, α, $\boldsymbol{\beta}$, τ, ρ_c, ρ_f, $\boldsymbol{\pi}$, $\boldsymbol{\sigma}, \mathrm{g}, \boldsymbol{\vartheta}$, t and Θ. To estimate the proposed model, we apply Quasi-Monte Carlo simulation

techniques based on the scrambled Halton sequence to approximate this integral in the likelihood function and maximize the logarithm of the resulting simulated likelihood function across individuals (please see Bhat, 2001; Yasmin and Eluru, 2013 for details). We use the GAUSS matrix programming software to run the models (Aptech, 2015).

4. DATA PREPARATION

The current research employs 2019 data from Orlando, Florida composed of 300 traffic analysis zones with a total of 20,204 crash records. At the crash level, crash specific variables (such as crash types, first harmful events), temporal characteristics (such as time of the day, seasons), roadway factors (such as location of the crashes, speed limit, shoulder type), vehicle factors (such as presence of passengers), driver factors (such as driving under influence related, distraction related), road environmental factors and weather information (such as clear, rain, fog) are considered. These crashes could be classified into 5 categories by crash severity outcomes: fatal, incapacitating injury, non-incapacitating injury, possible injury, and no-injury crashes. The distribution of crashes by severity is 0.30% fatal, 1.75% incapacitating, 8.84% non-incapacitating, 18.62% possible injury, and 70.49% no-injury crashes. Since the reported number of fatal crashes is very low, this study combines fatal and incapacitating injury crashes as fatal + incapacitating crashes for disaggregate level model estimation. For aggregate level model, four severity levels are considered and the dependent variable for fractional split component (represented as OPFS model) can be represented as proportions (number of specific severity level/total number of all crashes) as follows: (1) proportion of no-injury crashes (2) proportion of possible injury crashes (3) proportion of non-incapacitating injury crashes and (4) proportion of fatal and incapacitating injury crashes. A comprehensive set of independent variables including roadway, traffic, land-use, built environment, and sociodemographic characteristics are considered in this study. This study selects 280 traffic analysis zones randomly for model estimation resulting in a crash record sample of 18,286 crash records. The remaining 20 zones and 1,918 crash records are set aside for the validation of the models.

4.1 Variables Considered

The variables for disaggregate and aggregate analysis were collected from different data sources including Signal Four Analytics (S4A), Florida Department of Transportation (FDOT) Transportation Statistics Division, US Census Bureau and American Community Survey, and Florida Geographic Data Library databases. These explanatory variables were aggregated at the zonal level using the ArcGIS for aggregate level dataset. For example, annual average daily traffic (AADT) of a zone is obtained by computing a weighted average of AADT values across the road facilities in the zone. Aggregate level analyses use roadway and traffic factors (such as proportion of roads by functional class, number of lanes, average speed limit, average shoulder width, average sidewalk width and median width, intersection density, traffic signal density, AADT, and truck AADT), land-use attributes (such as proportion of residential, commercial, institutional, industrial, recreational and mixed areas), built environment characteristics (such as number of restaurants, business centers, commercial centers, educational centers, and shopping centers), and sociodemographic characteristics (such as population density, proportion of males and females, household density, median household income, proportion of car, drive alone, non-motorized means of transport, different population group by age level, household with vehicle availability, and population with different races). Land-use mix is defined as: $\left[-\frac{\sum(p_l(\ln p_l))}{\ln R}\right]$, where l is the category of land-use, p_l is the proportion of the developed land area devoted to a specific land-use l, R is

the number of land-use categories in an analysis zone. In our study, five land-use types were considered including residential, industrial, institutional, commercial (including office areas) and recreational areas. The value of this index ranges from zero to one - zero (no mix) corresponds to a homogenous area characterized by single land-use type and one to a perfectly heterogeneous mix.

In estimating the model, several functional forms, and combination of variables are considered and those that provide the best fit are retained in the final specification. The final specification of the model was based on removing the statistically insignificant variables in a systematic process based on 90% confidence level. Figure 1 shows the sample share of the variables at disaggregate level considered for the final model estimation while the aggregate level variables are presented in Table 1 with the appropriate definition and summary statistics.

5. EMPIRICAL ANALYSIS

5.1 Model Specification and Overall Measure of Fit

Several models are estimated for the empirical analysis of the proposed framework. First, we estimated the ordered probit model for disaggregate level severity analysis and negative binomial-ordered probit fractional split (NB-OPFS) model for aggregate level crash count by severity. Second, we developed our proposed integrated model system following two approaches: a) integrated aggregate and disaggregate model 1 (IADM1): focusing on optimizing the joint log-likelihood of the aggregate and disaggregate level models by only estimating the parameter for propensity aggregated from the disaggregate model (one parameter per model component i.e., count and proportion) as shown in equations 6 and 9, and b) integrated aggregate and disaggregate model 2 (IADM2): the disaggregate level parameters are estimated based on their contribution to the disaggregate level and the aggregate level models through the disaggregate level propensity component embedded within the aggregate level propensity equation (as shown in equations 7 and 10). Third, we identify the best model by comparing model performance based on Bayesian Information Criterion (BIC). The BIC for a given empirical model is equal to:

$$BIC = -2LL + N_p \ln(O) \tag{16}$$

where, LL is the log-likelihood value at convergence, N_p is the number of parameters and O is the number of observations. The model with the lower BIC is the preferred model.

