A Novel Integrated Approach to Modeling and Predicting Crash Frequency by Crash Event State

Angela Haddad
The University of Texas at Austin
Dept of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712-1172
Email: angela.haddad@utexas.edu

Aupal Mondal
The University of Texas at Austin
Dept of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712-1172
Email: aupal.mondal@utexas.edu

Naveen Eluru
University of Central Florida
Dept of Civil, Environmental and Construction Engineering
12800 Pegasus Drive, Room 301D, Orlando, Florida 32816, USA
Email: naveen.eluru@ucf.edu

Chandra R. Bhat*
The University of Texas at Austin
Dept of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712-1172
Phone: 512-471-4535, Fax: 512-475-8744
Email: bhat@mail.utexas.edu

*corresponding author

ABSTRACT

In this study, we propose a novel integrated parametric framework for analyzing multivariate crash count data based on linking a univariate count model for the total count of motor vehicle crashes across all possible crash states with a discrete choice model for crash event state given a crash. In doing so, we are able to use information at the disaggregate crash-level from an unordered model structure in analyzing the aggregate level crash count. To our knowledge, this is the first such model proposed in the econometric literature. We apply this approach in a demonstration exercise to examine the number of motor vehicle crashes in Census Block Groups (CBGs) in Austin, Texas, considering four injury severity levels. At the disaggregate level, we incorporate several explanatory variables such as the characteristics of the most severely injured individual and at-fault vehicle's parties, crash time variables (time of day, weather), crash location variables, and CBG level variables. At the aggregate level, we consider CBG level variables, including road design factors, land-use variables, crash exposure factors, aggregate sociodemographic attributes, and crime and traffic violations related measures. Importantly, our results indicate a significant and positive linkage between the disaggregate crash event state dimensions and the total crash count. Through the use of elasticity measures, our results also clearly highlight the improved policy sensitivity of the integrated model framework.

Keywords: Multivariate Crash Count, Injury Severity, Integrated Framework, Crash Analysis

1. INTRODUCTION

Traffic crashes represent an enormous cost to society in terms of property damage, productivity loss, emotional trauma, injury, and even death. According to the National Highway Traffic Safety Administration (National Center for Statistics and Analysis, 2023), an estimated 42,939 individuals died in roadway crashes in the U.S. in 2021, an increase of 11.1% over 2020 and the highest year-over-year increase since 2007. An additional 2.5 million individuals are estimated to have incurred serious injuries in 2021, underscoring the continued (and even rising) dangers associated with being a road user in the U.S. Thus, it continues to be critical to analyze the occurrence and the severity of crashes, so that appropriate geometric countermeasures and behavioral interventions may be designed to reduce both the number and the severity of injuries sustained if a crash occurs.

In safety research, crash frequency analysis is typically undertaken by aggregating crashes over a certain spatial and temporal scale (such as the number of crashes per year at intersections, specific roadway segments, census tracts, or traffic analysis zones). Various attributes of the spatial unit of analysis are used as determinants of crash frequency, including land-use and built environment (BE) factors, vehicle volumes and vehicle-mix factors, roadway geometry, and traffic control type at the location. While many earlier safety studies focused on total crashes at the spatial location (Lord and Mannering, 2010, Savolainen et al., 2011, and Shin and Washington, 2012 provide a good review), recent studies have increasingly recognized that the determinants of crash counts likely vary by crash type and severity, and that aggregating all crashes into a single "total crash" dependent variable invites the pitfalls of the classic ecological fallacy; that is, if the dependent variable represents total crashes, the implicit implication is that the effect of a particular variable is the same regardless of crash type. As a result, many studies have focused on the count of crashes of a specific type, such as based on road user type (for example, pedestrian crashes or bicyclist crashes), injury type (for example, fatal or serious injuries), impact type (head-on or rear-end or angle), or vehicle number and type (single vehicle crashes or multiple vehicle crashes, and heavy trucks versus passenger vehicles (see, for example, Abdel-Aty et al., 2011, Narayanamoorthy et al., 2013, and Hosseinpour et al., 2014 for such studies). But most such studies still consider univariate count models, and, if they differentiate among different crash types, do so by estimating independent univariate count models.

The safety analysis field has long matured in the area of univariate count models, with many different approaches ranging the smorgasbord from the simple Poisson and negative binomial models to discrete distribution models with binomial/logarithmic distributions to zero-inflated count models (see, for example, Musio et al., 2010) and hurdle-count models (see, for example, Bethell et al., 2010). More recently, though, there has been a literal explosion in the use of multivariate count models, which can arise in one or more of several ways in crash analysis: (1) Repeated univariate count data from the same spatial unit, so that some unobserved factors relevant to the spatial unit may impact the univariate count observations across different observation periods (causing the entire set of count observations at the spatial unit to be correlated, thus resulting in a multivariate crash event), (2) Spatially-correlated univariate count data, so that observed and unobserved factors impacting the univariate count at one spatial unit also impact the univariate count at proximal locations (causing the entire set of count observations across different spatial units to become correlated, thus again resulting in a multivariate count set), (3) Explicit multivariate count data of different crash types at a spatial unit at a specific cross-section of time, because of common unobserved factors affecting multiple crash types simultaneously (causing multiple count dependent variables based on, for example, different road user types, injury severity levels, impact types, or vehicle types to become stochastically dependent). In this paper, the focus will be on the last type corresponding to explicit multivariate count data models. Readers may refer to Castro et al., 2012, Narayanamoorthy et al., 2013, Cai et al., 2016, and Ziakopoulos and Yannis, 2020, for examples of the implicit multivariate count data arising from recognizing temporal and/or spatial correlation in univariate count data, though some of these methods have also been extended to multiple dependent variables; Cui and Xie (2021) provide a good overview of such implicit multivariate count data crash models.[1]

1.1. Explicit Multivariate Count Data Models

In the context of explicit multivariate count data, one may consider a simple Poisson or negative binomial discrete distribution, and develop multivariate versions of these discrete distributions to accommodate correlated counts (see Buck et al., 2009, and Bermúdez and Karlis, 2011 for

[1] In the literature, there does not appear to be adequate recognition of the fact that spatial/temporal stochasticity dependencies of univariate counts also lead to multivariate count models; the use of the term "multivariate counts" is usually reserved, based on our taxonomy in this study, for explicit multivariate count data models.

applications of these methods). While having the advantage of a closed form, these become cumbersome as the number of correlated counts increases and they also represent the undesirable property that they can only accommodate a positive correlation in the counts.

Alternatively, one may use a discrete or continuous mixing structure, in which one or more random terms are introduced in the parameterization of the mean for the count in each crash event state (so that the mean is not only a function of exogenous variables, but also includes one or more random terms within the exponentiated mean function of the Poisson distribution; see, for example, Barua et al., 2014, Yasmin and Eluru, 2018, Bhowmik et al., 2021, and Pervaz et al., 2023 for extensive reviews). The most common form of such a mixture is to include normally distributed terms. If a multivariate distribution is assumed for these normal error terms across the different count event states, this leads to a multivariate count model. The advantage of this method is that it permits both positive and negative dependency between the counts, but the limitation is that the approach gets quickly cumbersome in the presence of several crash event states. Another related problem with these multivariate count models is that there are likely to be excess zeros in each crash event category. This necessitates the use of zero-inflated and hurdle-count techniques. Unfortunately, such techniques, while simple to implement in a univariate count setting, become extremely difficult, if not infeasible, in a multivariate setting (see Mannering et al., 2016 for a detailed discussion).[2] Moreover, these multivariate count models are not able to capture crash-specific variables. For example, when directly modeling the number of crashes by injury severity at an intersection using a multivariate crash model, the analyst is unable to consider the intoxication/sober state of a driver traveling through the intersection at a particular time. But we might expect that an inebriated driver at an intersection would be more prone to severe injury risk if in a crash, while also increasing crash risk at the intersection. Similarly, consider a driver not wearing their seat belt and traveling through an intersection at a particular time. This may lead to a higher crash risk at the intersection (because, in general, unbuckled drivers tend to be more aggressive drivers; see Eluru and Bhat, 2007) as well as result in a higher injury severity risk conditional on a crash. Again, the "buckled or not buckled" state of a driver at a specific crash

[2] An alternative approach to analyze crash rates (for example, number of crashes per 100 million vehicle miles of travel) by injury severity level in the presence of excess zeros is to translate the dependent variable vector from a multivariate count to a multivariate continuous variable. For example, to address the preponderance of zero values, Anastasopoulos and Mannering (2011) developed a multivariate Tobit-regression model to analyze crash rates by injury severity level. However, the likelihood estimation approach again becomes cumbersome and presents a computational challenge when there are many tobit regressions in the multivariate set-up.

(time) instance at the intersection cannot be considered in a multivariate crash model by injury severity.[3] Fundamentally, multivariate count models are not able to adequately accommodate the effects of variables on crash counts through their effects on crash event state.