The corresponding BIC (LL) values are: (1) non-integrated model (ordered probit and NB-OPFS) (with 45 parameters): 32,836.046 (-16,291.240), (2) IADM1 (with 41 parameters): 32,684.538 (-16,226.756), (3) IADM2 (with 44 parameters): 32,700.379 (-16,226.224) and (4) IADM1 with unobserved heterogeneity (with 44 parameters): 32,646.507 (-16,199.288). Based on these BIC values, two specific observations could be drawn. First, all the integrated systems provide improved data fit as evidenced by the lower BIC values in comparison to the non-integrated model. Second, within the integrated systems, our proposed IADM1 provides the lowest BIC indicating the best data fit in comparison to the proposed IADM2. Finally, we accommodate additional unobserved heterogeneity ($\boldsymbol{\theta}_{ij}$) in our IADM1 (the best model in terms of data fit) and find that the model accommodating this additional unobserved heterogeneity provides further improved BIC (lower) compared to the IADM1 framework without $\boldsymbol{\theta}_{ij}$ elements.

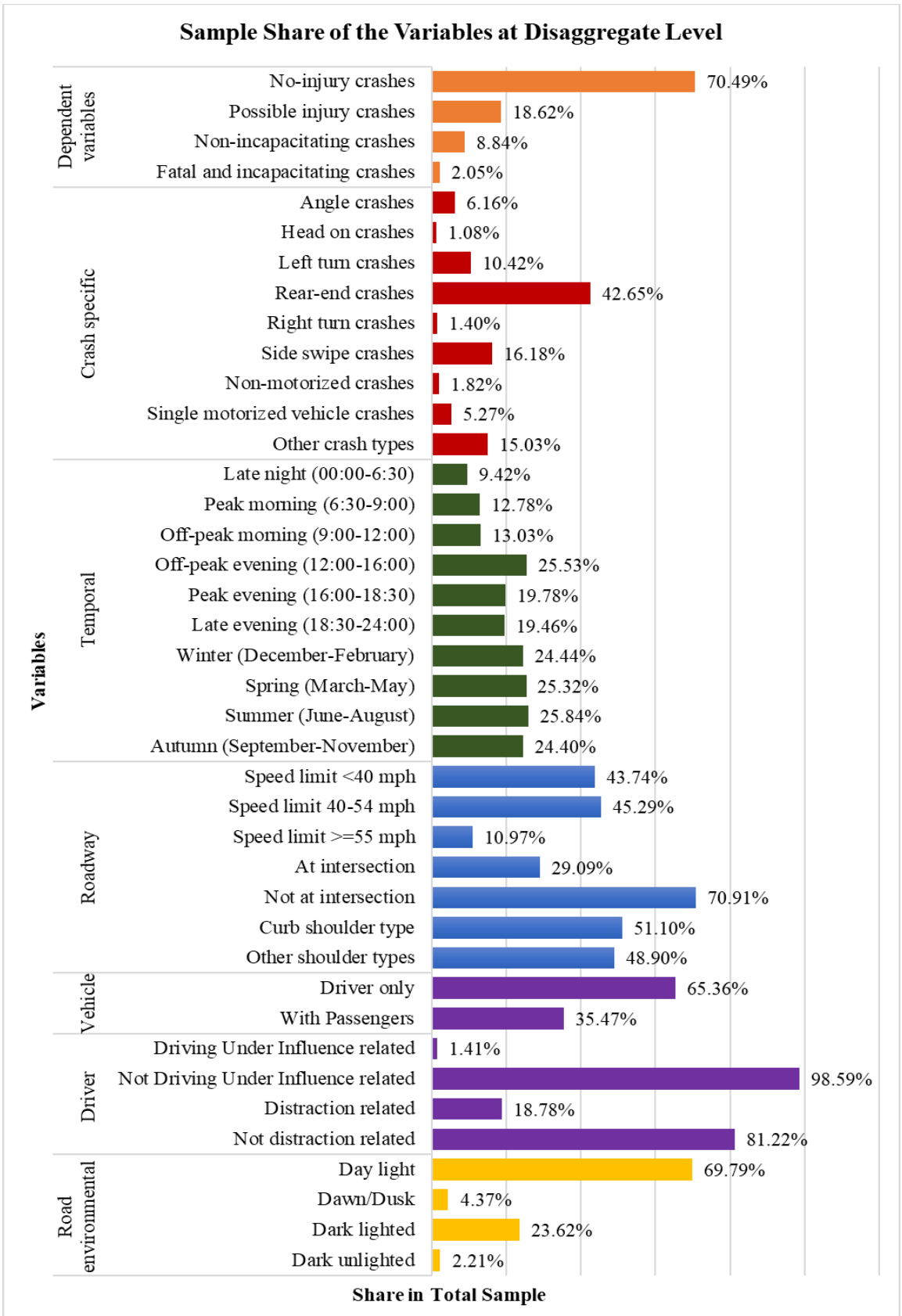

Figure 1: Sample Share of the Variables at Disaggregate Level (n = 20,204)

Table 1: Summary Statistics of the Variables at Aggregate Level (N = 300)