Another approach uses a strictly hierarchical combination of a count model to analyze total crashes and a discrete choice model or another count model that allocates the total count to different crash event states given a crash (see, for example, Kim et al., 2007, Milton et al., 2008, Yamamoto et al., 2008, Huang and Abdel-Aty, 2010, and Fu et al., 2023). Also, the many studies in the literature that focus solely on total crashes or solely on injury severity/crash type conditioned on a crash implicitly assume such a strictly hierarchical mechanism for predicting crashes by injury severity level/crash type. In this hierarchical setting, the probability of the observed counts in each injury severity level/crash type, given the total count, takes a multinomial distribution form (see Terza and Wilson, 1990). This structure, while easy to estimate and implement, may not be very appropriate for crash analysis. Thus, for example, consider the presence of wide inside shoulders or even physical barriers (between opposing directions of movement) at a rural highway segment site. Such wide shoulders/barriers are likely to reduce the number of head-on crashes and fatal injuries if there is a crash at the site (see, for example, Castro et al., 2012). But, in the pool of total crashes, the number of fatal injuries relative to the non-fatal injuries is small, and thus it is possible that, in a total count model, the effect of wide inside shoulders or a physical barrier does not turn out to be statistically significant.[4] In such a case, because the total count does not turn out to be affected by wide inside shoulders/barriers (of course, incorrectly so), the net result in a strictly hierarchical model of crash count and injury severity would be a decrease in the count of fatal injuries, but a necessary increase in the count of non-fatal injuries. The latter result would not make much sense at all.

An alternate and more appealing structure is one that explicitly links the event state discrete choice model with the total crash count model. A recent effort by Pervaz et al. (2023) attempts to do so by first estimating an injury severity ordered-response model at the event state level (that is,

[3] It may be argued that such effects could be considered in a multivariate count model by continually disaggregating outcomes, such as by analyzing crashes by injury severity by drunken state by seat belt use and so on. But the number of determinants of injury severity conditional on a crash can be quite a few, leading to a multivariate count model with too many crash event states. In addition to estimation problems as discussed earlier, this also has the effect of "thinning" the sample of non-zero crash counts within each of the very disaggregated crash event state.

[4] Such occurrences will be especially commonplace as the number of disaggregate event states (crash severity level and crash types) increases, since the number of crashes in each event state will be but a small fraction of total crashes.

a crash level), and then including the sum of the estimated underlying injury propensities across crashes within a joint model system of total count and a fractional split model to disaggregate the total counts into different injury severity types. But, as discussed later, their approach is applicable only for event type models that may be viewed as an ordered-response outcome; they also use a fractional split model for crash type rather than directly using a discrete outcome model at the crash level to partition the total count into its component crash types.

2. THE CURRENT PAPER

In the current paper, we consider a flexible unordered-response process as underlying the event state. The highest event state risk propensity from this event state model is then used as an explanatory variable in the total crash count model. In doing so, the factors in the unobserved portions of event state crash propensities must also influence the total crash count intensity just as the observed factors in the event state crash propensities do. This is essential to recognize the full econometric jointness between the event state (given a crash) and the total crash count, as does the model proposed in this paper. In doing so, we use a multinomial probit (MNP) model for the crash event state discrete model (conditional on a crash), rather than the traditional multinomial logit (MNL) or nested logit (NL) kernel used in earlier studies. The use of the MNP kernel allows a more flexible covariance structure for the event states relative to traditional GEV (Generalized Extreme Value) kernels. In addition, the model system allows random variations (or unobserved heterogeneity) in the sensitivity to exogenous factors in both the crash event state model as well as the total crash count components. The formulation also allows handling excess zeros in a straightforward manner (or excess counts of any value), which is a common characteristic of crash counts (see Lord, 2006). In contrast, a more recent group of multivariate crash count studies (see, for example, Yasmin and Eluru, 2018, Afghari et al., 2020, Bhowmik et al., 2021, Wang et al., 2021, Pervaz et al., 2023) that use a fractional split approach (see Papke and Wooldridge, 1996, and Sivakumar and Bhat, 2002) for crash event state combined with a count model do not account for such excess crash count values. Besides, they also use a statistical stitching mechanism similar to the mixing approaches discussed earlier to generate stochastic jointness, and use a restrictive ordered-response mechanism to model event state (such as injury severity). The ordered-response mechanism, while potentially parsimonious, can lead to severe inconsistencies in model estimates if the rigid implied restrictions of the effects of exogenous variables on the multivariate counts do

not hold (Bhat and Pulugurta, 1998). Also, even if there may be some basis for considering certain event states (such as injury severity) as being ordered, many event states must be modeled as unordered responses anyway (for example, if user type or crash type are event states). In this paper, and for the first time that we are aware of, we develop a multivariate count data framework that is able to use information at the disaggregate crash-level from an unordered model structure, given a crash. Further, by explicit modeling of the event state outcome at the event state level, rather than aggregating these outcomes into fractions, we consider unobserved heterogeneity at the event state outcome level. Additionally, our linkage from the event state to the total count naturally arises as the maximum risk across all event states, lending conceptual and theoretical support to the methodology. The linkage captures the intuitive notion that variables that positively impact the highest risk injury severity level given a crash at a particular spatial unit at a particular instant will also positively impact the total crash risk at that spatial unit.

Overall, we propose an integrated parametric framework for multivariate crash count data that is based on linking a univariate count model for the total count of crashes across all possible crash type states (that we will henceforth refer to as crash event states) with a discrete choice model for crash event state given a crash. The approach is applied in a demonstration exercise to examine the number of motor vehicle crashes in Census Block Groups (CBGs) in Austin, Texas by four injury severity levels. The data for the analysis is drawn from the Texas Department of Transportation crash incident files. Explanatory variables considered at the disaggregate level include the characteristics of the most severely injured individual and at-fault vehicle's parties, crash time variables (time of day, weather), crash location variables, and CBG level variables. At the aggregate level, we consider CBG-level variables, including road design factors, land-use variables, crash exposure factors, aggregate sociodemographic attributes, and crime and traffic violation-related measures.

The rest of this paper is structured as follows. The next section presents the model structure and estimation procedure. Section 4 describes the study area for our analysis of crashes, the data source, and sample characteristics. Section 5 presents the empirical estimation results and their implications for safety analysis. Finally, Section 6 concludes the paper.

3. MODELING FRAMEWORK

3.1. Model Formulation

Let q ($q = 1, 2, ..., Q$) be an index to represent CBGs and let i ($i = 1, 2, ..., I$) be an index to represent crash event states (for example, combinations of crash types and injury severity levels). In the empirical demonstration exercise in this paper, there are four event states. The precise definitions of the event states, used in this study, are provided later in Section 4. Let k $(k = 0, 1, 2, ..., \infty)$ be the index to represent total crash frequency and let n_q be the total number of crashes at CBG q over a certain period of interest (n_q takes a specific value in the domain of k). Each count unit contribution to the total count n_q of crashes at CBG q corresponds to a crash instance in which one of the I event states is manifested. Let t_q be an index for crash instance, so that t_q takes the values from 1 to n_q for CBG q. As a result, the crash event discrete model takes the form of a panel discrete choice model, with n_q crash observations ($t_q = 1, 2, \ldots, n_q$) from CBG q. The resulting data allows the estimation of CBG-specific unobserved factors that influence the intrinsic propensity risk of each crash event state as well as the effects of other exogenous variables.

In the rest of this paper, we will use the following notations: $MVN_R(\boldsymbol{b}, \boldsymbol{\Sigma})$ for the multivariate normal distribution of R dimensions with mean vector $\boldsymbol{b}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathbf{IDEN}_R$ for an identity matrix of dimension R, $\mathbf{1}_R$ for a column vector of ones of dimension R, $\mathbf{0}_R$ for a column vector of zeros of dimension R, and $\mathbf{1}_{RR}$ for a matrix of ones of dimension $R \times R$.

3.1.1. Crash event state model

Let the risk propensity of observing crash event state i at crash instance t_q at CBG q be S_{qt_qi} and write this propensity as a function of an exogenous variable vector $\mathbf{x}_{qt_qi}$ that may include both crash-level variables (such as time-of-day, day-of-week, season-of-year, crash type, number and type of vehicles involved, crash location attributes, and weather conditions) contained in an $(D_1 \times 1) -$ vector $\mathbf{z}_{qt_qi}$ as well as broader CBG-level exogenous variables (such as aggregated CBG road network and built-environment characteristics, and vehicle ownership trends) contained in another $(D_2 \times 1) -$ vector $\boldsymbol{\varpi}_{qi}$. Thus, $\mathbf{x}_{qt_qi} = \left(\mathbf{z}'_{qt_qi}, \boldsymbol{\varpi}'_{qi}\right)'$ is a $(D \times 1) -$ vector, where $D = D_1 + D_2$. $\mathbf{z}_{qt_qi}$

includes a constant for all event states except one, and may include interactions of crash-level and CBG-level attributes as they affect the risk propensities of the crash event states (for later use in forecasting, we will introduce variables in $\mathbf{z}_{qt_q i}$ as categorical variables). We then write the following for $S_{qt_q i}$:

$$S^*_{qt_q i} = \boldsymbol{\beta}'_q \mathbf{x}_{qt_q i} + \tilde{\boldsymbol{\varepsilon}}_{qt_q i};\ \boldsymbol{\beta}_q = \mathbf{b} + \tilde{\boldsymbol{\beta}}_q\ ,\ \tilde{\boldsymbol{\beta}}_q \sim MVN_D(\mathbf{0}_D, \boldsymbol{\Omega}), \tag{1}$$