Variables	Definition	Min	Max	Mean	SD
Dependent variables					
Total crashes	Total number of crashes in zone	0.000	292.000	67.347	56.300
Proportion of fatal and incapacitating injury crashes	Total fatal and incapacitating injury crashes/total crashes	0.000	0.333	0.022	0.030
Proportion of non-incapacitating injury crashes	Total non-incapacitating injury crashes/total crashes	0.000	0.333	0.090	0.053
Proportion of possible injury crashes	Total possible injury crashes/total crashes	0.000	1.000	0.187	0.095
Proportion of no-injury crashes	Total no-injury crashes/total crashes	0.000	1.000	0.698	0.116
Roadway characteristics					
Road density	Total road length in mile/total area of zone in sq. mile	0.000	27.895	4.447	4.341
Average sidewalk width	Ln (Average sidewalk width, feet + 1)	0.000	2.646	1.775	0.506
Average inside shoulder width	Average inside shoulder width, feet	0.000	18.000	3.008	3.742
Proportion of <40 mph roads	Total road length with speed limit <40 mph/total road length in zone	0.000	1.000	0.496	0.397
Intersection density	Number of intersections/total area of zone in acre	0.000	0.770	0.085	0.115
Traffic signal density	Number of traffic signals/total number of intersections in zone	0.000	1.000	0.058	0.106
Proportion of >=3 lane roads	Total length of >=3 lane roads/total road length in zone	0.000	1.000	0.231	0.274
Proportion of divided roads	Total divided road length/total road length in zone	0.000	1.262	0.610	0.357
Traffic characteristics					
AADT	Ln (AADT of zone + 1)	0.000	13.507	11.189	1.864
Proportion of heavy vehicles	Total truck AADT/total AADT	0.000	0.170	0.056	0.024
Land-use characteristics					
Proportion of residential areas	Residential areas/total land-use areas	0.000	0.998	0.490	0.350
Proportion of commercial areas	Commercial areas/total land-use areas	0.000	1.000	0.242	0.274
Land-use mix	Mixed land-use areas/total land-use areas	0.000	0.957	0.418	0.242
Built environment characteristics					
Number of restaurants	*Z score: Number of restaurants	-0.597	6.690	0.000	1.000
Number of educational centers	Z score: Number of educational centers	-0.649	3.879	0.000	1.000
Sociodemographic factors					
Household density	Number of households/total area of zone in acre	0.084	8.621	2.016	1.574
Non-motorized means of transport	Ln (Non-motorized means of transport + 1)	0.000	5.366	2.152	1.166
Proportion of African American population	Total African American population /total population in zone	0.000	0.978	0.222	0.246

**For the built environment characteristics, the results provided a superior fit for the standardized form (represented by Z-scores) of the actual variables than count, and number per area or density forms.*

5.2 Model Estimation Results

This section provides a brief description of the factors affecting crash count by severity at aggregate level as well as factors influencing crash severities at the disaggregate level model. For the sake of brevity, the results of the proposed integrated aggregate and disaggregate model 1 (IADM1) with unobserved heterogeneity are discussed in this section. Table 2 presents the model estimation results for the proposed model. The reader would note that a positive (negative) sign for a variable in Table 2 indicates that an increase in the variable is likely to result in more (less) crashes as well as exhibit a higher (lower) impact on severity. The results of the non-integrated ordered probit and negative binomial-ordered probit fractional split models are presented in Table A1 in the Appendix.

Table 2: Estimation Results of the Proposed IADM1 Considering Unobserved Heterogeneity (N = 280 Traffic Analysis Zones with 18,286 Crash Records)

Parameters	Estimates	t-stat
Disaggregate level		
Threshold between NI-PI	0.809	31.702
Threshold between PI-NII	1.576	58.048
Threshold between NII-FII	2.479	72.259
Crash type (Base: Rear-end and other crash types)		
Angle	0.303	7.735
Head on	0.478	5.385
Left turn	0.305	9.346
Right turn	-0.243	-2.775
Sideswipe	-0.553	-16.725
Non-motorized	1.705	31.014
Single motorized vehicle	0.253	5.557
Time (Base: Peak morning, off-peak morning, off-peak evening and late evening)		
Late night (00:00 to 6:30)	0.103	2.713
Peak evening (16:00 to 18:30)	-0.053	-2.208
Season (Base: Summer and Autumn)		
Spring (March to May)	-0.049	-2.097
Winter (December to February)	-0.040	-1.715
Speed limit (Base: Speed limit < 40 mph)		
Speed limit 40-54 mph	0.091	4.485
Speed limit >=55 mph	0.116	3.398
Location (Base: Not at intersection)		
Intersection	0.095	3.972
Shoulder type (Base: Other shoulder types)		
Curb shoulder type	-0.036	-1.815
Presence of passengers (Base: Driver only)		
With passenger	0.338	17.199
Driving under influence (Base: Not DUI related)		
DUI related	0.470	6.179
Distraction (Base: Not distraction related)		
Distraction related	0.234	10.019
Lighting condition (Base: Daylight and dawn/dusk)		
Dark lighted	0.066	2.564
Dark not lighted	0.169	2.542
Aggregate level		
Count component		

Parameters	Estimates	t-stat
Constant	-0.326	-6.444
Average sidewalk width	-0.032	-1.999
Intersection density	0.204	3.150
Traffic signal density	0.119	2.267
Proportion of commercial areas	0.079	2.975
Household density	0.009	1.734
Overdispersion parameter	0.697	5.229
Proportion component		
Threshold between NI-PI	0.605	2.984
Threshold between PI-NII	1.303	6.622
Threshold between NII-FII	2.111	10.518
Average inside shoulder width	-0.018	-4.098
Proportion of <40 mph roads	-0.199	-4.266
Traffic signal density	-0.222	-1.901
Proportion of heavy vehicles	2.030	1.931
Proportion of commercial areas	-0.116	-1.655
Number of restaurants	-0.052	-2.461
Parameter for disaggregate level model propensity		
Propensity in the count component	0.993	81.483
Propensity in the proportion component	0.037	0.827
Unobserved heterogeneity		
Constant	0.109	6.749
Road density	0.010	3.570
Land-use mix	0.030	1.693
BIC	32,646.507	
Log-likelihood	-16,199.288	
Number of parameters	44	

Note: NI = no-injury crashes, PI = possible injury crashes, NII = non-incapacitating injury crashes, and FII = fatal and incapacitating injury crashes.

5.3 Crash Specific Constant

The model constant in the count component does not have any substantive interpretation.

5.4 Disaggregate Level Attributes

The results of the proposed model show that among the disaggregate level crash specific variables, angle, head on, left turn, non-motorized and single motorized vehicle crash types have positive impact on crash severity while right turn and sideswipe crash types have negative impact compared to the rear-end and other crash types. These results are consistent with many previous studies (Abdel-Aty and Keller, 2005; Danesh et al., 2022; Marcoux et al., 2018; Wang and Kim, 2019; Yasmin and Eluru, 2013; Zeng et al., 2019). The findings are quite intuitive as in the case of angle crashes and head on crashes, the dissipation of kinetic energy and deformation of motor vehicle bodies are greater resulting in severe consequences. Left turn crashes generally occur while drivers tend to make left maneuvers. The severity of these crashes is also higher due to the greater force of impact exerted while colliding with oncoming vehicles and a similar result was found in other studies (Abdel-Aty and Keller, 2005). Further, non-motorized involved crashes, such as pedestrian and bicycle crashes, are severe in nature as these users are more vulnerable on roadways. In addition, single motorized vehicle crashes, such as roll over and run-off-road crashes are likely to result in severe crashes compared to the rear-end and other crash types (Yasmin and Eluru, 2013). On the contrary, right turn crashes and sideswipe crashes are likely to be less severe as a lower

amount of kinetic energy is dissipated within the vehicles due to the direction of impact force during these crashes (see Abdel-Aty and Keller, 2005 for similar findings).

Among the temporal factors, late nighttime has a positive impact on crash severity while evening peak has a negative impact (compared to other times of the day). This is plausible as the volume of traffic is low in the late night and vehicle operating speeds are higher (Marcoux et al., 2018). During peak hours, the higher traffic volume and lower speeds reduces the probability of severe crashes. In addition, spring and winter season have a lower impact on crash severity compared to summer and autumn. The findings are similar to those reported in other studies (Zeng et al., 2019).

The model results clearly showed that crashes occurred at road sections with speed limit 40-54 mph and >= 55 mph have a higher probability of severe crashes compared to the road section with speed limit less than 40 mph. This is quite intuitive as the operating speeds are also higher at the road sections with higher speed limit (K. Wang et al., 2019; Wang and Kim, 2019; Yasmin et al., 2014; Yasmin and Eluru, 2013). The findings, as expected, indicate that crashes at intersections are more severe than segment locations. Further, the curb shoulder type has a negative impact on crash severity as this shoulder type provides additional safety by reducing vehicle speed (Jiang et al., 2013).

The results demonstrate that the presence of passengers in the vehicle, driving under influence and distracted conditions have a positive impact on crash severity. These results are expected and have been documented in previous work (Das et al., 2009; Marcoux et al., 2018; Paleti et al., 2010; X. Wang et al., 2019; Weiss et al., 2014; Yasmin and Eluru, 2013).

The model results also show that compared to the daylight and dawn/dusk conditions, dark conditions irrespective of light have a positive impact on severity. This is because dark conditions often reduce visibility and increase reaction time on the roads (see Marcoux et al., 2018; Wang and Kim, 2019 for similar findings).

5.5 Aggregate Level Attributes

In the aggregate level count component, wider sidewalk has a negative impact on likelihood of crashes while intersection density and traffic signal density have positive impacts. This is plausible as the presence of wider sidewalks provides additional margin for error and thus contribute to a lower risk (Bhowmik et al., 2019). Higher number of intersections in a zone have higher traffic conflicts increasing crash risk. Along the same line, higher number of signalized intersections lead to increase in crashes (Wang and Huang, 2016).

Among the land-use and sociodemographic attributes, proportion of commercial areas and household density show positive impacts on crash count. These findings are intuitive as commercial areas have commerce related activities such as loading/unloading, movement of heavy vehicles and increased traffic conflicts that might contribute to higher crash risk (Cui and Xie, 2021; Mohammadnazar et al., 2021; Soroori et al., 2019; Xie et al., 2019). In the case of household density, as the density increases, traffic is likely to increase and contribute to additional crash risk (Yasmin and Eluru, 2018).

In the proportion component, wider inside shoulder width, proportion of roads with <40 mph speed limit and traffic signal density indicate a negative impact on crash severity. This is intuitive as wider shoulder provides additional safety margin on the road, thus, contributing to reduced severity (see Chen et al., 2017). As described earlier, it is evident that roads with lower speed limit have lower probability of severe crashes (Afghari et al., 2020; Yasmin and Eluru, 2018). Further, higher traffic signal density is associated with lower severity (Bhowmik et al.,

2021b). The results indicate that a higher proportion of heavy vehicles increases the severity of crashes as found in many studies (Yasmin and Eluru, 2018).

Among land-use attributes, higher proportion of commercial areas reduces the severity of crashes. The presence of more restaurants in a zone reduces the severity of the crashes. Similar findings are reported in previous studies (Yasmin and Eluru, 2018).

The coefficients for the fixed propensity from the disaggregate model in the count component and severity component are presented in the lower row panel of Table 2. The results indicate that an increased disaggregate level model propensity is associated with increased total crash count at the aggregate level. On the other hand, the results indicate that disaggregate level propensity is not significantly impacting the proportion of crashes by severity at 90% confidence level. The coefficient for severity component while insignificant was retained to show the direction of effect. The positive sign for both parameters indicates that a higher value of disaggregate level model propensity is likely to increase the number of total crashes and the crash severity i.e., higher propensity for severe crashes at the disaggregate level is directly associated with an increased number of total and severe crashes at the aggregate level.

5.6 Unobserved Heterogeneity

The proposed model system can capture unobserved heterogeneity in the form of spatial variations for crash records in a zone through a common spatial correlation between all crash records from a zone. The unobserved heterogeneity variable constant presented in Table 2 corresponds to this common zone spatial correlation. The significant effect of this parameterized correlation parameter ($\boldsymbol{\theta}_{ij}$) clearly highlights the presence of common unobserved factors across crash records in the same zone. Capturing this common spatial correlation across the crash records is an important contribution of this study. The correlation was also parameterized and tested for several independent variables. In our testing, we found two zonal variables – road density and land-use mix - exhibit significant unobserved correlation across crash records within the same zone. Further, we attempted to capture the unobserved correlation (η_i) between total crashes and crash proportions by severity levels, and random parameter effects ($\boldsymbol{\zeta}$ and $\boldsymbol{\rho}$) in our proposed model system. However, among these parameters tested no statistically significant effect was recovered in our dataset.