where $\boldsymbol{\beta}_q$ is a crash-specific (D×1)-column vector of corresponding coefficients. $\boldsymbol{\beta}_q = \left(\boldsymbol{\beta}'_{qz}, \boldsymbol{\beta}'_{q\varpi}\right)'$, where $\boldsymbol{\beta}_{qz}$ is the coefficient vector on $\mathbf{z}'_{qt_q i}$ and $\boldsymbol{\beta}_{q\varpi}$ is the coefficient vector on $\boldsymbol{\varpi}_{qi}$. $\boldsymbol{\beta}_q$ is assumed to be a realization from a multivariate normal density function with mean vector $\mathbf{b}$ and covariance matrix $\boldsymbol{\Omega}$ (this specification allows heterogeneity in the effects of exogenous variables due to unobserved CBG and crash-level attributes). $\tilde{\boldsymbol{\varepsilon}}_{qt_q i}$ is assumed to be an independently and identically distributed (across crash instances and across CBGs) error term, but having a general covariance structure across crash categories at each crash instance. Thus, consider the $(I \times 1)$-vector $\tilde{\boldsymbol{\varepsilon}}_{qt_q} = (\tilde{\varepsilon}_{qt_q 1}, \tilde{\varepsilon}_{qt_q 2}, \tilde{\varepsilon}_{qt_q 3}, \ldots, \tilde{\varepsilon}_{qt_q I})'$ and assume that $\tilde{\boldsymbol{\varepsilon}}_{qt_q} \sim MVN_I(\mathbf{0}_I, \tilde{\boldsymbol{\Theta}})$. Define $\mathbf{S}_{qt_q} = (S_{qt_q 1}, S_{qt_q 2}, ..., S_{qt_q I})'$ (I×1 vector), and $\mathbf{x}_{qt_q} = (\mathrm{x}_{qt_q 1}, \mathrm{x}_{qt_q 2}, ..., \mathrm{x}_{qt_q I})'$ (I×D matrix). Then, we can write:

$$\mathbf{S}_{qt_q} = \mathbf{x}_{qt_q} \boldsymbol{\beta}_q + \tilde{\boldsymbol{\varepsilon}}_{qt_q} = \tilde{\mathbf{V}}_{qt_q} + \tilde{\boldsymbol{\varepsilon}}_{qt_q}\ . \tag{2}$$

Next, let the crash event type observed at the t^{th} crash instance at CBG q be c_{qt_q} ($c_{qt_q} \in 1,2,...,I$). Define $\mathbf{M}_{qt_q}$ as a $[(I-1)\times I]$ matrix, corresponding to an $(I-1)$ identity matrix with an extra column of -1 values added as the $c^{\text{th}}_{qt_q}$ column. In the propensity differential form (where the propensity differentials are taken with respect to the observed crash event state c_{qt} at each crash instance), we may write the risk propensities of Equation (2) in differenced form (differenced from the risk propensity of the actually observed crash event observed at the $t_q{}^{th}$ crash instance at CBG q as:

$$\begin{aligned} \mathbf{s}^*_{qt_q} &= \mathbf{M}_{qt_q} \mathbf{S}^*_{qt_q} = \mathbf{M}_{qt_q} \left(\mathbf{x}_{qt_q} \boldsymbol{\beta}_q + \tilde{\boldsymbol{\varepsilon}}_{qt_q}\right) = \mathbf{M}_{qt_q} \mathbf{x}_{qt_q} \boldsymbol{\beta}_q + \mathbf{M}_{qt_q} \tilde{\boldsymbol{\varepsilon}}_{qt_q} \\ &= \mathbf{V}_{qt_q} + \boldsymbol{\varepsilon}_{qt_q},\ \mathbf{V}_{qt_q} = \mathbf{M}_{qt_q} \mathbf{x}_{qt_q} \boldsymbol{\beta}_q = \mathbf{M}_{qt_q} \tilde{\mathbf{V}}_{qt_q},\ \boldsymbol{\varepsilon}_{qt_q} = \mathbf{M}_{qt_q} \tilde{\boldsymbol{\varepsilon}}_{qt_q}. \end{aligned} \tag{3}$$

Next, we obtain the following result:

$$\mathbf{s}^*_{qt_q} \mid \boldsymbol{\beta}_q \sim MVN_{(I-1)}(\mathbf{V}_{qt_q}, \boldsymbol{\Theta}_{qt_q}), \text{ where } \boldsymbol{\Theta}_{qt_q} = \mathbf{M}_{qt_q}\tilde{\boldsymbol{\Theta}}\mathbf{M}'_{qt_q}. \quad (4)$$

The likelihood of observing the sequence of observations across the n_q crash instances at CBG *q*, conditional on the coefficient vector $\boldsymbol{\beta}_q$, may be written as:

$$L_{q,\text{crash event state}}(\tilde{\boldsymbol{\Theta}})\big|\boldsymbol{\beta}_q = \prod_{t_q=1}^{n_q} \Phi_{(I-1)}\left(-\mathbf{V}^*_{qt_q}; \boldsymbol{\Theta}^*_{qt_q}\right), \text{ where } \mathbf{V}^*_{qt_q} = \boldsymbol{\omega}^{-1}_{\boldsymbol{\Theta}_{qt_q}}\mathbf{V}_{qt_q} \text{ and } \boldsymbol{\Theta}^*_{qt_q} = \boldsymbol{\omega}^{-1}_{\boldsymbol{\Theta}_{qt_q}}\boldsymbol{\Theta}_{qt_q}\boldsymbol{\omega}^{-1}_{\boldsymbol{\Theta}_{qt_q}}, \quad (5)$$

where $\Phi_{(I-1)}\left(-\mathbf{V}^*_{qt_q}; \boldsymbol{\Theta}^*_{qt_q}\right)$ represents the standard multivariate normal cumulative distribution (MVNCD) function of dimension *I*–1, with the upper truncation points given by the vector $-\mathbf{V}^*_{qt_q}$ and the correlation matrix given by $\boldsymbol{\Theta}^*_{qt_q}$. $\boldsymbol{\omega}_{\boldsymbol{\Theta}_{qt_q}}$ is a diagonal matrix of standard deviations of $\boldsymbol{\Theta}_{qt_q}$ (the equation above results because the event instance outcomes are stochastically independent conditional on $\boldsymbol{\beta}_q$). The parameters to be estimated include the **b** vector, and the elements of the covariance matrices $\boldsymbol{\Omega}$ and $\tilde{\boldsymbol{\Theta}}$.[5] However, note that the parameters from the event state model also appear in the total count model, and hence we discuss the overall estimation procedure for the total count-event type model in Section 3.2 after first discussing the linking function and total count model formulation in the next couple of sections.[6]

3.1.2. Linking function

At each crash instance, a measure of the overall crash propensity may be obtained as the maximum of the value across the crash event state (type/injury severity level) risk propensities (because the highest risk state is what will get manifested as the crash event state at each instance). Next, to

[5] Due to identification considerations (see Bhat et al., 2013), not all parameters of $\tilde{\boldsymbol{\Theta}}$ are estimable. To be precise, only the covariance matrix of the error differences (with respect to an alternative) are estimable, and even that after scaling the variance of one of the error term differences to zero. Please see Section 5.2.5.

[6] Of course, if the analyst uses a simpler multinomial logit (MNL) model for the crash event state, then the crash event state likelihood expression, conditional on $\boldsymbol{\beta}_q$, collapses to the following:

$$L_{q,\text{crash event state}} \mid \boldsymbol{\beta}_q = \prod_{t_q=1}^{n_q}\prod_{i=1}^{I}\left(\frac{\exp\left(\boldsymbol{\beta}'_q\mathbf{x}_{qt_qi}\right)}{\sum_{j=1}^{I}\exp\left(\boldsymbol{\beta}'_q\mathbf{x}_{qt_qj}\right)}\right)^{1(i=c_{qt_q})}$$

, where $1(i = c_{qt_q})$ is a dummy variable taking the value 1 if alternative *i* is equal to c_{qt_q} (that is, if alternative i is the event state alternative at crash occasion t_q), and 0 otherwise.

recognize the fact that different crash instances may have different values for the exogenous variables (including observed and unobserved individual-specific and crash-specific factors), one can compute the overall crash propensity at a specific location as the geometric mean over the maximum values of crash risk across all crash instances at a location, conditional on overall unobserved location-specific effects captured in $\boldsymbol{\beta}_q$. This variable can then be included as an explanatory variable in the crash frequency model along with other variables that impact the overall number of crashes without impacting the severity of injury if a crash were to happen. To develop this link, consider the expression for the crash risk propensity of crash event state *i* at crash instance t_q $(t_q = 1, 2, ..., \eta_q)$ at CBG *q* in Equation (1). Based on this expression at the crash event state level, we may write the aggregate risk of a crash occurrence due to the crash event state, given observed crash-level/CBG-level characteristics embedded in the vector $\mathbf{x}_{qt_q}$ as well as the impact of these exogenous variables as captured in the CBG-specific effects vector $\boldsymbol{\beta}_q$, as follows:

$$\eta_{qt_q} \mid \boldsymbol{\beta}_q = Max(S^*_{qt_q1}, S^*_{qt_q2}, ..., S^*_{qt_qI} \mid \boldsymbol{\beta}_q) = \underset{i}{Max}(S^*_{qt_qi} \mid \boldsymbol{\beta}_q),\ i = 1, 2, ..., I. \tag{6}$$