5.7 Predictive Performance of the Model

To demonstrate the applicability of the model, we undertake a comparison exercise between the proposed integrated aggregate and disaggregate model 1 (IADM1) and the negative binomial-ordered probit fractional split (NB-OPFS) model by testing model performance on estimation and holdout samples. We compare the models by employing three different measures of fit: mean prediction bias (MPB), mean absolute deviation (MAD), and mean squared prediction error (MSPE). MPB represents the magnitude and direction of average bias in model prediction. The model with the lower MPB provides better prediction of the observed data and is computed as:

$$\text{MPB} = \ mean\ (\hat{y}_i - y_i) \tag{17}$$

where, $\hat{y}_i$ and y_i are the predicted and observed number of crashes occurring over a period of time in a zone i. 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:

$$\text{MAD} = \ mean\ |\hat{y}_i - y_i| \tag{18}$$

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

$$\text{MSPE} = \ mean\ (\hat{y}_i - y_i)^2 \tag{19}$$

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

Table 3 presents the values of these measures for the proposed integrated model and NB-OPFS model. The results clearly highlight that the proposed integrated model performs better than NB-OPFS model across all fit measures computed for both estimation and validation datasets.

Table 3: Predictive Performance of the Models

Dataset	Models	Measures	NI	PI	NII	FII	Total
Estimation (N=280 traffic analysis zones with 18,286 records)	IADM1 with unobserved heterogeneity	MPB	-0.429	0.028	0.152	0.079	-0.169
	NB-OPFS		-0.459	-0.168	0.002	0.031	-0.594
	IADM1 with unobserved heterogeneity	MAD	6.400	2.837	1.923	0.972	5.718
	NB-OPFS		20.957	6.010	3.205	1.191	29.654
	IADM1 with unobserved heterogeneity	MSPE	88.783	15.950	7.056	1.942	75.567
	NB-OPFS		859.650	69.867	18.628	2.935	1,631.054
Validation (N=20 traffic analysis zones with 1,918 records)	IADM1 with unobserved heterogeneity	MPB	-1.972	0.509	-1.058	0.042	-2.479
	NB-OPFS		-21.109	-3.606	-2.784	-0.301	-27.800
	IADM1 with unobserved heterogeneity	MAD	11.211	3.007	3.148	1.156	11.275
	NB-OPFS		42.281	9.388	6.440	1.421	57.772
	IADM1 with unobserved heterogeneity	MSPE	424.191	16.800	16.387	3.019	391.580
	NB-OPFS		3,961.385	184.153	63.468	4.195	6,915.558

Note: NI = no-injury crashes, PI = possible injury crashes, NII = non-incapacitating injury crashes, and FII = fatal and incapacitating injury crashes.

The values of the measures shown in Table 3 demonstrate the model performance at the aggregate level. To further evaluate the predictive performance of the estimated models, we carried out a comparison exercise between the proposed IADM1 with unobserved heterogeneity and the NB-OPFS model by comparing absolute deviation from observed values across the zones in our study region (see Figure 2). The heat maps show that our proposed model exhibits lower predictive deviation values compared to the NB-OPFS model across the study region. The mean and the distribution of the deviation values further highlight the superiority of the proposed model in both estimation and validation samples.

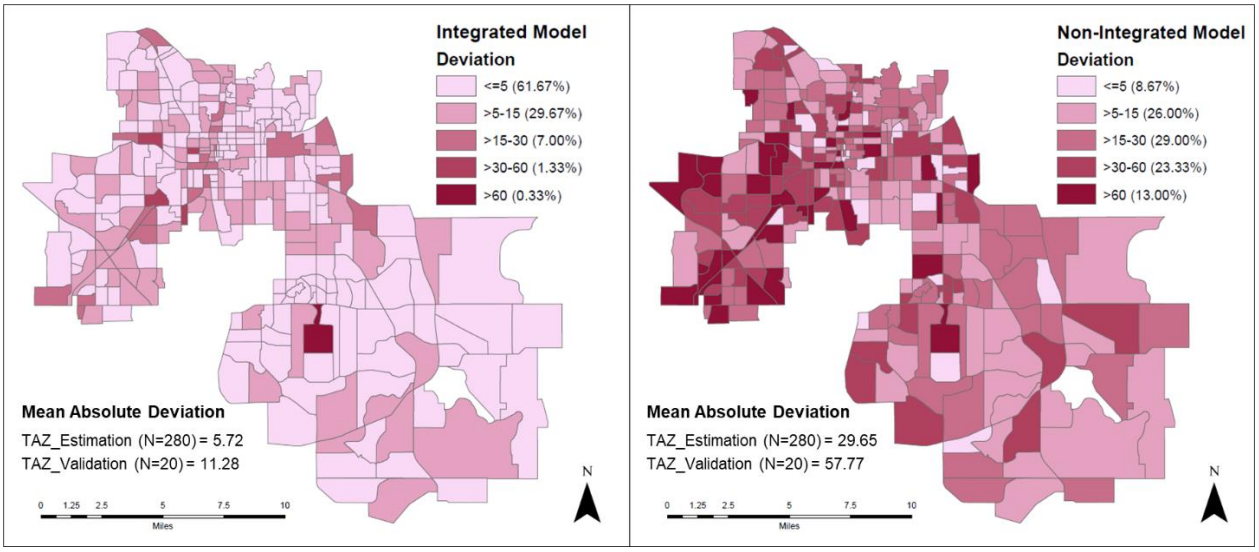

Figure 2: Distribution of the Model Deviations across Traffic Analysis Zones (TAZs): Proposed IADM1 Model (Left) and Non-integrated (NB-OPFS) Model (Right)