$\eta_{qt_q} \mid \boldsymbol{\beta}_q$ is, of course, distributed with a cumulative distribution function given by:

$$\text{Prob}(\eta_{qt_q} \mid \boldsymbol{\beta}_q < z) = \text{Prob}\left[\left(S^*_{qt1} \mid \boldsymbol{\beta}_q\right) < z, \left(S^*_{qt2} \mid \boldsymbol{\beta}_q\right) < z, ..., \left(S^*_{qtI} \mid \boldsymbol{\beta}_q\right) < z\right] = F_I\left[z\mathbf{1}_I;\ \tilde{\mathbf{V}}_{qt_q},\ \tilde{\boldsymbol{\Theta}}\right], \quad \text{where}$$

F_I is the multivariate normal cumulative distribution function of dimension *I* with mean $\tilde{\mathbf{V}}_{qt_q}$ and covariance matrix $\tilde{\boldsymbol{\Theta}}$. Next, define $\boldsymbol{\delta}_q \mid \boldsymbol{\beta}_q = Max\left(\eta_{q1}, \eta_{q2}, ..., \eta_{qn_q}\right) \Big| \boldsymbol{\beta}_q$, which is the aggregate crash risk at CBG *q* in a given time period (that is, over all the crash occurrences observed over the time period). This random variable has a cumulative distribution function as follows:

$$\text{Prob}\left[\left(\boldsymbol{\delta}_q \mid \boldsymbol{\beta}_q\right) < z\right] = \text{Prob}\left(\eta_{q1} < z, \eta_{q2} < z, ..., \eta_{qn_q} < z\right) \Big| \boldsymbol{\beta}_q = \left(\prod_{t_q=1}^{n_q} F_I\left[z\mathbf{1}_I;\ \tilde{\mathbf{V}}_{qt_q},\ \tilde{\boldsymbol{\Theta}}\right]\right) \Big| \boldsymbol{\beta}_q \tag{7}$$

Next, to control for the different number of crash occasions at each CBG *q*, we develop another constant risk variable that operates at the individual crash level at CBG *q* in such a way that the probability of the per-crash individual risk, when aggregated across all crash occasions, provides the same probability as the aggregate crash risk across all observed crash occasions at that CBG. Statistically speaking, the objective is to define a random variable at the individual crash occurrence level at CBG *q* with a cumulative distribution function such that, across all crash

occurrences, the aggregate crash risk has the same cumulative distribution function (CDF) as that of $\boldsymbol{\delta}_q \,|\, \boldsymbol{\beta}_q$. Define such a per-crash risk at CBG q, given $\boldsymbol{\beta}_q$, as $\tau_q \,|\, \boldsymbol{\beta}_q$. Let the CDF of $\tau_q \,|\, \boldsymbol{\beta}_q$ be $\text{Prob}(\tau_q < z \,|\, \boldsymbol{\beta}_q) = G(z) \,|\, \boldsymbol{\beta}_q$. Then, across all crash occurrences at CBG q, the distribution of the aggregate crash risk would be $\left[G(z)\right]^{n_q} \Big| \boldsymbol{\beta}_q$. Equating this to the expression in Equation (7), we get the following as the distribution function of $\tau_q \,|\, \boldsymbol{\beta}_q$:

$$\left[G(z)\right]^{n_q} \Big| \boldsymbol{\beta}_q = \left(\prod_{t_q=1}^{n_q} F_I\left[z\mathbf{1}_I ;\, \tilde{\mathbf{V}}_{qt_q}, \tilde{\boldsymbol{\Theta}} \right] \right) \Big| \boldsymbol{\beta}_q \text{, or } \quad \left[G(z)\right] \Big| \boldsymbol{\beta}_q = \left(\prod_{t_q=1}^{n_q} F_I\left[z\mathbf{1}_I ;\, \tilde{\mathbf{V}}_{qt_q}, \tilde{\boldsymbol{\Theta}} \right] \right)^{\frac{1}{n_q}} \Big| \boldsymbol{\beta}_q . \tag{8}$$

The reader will note that $\tau_q \,|\, \boldsymbol{\beta}_q$ is itself a stochastic variable, and it is important to consider this stochasticity in the effect on crash occurrence. It is this $\tau_q \,|\, \boldsymbol{\beta}_q$ that serves as a CBG-specific linking term between the crash events and the CBG counts. To our knowledge, this linking function for the per-instance aggregate crash risk across multiple crash instances is a first in the econometric literature, and is the fundamental vehicle that allows us to consider conditional-on-crash attributes in a crash occurrence model.

3.1.3. Crash frequency model

The crash frequency model is based on a Generalized Ordered Response Probit (GORP) representation for count models (see Bhat, 2015, and Castro et al., 2012 who show that any count model may be reformulated as a special case of a GORP model in which a single latent continuous variable is partitioned into mutually exclusive intervals). This representation generalizes traditional count models, can exactly reproduce any traditional count data model, and allows handling excess zeros with ease.

Define the latent crash propensity for CBG q as y_q^* and consider the following structure:

$$y_q^* \,|\, \boldsymbol{\beta}_q = \left(\boldsymbol{\theta} + \tilde{\boldsymbol{\theta}}_q\right)' \mathbf{w}_q + \vartheta\left(\tau_q \,|\, \boldsymbol{\beta}_q\right) + \zeta_q, \quad y_q = k \quad \text{if} \quad \psi_{q,k-1} < y_q^* \,|\, \boldsymbol{\beta}_q < \psi_{qk}, \tag{9}$$

$$\psi_{qk} = f_k(\tilde{\mathbf{z}}_q) + \sum_{l=0}^{k} \alpha_l$$

The parameter ϑ is the linkage parameter. $\mathbf{w}_q$ is an (L×1)-column vector of exogenous attributes (excluding a constant), $\boldsymbol{\theta}_q$ is a corresponding (L×1)-column vector of CBG-specific variable

effects, and ζ_q is a random error term assumed to be identically and independently standard normal distributed across CBGs. $\boldsymbol{\theta}_q$ is a realization from a multivariate normal density function with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Xi}$, such that $\boldsymbol{\theta}_q = \boldsymbol{\theta} + \tilde{\boldsymbol{\theta}}_q$ and $\tilde{\boldsymbol{\theta}}_q \sim MVN_L(\mathbf{0}_L, \boldsymbol{\Xi})$ is independent of ζ_q ($\tilde{\boldsymbol{\theta}}_q$ is a CBG-specific coefficient vector introduced to account for unobserved heterogeneity in the latent crash propensity). The latent crash propensity y_q^* is mapped to the observed count variable y_q by the thresholds ψ_{qk}, which satisfy the ordering conditions ($\psi_{q,-1} = -\infty; -\infty < \psi_{q0} < \psi_{q1} < \psi_{q2} < ...$) in the usual ordered-response fashion, $f_k(\mathbf{z}_q)$ is a non-linear function of a vector of CBG-specific variables $\tilde{\mathbf{z}}_q$ ($\tilde{\mathbf{z}}_q$ includes a constant), and α_l is a scalar similar to the thresholds in a standard ordered-response model ($\alpha_{-1} = -\infty$; $\alpha_K = 0$ for identification, where *K* is the largest count value observed in the estimation sample). As indicated by Castro et al. (2013), traditional count models do not consider the vector $\mathbf{w}_q$, and only consider the vector $\tilde{\mathbf{z}}_q$. Write $f_k(\tilde{\mathbf{z}}_q) = \Phi^{-1}\left(e^{-\lambda_q}\sum_{l=0}^{k}\frac{\lambda_q^l}{l!}\right)$, so that the thresholds in Equation (9) take the following form:

$$\psi_{qk} = \Phi^{-1}\left(e^{-\lambda_q}\sum_{l=0}^{k}\frac{\lambda_q^l}{l!}\right) + \sum_{l=0}^{k}\alpha_l, \text{ with } \lambda_q = e^{\boldsymbol{\gamma}'\tilde{\mathbf{z}}_q}, \text{ and } \alpha_l = 0 \text{ if } l > K^*, \tag{10}$$

where Φ^{-1} is the inverse function of the univariate cumulative standard normal, $\boldsymbol{\gamma}$ is a coefficient vector to be estimated, and K^* is an appropriate count level that may be determined based on the empirical context under consideration and empirical testing. The presence of the α_l terms provides flexibility to accommodate high or low-probability masses for specific count outcomes without the need for using hurdle or zero-inflated mechanisms. Also, note that $\mathbf{w}_q$ and $\tilde{\mathbf{z}}_q$ can have common elements. The presence of intersection characteristics in $\tilde{\mathbf{z}}_q$ allows CBG with the same latent crash propensity to have different observed crash frequency outcomes.