5.8 Elasticity Effects

The model estimates presented in Table 2 represent a complex joint interaction of aggregate and disaggregate level models and do not directly provide the magnitude of the effects of the variables on crash counts. The variable impact magnitudes can be obtained by computing elasticity effects of those variables. As the primary focus of our analysis is on illustrating how disaggregate level variables contribute to aggregate level model, we focus on the elasticity effects of disaggregate level variables by following the procedure demonstrated in Eluru and Bhat (2007) (Eluru and Bhat, 2007). By this procedure, the percentage changes in the expected total zonal crash counts by severity caused by the change in the disaggregate level exogenous variable were computed. As all the exogenous variables in the disaggregate level are indicator variables, we obtain these changes by changing the value of the variable to one for the subsample of observations for which the variable takes a value of zero and to zero for the subsample of observations for which the variable takes a value of one. Specifically, the elasticity effects computed in this procedure are aggregated percentage elasticity based on the aggregated change and the overall shares of the sample.

The computed elasticities are presented in Table 4. The results show the percentage change in the number of crashes by different severities due to the changes in the disaggregate level exogenous variable of interest. For example, the elasticity estimate for the driving under influence related variable indicates that driving under influence increases a total crash by 58.74%. The effects of all the variables presented in Table 4 can be interpreted in similar fashion. By analyzing these effects, several observations can be drawn. First, there are differences in the elasticity effects across the expected number of crashes for different severities. Second, the most significant variables with respect to an increase in the expected number of total crashes are non-motorist involved crash types, head on crash type, driving under influence related, driving with passenger, angle, left turn and single motorized vehicle crash types, distraction related, and dark conditions. Third, the most significant variables with respect to an increase in the expected number of fatal and incapacitating injury crashes are non-motorist involved crash types, head on crash type, driving under influence related, driving with passenger, angle, left turn and single motorized vehicle crash types, distraction related, dark conditions and road section with speed limit >=

55mph. Finally, the elasticity effects demonstrate that the influence of crash specific and driver related variables is substantially larger than the influence of temporal and roadway characteristics.

Table 4: Aggregate Elasticity Effect for Disaggregate Level Variables

Parameters	%Total	%NI	%PI	%NII	%FII
Crash type (Base: Rear-end and other crash types)					
Angle	34.35	33.62	35.55	36.58	37.85
Head on	60.32	58.92	62.63	64.62	67.08
Left turn	34.08	33.38	35.26	36.25	37.48
Right turn	-21.48	-21.13	-22.07	-22.55	-23.15
Sideswipe	-44.83	-44.17	-45.92	-46.83	-47.92
Non-motorized	412.00	396.80	436.58	459.00	487.74
Single motorized vehicle	28.19	27.61	29.16	30.00	31.00
Time (Base: Peak morning, off-peak morning, off-peak evening and late evening)					
Late night (00:00 to 6:30)	10.69	10.48	11.03	11.32	11.67
Peak evening (16:00 to 18:30)	-5.19	-5.10	-5.35	-5.48	-5.64
Season (Base: Summer and Autumn)					
Spring (March to May)	-4.82	-4.73	-5.00	-5.09	-5.24
Winter (December to February)	-3.91	-3.84	-4.03	-4.13	-4.26
Speed limit (Base: Speed limit < 40 mph)					
Speed limit 40-54 mph	9.04	8.88	9.31	9.54	9.81
Speed limit >=55 mph	12.10	11.87	12.50	12.82	13.23
Intersection (Base: Not at intersection)	9.58	9.40	9.88	10.13	10.44
Curb shoulder (Base: Other shoulder types)	-3.53	-3.47	-3.64	-3.73	-3.84
With passenger (Base: Driver only)	35.16	34.47	36.31	37.28	38.47
Driving under influence (DUI) related (Base: Not DUI related)	58.74	57.39	61.00	62.88	65.26
Distraction related (Base: Not distraction related)	24.93	24.45	25.74	26.41	27.24
Lighting condition (Base: Daylight and dawn/dusk)					
Dark lighted	6.70	6.58	6.92	7.09	7.31
Dark not lighted	18.13	17.77	18.74	19.24	19.86

Note: NI = no-injury crashes, PI = possible injury crashes, NII = non-incapacitating injury crashes, and FII = fatal and incapacitating injury crashes.

These elasticity effect results can contribute to improving road safety. For instance, results indicate that crash types such as angle, head on, left turn, non-motorized and single motorized vehicle should be considered during countermeasure development. Effective traffic signals, medians, facilities for non-motorists and roadside guidepost/guard rail could be some suitable solutions for mitigating these crash types. In addition, results indicate that driving at higher operating speeds, especially on road sections with posted speed limit >= 55mph, driving with passengers, under influence, distracted condition and dark conditions should be considered as serious concerns. Strategies such as continuous monitoring and targeted enforcement, road safety awareness campaigns, traffic education, roadway lighting improvement and maintenance should be accelerated in the zones with over-speeding, driving under influence incidence, and distraction. Overall, the elasticity analysis demonstrates how the proposed model can be applied to determine critical disaggregate level factors contributing to the increase of the total crashes and crash severity.