Equation (9) may be rewritten after some straightforward algebraic manipulations as follows:

$$y_q^* \mid \boldsymbol{\beta}_q = \vartheta\left(\tau_q \mid \boldsymbol{\beta}_q\right) + H_q, \text{ where } H_q \sim N(\mu_q, \upsilon_q^2),\ \mu_q = \boldsymbol{\theta}'\mathbf{w}_q,\ \upsilon_q^2 = \mathbf{w}_q'\boldsymbol{\Xi}\mathbf{w}_q + 1. \tag{11}$$

3.1.4. Model building

To proceed, we need the cumulative distribution function of $y_q^* \mid \boldsymbol{\beta}_q$. The cumulative distribution function of $y_q^* \mid \boldsymbol{\beta}_q$ for CBGs with $k > 0$ may be obtained from the following theorem:

Theorem 1: The distribution of $y_q^* \mid \boldsymbol{\beta}_q$, which includes the stochastic maximum over crash instances and crash risk propensities for each event state at each crash instance, takes the following cumulative distribution function form for $\vartheta > 0$:

$$R_q(u;\tilde{\boldsymbol{\Theta}},\vartheta,\mu_q,\upsilon_q^2) = \left[\int_{h_q=-\infty}^{+\infty} \left[\left(\prod_{t_q=1}^{n_q} F_I \left[\left(\frac{u-h_q}{\vartheta} \right) \mathbf{1}_I ;\, \tilde{\mathbf{V}}_{qt_q}, \tilde{\boldsymbol{\Theta}} \right] \right)^{\frac{1}{n_q}} \right] f_{H_q}\left(h_q;\mu_q,\upsilon_q^2\right) dh_q \right] \Bigg\| \boldsymbol{\beta}_q,\ \tilde{\mathbf{V}}_{qt_q} = \mathbf{x}_{qt_q}\boldsymbol{\beta}_q, \quad (12)$$

where F_I is the multivariate normal cumulative distribution function of dimension *I* with mean $\tilde{\mathbf{V}}_{qt_q}$ and covariance matrix $\boldsymbol{\Theta}$, and f_{H_q} is the univariate normal density function with mean μ_q and variance υ_q^2.

$$\begin{aligned} \text{Proof}: \quad R_q(u;\tilde{\boldsymbol{\Theta}},\vartheta,\mu_q,\upsilon_q^2) &= \text{Prob}\left(y_q^* \mid \boldsymbol{\beta}_q < u\right) = \text{Prob}\left[\, \vartheta\left(\tau_q \mid \boldsymbol{\beta}_q\right) < u - H_q \,\right] \\ &= \text{Prob}\left[\left(\tau_q \mid \boldsymbol{\beta}_q\right) < \frac{u-H_q}{\vartheta} \right] \\ &= \left[\int_{h_q=-\infty}^{+\infty} \left[\left(\prod_{t_q=1}^{n_q} F_I \left[\left(\frac{u-h_q}{\vartheta} \right) \mathbf{1}_I ;\, \tilde{\mathbf{V}}_{qt_q}, \tilde{\boldsymbol{\Theta}} \right] \right)^{\frac{1}{n_q}} \right] f_{H_q}\left(h_q;\mu_q,\upsilon_q^2\right) dh_q \right] \Bigg\| \boldsymbol{\beta}_q \end{aligned} \quad (13)$$

For CBGs with *k*=0, there are no crash-specific variables, but CBG-level exogenous variables are still available. Thus, for such CBGs, Equation (12) holds with $n_q = 1$, and $\tilde{\mathbf{V}}_{qt_q} = \boldsymbol{\varpi}_{qi}\boldsymbol{\beta}_{q\varpi}$, and the conditionality being taken only with respect to the $\boldsymbol{\beta}_{q\varpi}$ vector. Next, the likelihood function from the total count model, given that the observed count level of CBG *q* is n_q conditional on $\boldsymbol{\beta}_q$, may be written as:[7]

[7] If the event state model is analyzed using a simpler MNL model, the likelihood expression below simplifies to a one-dimensional integral:

$$L_{q,count}(\tilde{\boldsymbol{\Theta}},\boldsymbol{\theta},\boldsymbol{\Xi},\boldsymbol{\gamma},\vartheta)\Big|\boldsymbol{\beta}_q = R_q(\psi_{q,n_q};\tilde{\boldsymbol{\Theta}},\vartheta,\mu_q,\upsilon_q^2) - R_q(\psi_{q,n_q-1};\tilde{\boldsymbol{\Theta}},\vartheta,\mu_q,\upsilon_q^2) \quad (14)$$

The likelihood function above involves the computation of an I+1-dimensional integral.[8]

3.2. Model Estimation

The conditional likelihood function for the joint crash frequency-crash category model may be obtained from Equations (5) and (14) as follows:

$$L_q(\tilde{\boldsymbol{\Theta}},\boldsymbol{\theta},\boldsymbol{\Xi},\boldsymbol{\gamma},\vartheta)\Big|\boldsymbol{\beta}_q = \left[\left(L_{q,crash\ event\ state}(\tilde{\boldsymbol{\Theta}})\right)\times\left(L_{q,count}(\tilde{\boldsymbol{\Theta}},\boldsymbol{\theta},\boldsymbol{\Xi},\boldsymbol{\gamma},\vartheta)\right)\right]\Big|\boldsymbol{\beta}_q . \quad (15)$$

Defining $\ddot{\boldsymbol{\beta}}_q = (\boldsymbol{\beta}_q', H_q)'$, $\ddot{\mathbf{b}}_q = (\mathbf{b}', \mu_q)' \ddot{\boldsymbol{\Omega}} = \begin{bmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \upsilon_q^2 \end{bmatrix}$, the unconditional likelihood from CBG q may be rewritten as follows:

$$L_q\left(\tilde{\boldsymbol{\Theta}},\boldsymbol{\theta},\boldsymbol{\Xi},\boldsymbol{\gamma},\vartheta,\mathbf{b},\boldsymbol{\Omega}\right) = \int_{\ddot{\boldsymbol{\beta}}_q} \left\{\left[\left(L_{q,crash\ event\ state}(\tilde{\boldsymbol{\Theta}})\right)\times\left(L_{q,count}(\tilde{\boldsymbol{\Theta}},\boldsymbol{\theta},\boldsymbol{\Xi},\boldsymbol{\gamma},\vartheta)\right)\right]\Big|\ddot{\boldsymbol{\beta}}_q\right\} f_{D+1}(\ddot{\boldsymbol{\beta}}_q \mid \ddot{\mathbf{b}},\ddot{\boldsymbol{\Omega}}) d\ddot{\boldsymbol{\beta}}_q, \quad (16)$$

where $f_{D+1}(.|\mathbf{b},\boldsymbol{\Omega})$ is the multivariate normal density function with mean vector $\ddot{\mathbf{b}}$ and covariance matrix $\ddot{\boldsymbol{\Omega}}$. The integrand in the likelihood function above comprises the evaluation of MVNCD functions of dimension (I–1) in the $L_{q,crash\ event\ state}(\tilde{\boldsymbol{\Theta}})$ component and dimension I in the

---

$$L_{q,count}(\boldsymbol{\theta},\boldsymbol{\Xi},\boldsymbol{\gamma},\vartheta)\Big|\boldsymbol{\beta}_q = \left(\tilde{R}_q(\psi_{q,n_q};\tau_q,\vartheta,\mu_q,\upsilon_q^2) - \tilde{R}_q(\psi_{q,n_q-1};\tau_q,\vartheta,\mu_q,\upsilon_q^2)\right)\Big|\boldsymbol{\beta}_q,$$

$$\tilde{R}_q(\psi_{q,n_q};\vartheta,\mu_q,\upsilon_q^2)\Big|\boldsymbol{\beta}_q = \int_{h_q=-\infty}^{+\infty} \left(\prod_{t_q=1}^{n_q} \left(\exp\left(-\exp\left[-\left\{\left(\frac{\left(\psi_{q,n_q}\right)-h_q}{\vartheta}\right)-\ln\left(\sum_{i=1}^{I}\exp\left(\boldsymbol{\beta}_q'\mathbf{x}_{qt_qi}\right)\right)\right\}\right]\right)\right)^{\frac{1}{n_q}} f_{H_q}\left(h_q;\mu_q,\upsilon_q^2\right) dh_q,$$

[8] If ϑ=0, the contribution from the event state model into the count model ceases to exist, and the count model likelihood expression collapses to:
$L_{q,count}(\boldsymbol{\Xi},\boldsymbol{\gamma},\vartheta)\big|\boldsymbol{\beta}_q = F_q(\psi_{q,n_q};\mu_q,\upsilon_q^2) - F_q(\psi_{q,n_q-1};\mu_q,\upsilon_q^2)$,
where $F_q(\psi_{q,n_q};\mu_q,\upsilon_q^2)$ is the univariate normal cumulative distribution function with the upper threshold at ψ_{q,n_q}, and a mean of μ_q and variance of υ_q^2. This corresponds to the case of an unlinked model. Of course, there is a computational time implication in the estimation of the linked and unlinked models. While the linked model (conditional on $\boldsymbol{\beta}_q$) requires the evaluation of the I+1-dimensional integral in the total crash component (based on Equation 14), the unlinked model (conditional on $\boldsymbol{\beta}_q$) entails only the estimation of a simple univariate normal cumulative distribution function. In our empirical analysis, the net result was that the unlinked model required only about a couple of minutes to estimate, while the corresponding linked model took about a couple of hours to estimate on the same machine. Such extended run times are expected given the multivariate integration involved in jointly estimating the aggregate and disaggregate models together. This is analogous to the increase in runtime observed when comparing a traditional fixed parameters model with its random parameters variant.

$L_{q,count}(\tilde{\boldsymbol{\Theta}},\boldsymbol{\theta},\boldsymbol{\Xi},\boldsymbol{\gamma},\vartheta)$ component. These are evaluated using the accurate analytic two-variate bivariate screening (TVBS) approximation proposed by Bhat, (2018). The integrals of dimension D+1 involved in unconditioning over the entire real line for the $\ddot{\boldsymbol{\beta}}_q = (\boldsymbol{\beta}_q', H_q)'$ vector are evaluated using Halton draws (Bhat, 2003, Halton, 1960).

One additional issue still needs to be dealt with. This concerns the positive definiteness of several matrices in Equation (16). Specifically, for the estimation to work, we need to ensure the positive definiteness of the following matrices: $\boldsymbol{\Omega}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Xi}$. This can be guaranteed in a straightforward fashion using a Cholesky decomposition approach (by parameterizing the function in Equation (16) in terms of the Cholesky-decomposed parameters).

3.3. Use of Model in Forecasting

Once estimated, the model may be used to forecast crash counts by crash event state. To do so, note that $\mathbf{z}_{qt_q{}^i}$ contains variables in categorical form, and we first develop all combinations of these categorical variables and populate these combinations as members of a set A_z. For example, if $\mathbf{z}_{qt_q{}^i}$ includes two time-of-day variables of day and night, and two weather-related variables of clear conditions versus rainy conditions, the cardinality (say CA_z) of the set A_z is four, with the membership being the combinations of a_z=1 (representing the combination of day, clear weather), a_z=2 (combination of day, rainy weather), a_z=3 (representing the combination of night, clear weather), and a_z=4 (representing the combination of night, rainy weather). Then, for each combination a_z (a_z=1,2,3,…, CA_z), assuming that combination a_z is the one that applies for each crash instance t_q, the multivariate probability of counts in each crash event state, conditional on the total count level for CBG q being k_q ($k_q > 0$) and conditional on $\boldsymbol{\beta}_q$, takes the following multinomial distribution form:

$$P[(y_{q1}=k_{q1}),(y_{q2}=k_{q2}),...,(y_{qI}=k_{qI})\,|\,k_q,a_z,\boldsymbol{\beta}_q] = \frac{k_q!}{\prod_{i=1}^{I} k_{qi}!}\prod_{i=1}^{I}(P_{qi}\,|\,a_z,\boldsymbol{\beta}_q). \quad (17)$$

To compute $P_{qi}\left|a_z,\boldsymbol{\beta}_q\right.$, define $\mathbf{R}_{qi}$ $(i=1,2,...,I)$ as an $(I-1)\times I$ matrix that corresponds to an $(I-1)$ identity matrix with an extra column of -1's added as the i^{th} column. Let $\mathbf{G}_{qi}=\mathbf{R}_{qi}\left(\tilde{\boldsymbol{\Theta}}\right)\mathbf{R}'_{qi}$. Then,

$$P_{qi}\left|a_z,\boldsymbol{\beta}_q\right.=P[\mathbf{R}_{qi}\mathbf{S}^*_{qa_z}<\mathbf{0}_{I-1}]\left|\boldsymbol{\beta}_q\right.=\Phi_{(I-1)}\left[(\boldsymbol{\omega}_{\mathbf{G}_{qi}})^{-1}(-\mathbf{x}_{qa_z}\boldsymbol{\beta}_q),(\boldsymbol{\omega}_{\mathbf{G}_{qi}})^{-1}\mathbf{G}_{qi}(\boldsymbol{\omega}_{\mathbf{G}_{qi}})^{-1}\right]. \quad (18)$$

where $\mathbf{S}^*_{qa_z}$ refers to the vector of injury severity risks given that the crash event-specific combination state of a_z applies, and $\mathbf{x}_{qa_z}$ includes the $(I\times D_2)-$intersection-specific matrix $\boldsymbol{\varpi}_q=(\boldsymbol{\varpi}_{q1},\boldsymbol{\varpi}_{q2},...,\boldsymbol{\varpi}_{qI})'$ and the $(I\times D_1)-$CBG/crash-specific matrix of explanatory variables based on the combination a_z. $\boldsymbol{\omega}_{\mathbf{G}_{qi}}$ is the diagonal matrix of standard deviations of $\mathbf{G}_{qi}$.

Next, to obtain the multivariate probability unconditional on the crash state context at each event instance, we write the following mixture based on the probability of occurrence of crash state a_z:[9]

$$P[(y_{q1}=k_{q1}),(y_{q2}=k_{q2}),...,(y_{qI}=k_{qI})\,|\,k_q,\boldsymbol{\beta}_q]=\sum_{a_z=1}^{CA_z}P_{a_z}\frac{k_q!}{\prod_{i=1}^{I}k_{qi}!}\prod_{i=1}^{I}(P_{qi}\left|a_z,\boldsymbol{\beta}_q\right.)^{k_{qi}} \quad (19)$$

The expression above needs the probability of the occurrence of the crash combination a_z. For CBGs in the estimation sample, this is easily computed as the probability of actual occurrence of the crash combination. For CBGs not in the estimation sample, this may be computed as the average of the corresponding probabilities from the CBGs in the estimation sample or using a more stratified grouping based on a similarity index for CBGs and then attributing the average of the probabilities of the appropriate grouping from the estimation sample, or using a separate supplementary model. In our joint crash frequency-crash event state model, the unconditional

[9] We are not aware of the earlier use, within the context of the multinomial distribution, of a discrete mixture probability for accounting for different probabilities across event instances as well as a continuous mixture to accommodate unobserved heterogeneity of the $\boldsymbol{\beta}_q$ coefficient vector. Of course, if crash-specific variables are entirely ignored in the crash count model (that is, the exogenous variable vector is restricted to the same across all crash instances, which is the current state-of-the-art in the crash literature), then (and only then) does the expression below in Equation (19) collapse to the well-known multinomial distribution (in this special case, $CA_z=1$, and P_{a_z}=1).

multivariate probability for any given set of values k_{qi} then takes the form indicated below: $k_q = \sum_{i=1}^{I} k_{qi}$, ($k_{qi} = 0,1,2,...,\infty$, $k_q = 0,1,2,...,\infty$):

$$P[(y_{q1} = k_{q1}),...,(y_{qI} = k_{qI})] = \int_{\boldsymbol{\beta}_q} P[y_q = k_q] \times \left(\sum_{a_z=1}^{CA_z} \left(P_{a_z} \times \frac{k_q!}{\prod_{i=1}^{I} k_{qi}!} \prod_{i=1}^{I} (P_{qi}|a_z, \boldsymbol{\beta}_q)^{k_{qi}} \right) \right) f_D(\boldsymbol{\beta}_q | \mathbf{b}, \boldsymbol{\Omega}) d\boldsymbol{\beta}_q, \quad (20)$$

with $P[y_q = k_q]$ as in Equation (14) after replacing n_q (the actual observed total crash count for CBG q in the estimation sample) with an arbitrary value k_q. Using the properties of the multinomial distribution, the marginal probability of k_{qi} counts for crash event state *i* is:

$$P[y_{qi} = k_{qi}] = \sum_{k_q=k_{qi}}^{\infty} \int_{\boldsymbol{\beta}_q} \left[P[y_q = k_q] \times \left(\sum_{a_z=1}^{CA_z} P_{a_z} \times \left[\frac{k_q!}{k_{qi}!(k_q - k_{qi})!} \left(P_{qi}|a_z, \boldsymbol{\beta}_q\right)^{k_{qi}} \left(1 - \left(P_{qi}|a_z, \boldsymbol{\beta}_q\right)\right)^{\left(k_q - k_{qi}\right)} \right] \right) \right] \quad (21)$$

In the above expression, the upper bound of the summation is $k_q = \infty$, though the probability values fade very rapidly beyond a k_q value of 5. So, the summation may be carried out up to a predetermined threshold (such as say $k_q = 20$) depending on the empirical context.

Equation (20) provides the probability for any given multivariate count category by event state at a CBG (for data fit assessment purposes on the estimation sample, Equation (20) can also be used to compute an average probability of correct prediction across CBGs). Similarly, Equation (21) provides the probability of any specified univariate marginal count for an event state, given a specified total count (again, for data fit assessment purposes on the estimation sample, Equation (21) can also be used to compute an average probability of correct prediction separately for each crash event state). For prediction at any given CBG, one can compute all possible combinations of the multivariate outcomes (from Equation (20)), and then translate those probabilities to a deterministic prediction at the CBG in the usual microsimulation-based fashion by (1) arranging the probabilities sequentially in a linear line between 0 and 1 in a cumulative fashion, (2) picking a random uniform number between 0 and 1, and (3) selecting the combination that corresponds to the random number realization on the cumulative probability line.

4. DATA

4.1. Sample Formation

Our study uses crash data from the city of Austin, located in Central Texas, encompassing a total of 671 CBGs. The data is extracted from the Texas Department of Transportation (TxDOT) crash database between January 1, 2018, and December 31, 2019. We specifically focus on motorized vehicle crashes occurring at intersections and involving two vehicles. For each crash, five levels of injury severity are reported: no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury. Due to the relatively limited number of observations in the fatal crash category, we combine fatal and incapacitating injuries into a single severe injury category. Also, the choice of a two-year aggregation period (2018-2019) was an informed decision based on the spatial resolution of CBG areas. A one-year aggregation often led to many CBGs recording zero or very few crashes, resulting in limited variation to extract a meaningful relationship between crash count and determining variables. Extending the aggregation period to two years allowed us to achieve better variation in crash counts. Finally, to ensure consistency and comparability across the two time periods, we implemented a logarithmic offset of ln(2) in our count model. This adjustment aligns the frequency model to an annual basis, enabling predictions to reflect annual count rates. In total, our dataset comprises 2,757 crashes, distributed across 469 CBGs.