6. CONCLUSIONS

The independent modeling approaches for crash frequency and severity models do not allow for interaction of variable impacts across the two systems. However, crashes employed in the frequency models are aggregated from the disaggregate level crash records. To bridge the gap between these two model systems, the current research proposes an integrated model framework that allows for the influence of disaggregate level variables within the aggregate level propensity estimation. The approach would involve summing up the crash propensity of each disaggregate level severity record within the aggregate resolution and adding the generated value as a new variable in the aggregate model. In this study, we employed a negative binomial-ordered probit fractional split (NB-OPFS) framework at the aggregate model to examine crash frequency by severity and the ordered probit model at the disaggregate model to examine the crash severity. In the NB-OPFS framework, the negative binomial component models the total number of crashes and the ordered probit fractional split component determines the proportion of each severity at a zone. These models were utilized by following two approaches of the proposed integrated framework. In the first approach, the ordered probit model propensity across the crashes in the zone is computed as a composite score and treated as an exogenous variable. In this approach, an additional parameter per component (each for count and proportion component) for the composite variable is estimated. In the second approach, the composite score of the ordered probit model propensity is treated as endogenous, allowed to vary, and estimated simultaneously within the NB-OPFS model. The empirical analysis was conducted using crash data drawn from the City of Orlando for the year 2019. The study considered a total of 20,204 crash records from 300 traffic analysis zones.

A series of models was estimated for the empirical analysis of the proposed framework, including ordered probit model for disaggregate level severity analysis, NB-OPFS model for aggregate level crash frequency by severity, integrated aggregate and disaggregate model 1 (IADM1) and integrated aggregate and disaggregate model 2 (IADM2). The model selection exercise was conducted based on the Bayesian Information Criterion (BIC) value. The results clearly highlighted the improved performance of the proposed integrated models over non-integrated model system. Within the integrated model approaches, IADM1 outperforms the IADM2 in terms of BIC value. Finally, we accommodated unobserved heterogeneity in the IADM1 and found that accommodating unobserved heterogeneity provides further improved BIC (lower) value. We also compared the performance of the proposed integrated model with the non-integrated model system by using several predictive performance measures. The measures also clearly highlighted the superiority of our proposed integrated model over the non-integrated model system. Further, an elasticity exercise was conducted to illustrate how the influence of disaggregate level variables on crash frequency dimensions can be examined.

The study is not without limitations. The proposed integrated approach requires substantial effort for data compilation in the region. The compilation can also be cumbersome as data from various sources are needed leading to the handling of large datasets and substantial data processing resources. Additionally, future research efforts can explore the efficacy of the proposed model examining crash frequency by crash type and severity simultaneously.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge the Signal Four Analytics (S4A), Florida Department of Transportation (FDOT) and other data sources for providing access to Florida crash and geospatial data.

AUTHOR CONTRIBUTION STATEMENT

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

7. REFERENCES

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accident Analysis and Prevention 37(3), 417–425.

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

Ahmed, S.S., Alnawmasi, N., Anastasopoulos, P.Ch., Mannering, F., 2022. The effect of higher speed limits on crash-injury severity rates: A correlated random parameters bivariate tobit approach. Analytic Methods in Accident Research 34, 100213.

Ahmed, S.S., Corman, F., Anastasopoulos, P.Ch., 2023. Accounting for unobserved heterogeneity and spatial instability in the analysis of crash injury-severity at highway-rail grade crossings: A random parameters with heterogeneity in the means and variances approach. Analytic Methods in Accident Research 37, 100250.

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

Anastasopoulos, P.Ch., Mannering, F.L., Shankar, V.N., Haddock, J.E., 2012a. A study of factors affecting highway accident rates using the random-parameters tobit model. Accident Analysis and Prevention 45, 628–633.

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

Aptech, 2015. Aptech Systems Inc.

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete choice analysis: theory and application to travel demand. MIT press.

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., Sen, S., Eluru, N., 2009. The impact of demographics, built environment attributes, vehicle characteristics, and gasoline prices on household vehicle holdings and use. Transportation Research Part B 43(1), 1–18.

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

Bhowmik, T., Yasmin, S., Eluru, N., 2021b. A New Econometric Approach for Modeling Several Count Variables: A Case Study of Crash Frequency Analysis by Crash Type and Severity. Transportation Research Part B 153, 172–203.

Bhowmik, T., Yasmin, S., Eluru, N., 2019. Do we need multivariate modeling approaches to model crash frequency by crash types? A panel mixed approach to modeling crash frequency by crash types. Analytic Methods in Accident Research 24, 100107.

Cai, Q., Abdel-Aty, M., Lee, J., Huang, H., 2019. Integrating macro- and micro-level safety analyses: a Bayesian approach incorporating spatial interaction. Transportmetrica A 15(2), 285–306.

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

Cui, H., Xie, K., 2021. An accelerated hierarchical Bayesian crash frequency model with accommodation of spatiotemporal interactions. Accident Analysis and Prevention 153, 106018.

Danesh, A., Ehsani, M., Moghadas Nejad, F., Zakeri, H., 2022. Prediction model of crash severity in imbalanced dataset using data leveling methods and metaheuristic optimization algorithms. International Journal of Crashworthiness, 27(6), 1869-1882.

Das, A., Abdel-Aty, M., Pande, A., 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. Journal of Safety Research 40(4), 317–327.

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

Eluru, N., Pinjari, A., Pendyala, R., Bhat, C., 2010. An econometric multi-dimensional choice model of activity-travel behavior. Transportation Letters 2(4), 217–230.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis and Prevention 42(6), 1556–1565.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography 54, 248–256.

Jiang, X., Huang, B., Zaretzki, R.L., Richards, S., Yan, X., Zhang, H., 2013. Investigating the influence of curbs on single-vehicle crash injury severity utilizing zero-inflated ordered probit models. Accident Analysis and Prevention 57, 55–66.

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

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

Marcoux, R., Yasmin, S., Eluru, N., Rahman, M., 2018. Evaluating temporal variability of exogenous variable impacts over 25 years: An application of scaled generalized ordered logit model for driver injury severity. Analytic Methods in Accident Research 20, 15–29.