4.2. Outcome Variables

The endogenous outcome variables correspond to the total count of crashes per CBG by the four severity levels. Figure 1 provides an overview of the distribution of crash counts by severity level. The black line represents the number of CBGs with a specific crash count (based on the left-side y-axis), while the color scheme of the stacked bars represents the allocation of the total crashes across the severity levels (based on the right-side y-axis). The figure shows that, among the 671 CBGs, 202 CBGs (30.1%) did not experience any crashes, while 115 CBGs (17.1%) had one crash, and 67 CBGs (10.0%) had two crashes. The remaining specific crash counts were observed in a smaller number of CBGs, as evident from the declining trend revealed in Figure 1. Notably, regardless of the total count of crashes per CBG (the x-axis in Figure 1), non-injury crashes were the most prevalent, accounting for an average of about 45.7% of all crashes. Possible injury crashes constituted an average of 26.0% of all crashes, while the corresponding figures for non-incapacitating injury crashes and severe injury crashes are 25% and 3.3% respectively.

Additionally, Figure 1 indicates a slightly increasing trend in the proportion of severe injury crashes as the total number of crashes per CBG increases. This finding suggests a linkage between injury outcomes and counts in CBGs.

While our current application does not encounter issues associated with high crash counts, our modeling framework is robust and versatile enough to be adapted for high-crash-count scenarios, typically in cases when larger geographic units such as Traffic Analysis Zones (TAZs) are employed as the spatial unit of analysis. In such situations, counts in the upper end of the spectrum may be grouped into bracketed categories without much loss in the accuracy of the estimated relationship.

4.3. Exogenous Variables

The study draws upon an array of data sources to compile a comprehensive set of variables influencing crash occurrence and severity. The exogenous variables and data sources include (a) crash location, time, and conditions as well as the characteristics of the involved parties from TxDOT's CRIS database, (b) road network and built-environment (BE) features from the roadway network inventory database of the Texas Department of Transportation (TxDOT), (c) land-use distribution by type from the City of Austin's Open Data Portal, (d) motorized vehicle ownership data from the U.S. Environment Protection Agency (EPA) Smart Location Database (or SLD; see Chapman et al., 2021, and Ramsey and Bell, 2014) (e) commute mode splits and sociodemographic data from the American Community Survey (ACS) 2021 five-year estimates, and (f) police-reported crime and traffic violations from the City of Austin's Open Data Portal.

The exogeneous variables are categorized into five categories (i) most severely injured individual characteristics, (ii) at-fault vehicle and parties characteristics, (iii) crash time variables, (iv) crash location variables, and (v) CBG level variables. First, the characteristics of the most severely injured individual in the crash, including gender, age, race, and ethnicity, are considered. Second, after determining the at-fault vehicle for each crash, the characteristics of the driver are reported, including their gender, age, race, and intoxication levels. Third, the time-varying covariates associated with each crash include the time of day, categorized as dawn, day, dusk, and night, as well as weather conditions, which include clear, cloudy, fog, rain, and other conditions. Fourth, crash location level variables reflect intersection characteristics and roadway attributes for the major and minor intersection approaches. Intersection characteristics include the number of

intersection legs, and the type of traffic control device. Roadway attributes include the functional classification, number of lanes, and posted speed limit. Fifth, the aggregate CBG level variables are further divided into several subcategories. Road/network BE features include the proportion of roads categorized by different functional classes, the number of lanes, and posted speed limits, as well as the number of intersections with different number of legs and signalized intersections. Note that these road and intersection level data were aggregated to a zonal CBG level using appropriate GIS tools. Land use distribution variables include the proportion of the CBG area corresponding to residential, commercial, office, industrial, civic, open space, utilities, and undeveloped land use types. Crash exposure variables are related to population density, vehicle ownership, and means of transport to work. The sociodemographic variables reflect the racial composition of the CBG, and income levels. Lastly, the crime and traffic violations category corresponds to the crime rate of the CBG and the proportion of yielding, intoxication, red light running, stop sign running, and speeding traffic violations.

Table 1 and Table 2 provide the descriptive statistics for the many variables considered at both the crash and CBG levels, respectively. Table 1 provides the percentage of observations corresponding to each crash-specific exogenous variable within each injury severity category (the percentages in the table are taken row-wise; that is, the percentages sum to 100% for each row). In general, the table indicates that women, older individuals, and minorities are more represented in the pool of severe injury crashes than their peers, while men, young individuals (≤25 years) and older individuals (>60 years), and individuals of Black and Hispanic origin are over-represented as drivers in the pool of severe injury crashes. Also, crashes that take place during clear/cloudy conditions (relative to crashes during rainy conditions) are more likely to be associated with severe injuries, as are crashes during night time (relative to day time). Of course, the table is a simple univariate crosstabulation of each crash-level exogenous variable with injury severity, which does not control for the effect of other variables at the same time. Thus, the relationships listed above should be viewed as mere associations rather than substantive causal effects. Table 2 provides the sample characteristics for other CBG-level exogenous variables.

5. ESTIMATION RESULTS

5.1. Variable Specification

The selection of variables included in the final model specification was based on previous research, intuitiveness, and parsimony considerations. We investigated different functional forms and combinations of explanatory variables. For variables in bracketed form (such as age and time) and those naturally discrete (such as gender, race, driving under the influence, weather, and traffic control variables), we created dummy variables in the most disaggregate form and progressively combined them based on statistical tests. This approach helps achieve parsimonious specifications without sacrificing essential information. In cases where certain levels of a categorical variable lacked sufficient observations, we combined them with other appropriate levels to enhance statistical reliability. Additionally, when two levels showed similar effects, we merged them into one level to simplify the model without compromising its accuracy. For variables in continuous form (most BE and CBG level variables), various functional forms were tested, including a continuous linear form, a continuous logarithm form, a piece-wise linear form, and a set of dummy variables for different ranges. Among the tested forms, the continuous linear form stood out as the most effective and efficient option for most variables. CBG attributes were considered both in the crash frequency model specification and the crash severity model specification.

The final estimation results are presented in Table 3. As may be observed from the table, not all variables included in the model are statistically significant at a 90% confidence level. This is to acknowledge the relatively small sample size used in our estimation, particularly for the severe injury category. In the context of crash severity models, it is common practice to group severity categories to increase the number of data points within each severity category (thus pinning down parameters more precisely and increasing confidence levels). For example, most studies group the five possible severity levels into three levels (refer to Zou et al. (2023), Gong et al. (2022), Yan et al. (2021), and Wu et al. (2014) for examples). While the rationale is understandable, these grouping approaches result in models that are less informative. In our analysis, we opted to maintain four alternatives to retain a more detailed classification of severity levels. Given the small share of the “severe injury” alternative, a low significance level for variables in this category is to be expected. In fact, in scholarly research, retaining variables with lower than the 0.05 level of confidence (t-statistic of 1.96) is not at all an uncommon practice, particularly when dealing with small or unbalanced sample sizes, as with the severe injury category in the current paper. Doing

so has the benefit of identifying variables that are suggestive and that may help inform future specifications with more balanced injury severity categories.

A few other notes about the estimation results in Table 3. We use the label "na" to indicate that the corresponding explanatory variables are not applicable to the outcome of interest. In contrast, a "—" is used to signify that a variable is not statistically significant for a given alternative even at the 75% confidence level. Additionally, we attempted random coefficients on several exogenous variables to control for unobserved heterogeneity effects in the injury severity model, as discussed in Section 3, but none of these turned out to be even marginally statistically significant. Finally, we also investigated the effect of exogenous variables in the latent propensity of the count model (that is, for the presence of exogenous variables in the $\mathbf{w}_q$ vector of Equation (9)), but none turned out to be statistically relevant. As in traditional count models, exogenous variables appeared only in the vector $\tilde{\mathbf{z}}_q$.

5.2. Crash Event State Model

The table is partitioned into two main components: the injury severity model and the crash count model. The second broad column represents the injury severity model component results, showing the impact of each exogenous variable on the crash risk propensity (elements of the $\boldsymbol{\beta}$ vector). The injury severity model was estimated with the "no injury" category as the base. A positive (negative) coefficient specific to a severity category represents an increased (decreased) risk propensity for the specific severity level relative to the "no injury" category. The estimation results of the severity model component demonstrate that the severity of a crash is closely related to various factors, including the characteristics of the individuals involved, crash time and location attributes, and the broader CBG characteristics.

5.2.1. Individual-level variables

The results indicate that gender exerts a notable influence on injury severity, revealing that women are more susceptible to injuries when involved in crashes compared to men. The increasing magnitude of the gender coefficient across various severity levels indicates a higher probability of women being involved in more severe crashes. A similar result has been reported in other studies (see, for example, Bose et al., 2011, and Fu et al., 2021). This trend can be attributed to several reasons, including gender-specific disparities in the effectiveness of vehicle safety devices, as well

as physical differences related to neck anthropometry, strength, and musculature (Bose et al., 2011). At the same time, when women are at-fault in a crash, these crashes tend to have lower severity than when men are responsible (as shown by the negative sign for the female variable under the at-fault vehicle and parties section in Table 3). Women tend to be more cautious drivers, avoiding risky maneuvers such as speeding, tailgating, and aggressive acceleration, all of which may lead to fewer high-impact crashes (see Rhodes and Pivik, 2011, and Song et al., 2021).

Age is another variable that significantly influences crash severity. Individuals who are either younger than 13 years old or older than 60 years old face a higher risk of being involved in a crash with injuries compared to other age groups. In particular, older vehicle occupants are more likely to be involved in an incapacitating/fatal crash. This finding aligns with earlier research by Kabli et al. (2020), and Regev et al. (2018), which also highlight that the presence of physical fragility among these older age groups likely contributes to the higher vulnerability and more severe injury outcomes in crashes. The age of the at-fault driver is associated with a lower likelihood of possible injury crashes.