Mohammadnazar, A., Mahdinia, I., Ahmad, N., Khattak, A.J., Liu, J., 2021. Understanding how relationships between crash frequency and correlates vary for multilane rural highways: Estimating geographically and temporally weighted regression models. Accident Analysis and Prevention 157, 106146.

Paleti, R., Eluru, N., Bhat, C.R., 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. Accident Analysis and Prevention 42(6), 1839–1854.

Pervaz, S., Bhowmik, T., Eluru, N., 2022. Integrating macro and micro level crash frequency models considering spatial heterogeneity and random effects. Analytic Methods in Accident Research 36, 100238.

Savolainen, P.T., Mannering, F.L., Lord, D., 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.

Soroori, E., Mohammadzadeh Moghaddam, A., Salehi, M., 2019. Application of local conditional autoregressive models for development of zonal crash prediction models and identification of crash risk boundaries. Transportmetrica A 15(2), 1102–1123.

Wang, J., Huang, H., 2016. Road network safety evaluation using Bayesian hierarchical joint model. Accident Analysis and Prevention 90, 152–158.

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

Wang, X., Kim, S.H., 2019. Prediction and Factor Identification for Crash Severity: Comparison of Discrete Choice and Tree-Based Models. Transportation Research Record 2673(9), 640–653.

Wang, X., Zhou, Q., Yang, J., You, S., Song, Y., Xue, M., 2019. Macro-level traffic safety analysis in Shanghai, China. Accident Analysis and Prevention 125, 249–256.

Weiss, H.B., Kaplan, S., Prato, C.G., 2014. Analysis of factors associated with injury severity in crashes involving young New Zealand drivers. Accident Analysis and Prevention 65, 142–155.

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

Xiong, Y., Mannering, F.L., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach. Transportation Research Part B 49, 39–54.

Yasmin, S., Eluru, N., 2018. A joint econometric framework for modeling crash counts by severity. Transportmetrica A 14(3), 230–255.

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis and Prevention 59, 506–521.

Yasmin, S., Eluru, N., Bhat, C.R., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Analytic Methods in Accident Research 1, 23–38.

Zeng, Q., Gu, W., Zhang, X., Wen, H., Lee, J., Hao, W., 2019. Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors. Accident Analysis and Prevention 127, 87–95.

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

APPENDIX

Table A1: Estimation Results of the Non-integrated (Ordered Probit and NB-OPFS) Models (N = 280 Traffic Analysis Zones with 18,286 Crash Records)

Parameters	Estimates	t-stat
Disaggregate level ordered probit model		
Threshold between NI-PI	0.809	31.702
Threshold between PI-NII	1.576	58.048
Threshold between NII-FII	2.479	72.259
Crash type (Base: Rear-end and other crash types)		
Angle	0.303	7.735
Head on	0.478	5.385
Left turn	0.305	9.346
Right turn	-0.243	-2.775
Sideswipe	-0.553	-16.725
Non-motorized	1.705	31.014
Single motorized vehicle	0.253	5.557
Time (Base: Peak morning, off-peak morning, off-peak evening and late evening)		
Late night (00:00 to 6:30)	0.103	2.713
Peak evening (16:00 to 18:30)	-0.053	-2.208
Season (Base: Summer and Autumn)		
Spring (March to May)	-0.049	-2.097
Winter (December to February)	-0.040	-1.715
Speed limit (Base: Speed limit < 40 mph)		
Speed limit 40-54 mph	0.091	4.485
Speed limit >=55 mph	0.116	3.398
Location (Base: Not at intersection)		
Intersection	0.095	3.972
Shoulder type (Base: Other shoulder types)		
Curb shoulder type	-0.036	-1.815
Presence of passengers (Base: Driver only)		
With passenger	0.338	17.199
Driving under influence (Base: Not DUI related)		
DUI related	0.470	6.179
Distraction (Base: Not distraction related)		
Distraction related	0.234	10.019
Lighting condition (Base: Daylight and dawn/dusk)		
Dark lighted	0.066	2.564
Dark not lighted	0.169	2.542
Log-likelihood: -14,637.577; BIC: 29,500.874; Number of parameters: 23		
Aggregate level NB-OPFS model		
Count component		
Constant	1.660	4.170
Average inside shoulder width	0.027	2.253
Intersection density	0.548	1.742
Proportion of >=3 lane roads	0.249	1.731
Proportion of divided roads	0.398	2.919
AADT	0.104	2.769
Proportion of residential areas	0.240	1.907
Proportion of commercial areas	0.684	3.755
Number of educational centers	0.124	3.854

Parameters	Estimates	t-stat
Non-motorized means of transport	0.171	5.507
Proportion of African American population	0.642	4.057
Overdispersion parameter	0.997	7.380
Proportion component		
Threshold between NI-PI	0.460	6.671
Threshold between PI-NII	1.158	18.760
Threshold between NII-FII	1.965	30.275
Average inside shoulder width	-0.016	-3.579
Proportion of <40 mph roads	-0.201	-4.275
Traffic signal density	-0.210	-1.727
Proportion of heavy vehicles	1.658	1.684
Proportion of commercial areas	-0.131	-1.907
Number of restaurants	-0.036	-2.243
Proportion of African American population	0.137	2.258
Log-likelihood: -1,653.646; BIC: 3,431.258; Number of parameters: 22		
Non-integrated model (combined) - BIC	32,836.046	
Non-integrated model (combined) - Log-likelihood	-16,291.240	
Non-integrated model (combined) - Number of parameters	45	

Note: NI = no-injury crashes, PI = possible injury crashes, NII = non-incapacitating injury crashes, and FII = fatal and incapacitating injury crashes.