Consistent with prior research (Adanu and Jones, 2017), the race of vehicle occupants is another significant factor correlated with crash severity. Our results reveal that Black individuals have a higher likelihood of being involved in injury-causing crashes. Specifically, the results in Table 3 indicate that being Black has a positive impact on the propensity of possible, non-incapacitating, and severe injury crashes compared to being white. Several confounding variables may contribute to this finding. From a sociodemographic perspective, on average, Black individuals may face limited access to newer and safer vehicles equipped with advanced safety features, thereby increasing their risk of injury in crashes (Hanks et al., 2018). Disparities in healthcare access and response times could also contribute to differences in injury severity and likelihood of fatality among racial groups (Hanks et al., 2018, Hanchate et al., 2019). Another factor to consider is social resistance, where racial minorities may resist norms perceived to be set by the majority group. A previous study by Factor et al. (2013) found that Black drivers and passengers who exhibited high social resistance were more likely to drive with unbuckled seatbelts, which can impact injury severity in crashes.

Lastly, drivers under the influence of alcohol or drugs are less likely to be involved in possible injury and non-incapacitating crashes. While it might appear counter-intuitive, the result

indicates that based on other crash characteristics, the individual will either sustain no injury or a severe injury.

5.2.2. Crash time variables

Among the temporal factors, crashes occurring during rainy weather are negatively associated with possible and non-incapacitating injury severities. This finding aligns with the many studies that have shown that drivers tend to be more cautious in adverse pavement surface conditions, driving at lower speeds with heightened attention (see, for example, Chen et al., 2016, and Pervaz et al., 2023).

In contrast, crashes occurring at night are more likely to be severe, possibly due to (a) drivers experiencing reduced visibility during nighttime driving, (b) increased fatigue among drivers, and (c) the lower traffic volumes that allow for higher driving speeds (consistent with Marcoux et al., 2018, and Pervaz et al., 2023).

5.2.3. Crash location (intersection) level variables

At the crash location level, the traffic control device present, and the functional class of intersecting road segments were found to significantly impact crash injury severity.

The model results indicate that crashes occurring at intersections with yield signs have a higher probability of resulting in non-incapacitating injuries compared to intersections with no traffic control or other types of traffic signs. This finding highlights the potential safety impacts of driver behavior at yield-controlled intersections. One contributing factor may be related to misunderstandings or misinterpretations of yield sign requirements by some drivers. Uncertainty about when to yield the right-of-way can lead to confusion and conflicting expectations among drivers approaching the intersection, increasing the risk of crashes. Furthermore, the higher speeds and reduced response times associated with drivers failing to yield may exacerbate the severity when crashes do occur. Additionally, it is worth noting that when drivers merge at a yield sign, the predominant crash type is likely angle crashes, known for having more severe injury consequences (Pervaz et al., 2023). However, the relatively slower speed associated with yielding likely prevents these crashes from resulting in the most serious injuries, such as fatalities. As a result, the combination of inherently more dangerous crash types occurring at lower speeds due to the

yielding contributes to the increased incidence of moderate non-incapacitating injuries at intersections with yield signs.

Conversely, intersections connecting two highway road segments experience a reduced likelihood of non-incapacitating injuries. At highway intersections, traffic is usually regulated by traffic signals or stop signs, which control the flow of vehicles and reduce the likelihood of high-speed crashes. Drivers approaching such intersections are more likely to come to a complete stop, ensuring that any potential crashes occur at lower speeds, thereby reducing the likelihood of non-incapacitating injuries.

5.2.4. CBG level variables

The severity of a given crash is significantly influenced by a combination of aggregate built environment, exposure, and sociodemographic variables. Notably, intersection density, proportion of industrial and agricultural land use, proportion of individuals commuting by car, and proportion of low-income households at the CBG level were found to be particularly significant factors.

Crashes occurring in CBGs with a higher intersection density are less likely to result in possible injuries, indicating that crashes at high intersection density CBGs are likely to be associated with injury severity spectrum extremes. This is an intriguing result that deserves additional investigation in future efforts, especially because the proportion of signalized intersections in the CBG has no effect on injury severity (suggesting that signal control at intersections has no bearing on injury severity, but the clustering of intersections does).

Land-use also impacts the crash severity levels. In regions with a higher fraction of industrial and agricultural land use, the likelihood of non-incapacitating and severe injuries is higher. This could be due to higher speed limits, larger vehicles, or less effective traffic control measures in these areas. The results indicate a positive correlation between car usage and crash severity, implying that the risk of more serious crashes rises as more individuals rely on cars for commuting within the CBG. Crashes in lower-income CBGs have a higher likelihood of resulting in serious non-incapacitating injuries. Several factors may contribute to this finding. Lower-income drivers are more likely to drive older, less safe vehicles that lack modern safety features, increasing their vulnerability to sustaining moderate injuries in a crash. Furthermore, infrastructure deficiencies (such as poorly designed roads, unmaintained surfaces, limited traffic control devices

like signals and signs, and insufficient lighting) that are more prevalent in lower-income areas can potentially exacerbate crash severity.

5.2.5. Correlation terms

The correlation results, presented in Table 3, provide the correlation among the risks associated with the injury severity alternatives in differenced form (with the difference taken with respect to the first alternative of "no injury"). As such, this differenced correlation matrix is not interpretable, unless assumptions are made about the correlation among the original four injury severity levels. One such reasonable assumption is that there is not much variance in the "no-injury" risk category and that unobserved factors that increase the chances of no-injury do not have any bearing on the risks of other injury severity categories. In this case, the differenced correlation directly represents the correlation matrix for the three higher injury risk categories. Specifically, as the crash risk of possible injury increases due to unobserved factors, so does the crash risk of non-incapacitating injury. On the other hand, as the crash risk of possible or non-capacitating injury increases, the crash risk of severe injury decreases.

5.3. Crash Frequency Model

The last column presents the results for the crash count model component. The exogenous effects in the count model correspond to the non-constant elements of the $\boldsymbol{\gamma}$ vector, which directly influence the count model after accounting for any indirect effects through the linking function. Regarding the direct effects, a positive coefficient in $\boldsymbol{\gamma}$ shifts the threshold to the right on the propensity scale, leading to a reduced probability of low or zero crash outcome. Conversely, a negative coefficient shifts the threshold to the left on the propensity scale, increasing the probability of low or zero crash outcomes. The antepenultimate row section of Table 3 provides the constant estimates corresponding to the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ vectors, while the penultimate section provides the threshold shifter terms (elements of the $\boldsymbol{\alpha}$ vector) embedded in the thresholds of the count model. The constants do not have any substantive interpretation and are primarily responsible for optimally mapping the latent propensity to the observed counts, given the coefficients on other variables embedded in the threshold function. Similarly, the threshold shifter elements of the vector $\boldsymbol{\alpha}$ do not have any substantive interpretation, though they allow the count model to flexibly accommodate high or low probability masses for specific outcomes. In the

current empirical analysis, the best model specification was achieved with one predominantly positive threshold shifter term between 0 and 1 counts, reflecting the significant number of CBGs with zero crashes. The reader will note that an offset of two was used in the count model to represent the crash data from two years.

5.3.1. CBG level variables

Our results, summarized in Table 3, suggest that a multitude of factors, including roadway characteristics, population density, sociodemographic variables, and crime rates, influence CBG crash counts.

The number of intersections per square mile is found to be positively associated with crash counts, suggesting that as intersection density increases so does the likelihood of crashes. Intersections contribute to increased navigational complexity and growth in conflict points. Similarly, a higher proportion of four or more leg intersections also positively correlates with the number of crashes in CBG. The complex vehicle interactions and turning movements associated with multi-leg crossings may lead to increased crash risk due to intersection complexity and related driver confusion (see Wang and Huang, 2016, Lee et al., 2017, and Pervaz et al., 2023). In contrast, a higher density of signalized intersections reduces the overall crash count. This outcome is likely attributable to the role these signals play in enforcing precise right-of-way protocols during conflicting vehicular movements.

Several critical patterns emerge when analyzing the results for roadway design variables. While a higher number of highway centerline miles increases the risk of crashes, a higher roadway density (road length relative to CBG area) lowers crashes. A plausible explanation is that a greater prevalence of roadways within an area naturally leads to higher exposure and, consequently, increased risk of crashes, while a dense road network induces slower vehicle speeds due to congestion and more organized traffic flow, thereby reducing the likelihood of crashes (see also Zeng et al., 2019). Another notable observation is that a higher proportion of freeway miles is associated with lower crash counts. Despite the high operating speeds on freeways, their design incorporates limited access points and controlled traffic flow, consequently leading to reduced clashes between traffic streams. Meanwhile, the proportions of principal and minor arterial miles significantly correlates with increased crash counts. Various attributes of arterial roads, such as elevated speeds and traffic volumes, alongside the prevalence of access points and traffic

interruptions, collectively lead to more conflicts among road users and subsequently increase the likelihood of crashes (Guerra et al., 2019). However, the proportion of minor arterial miles shows a less significant positive relationship with crash counts, suggesting that these roads, which usually experience less traffic and lower speeds, contribute less to the crash counts.

In the context of crash exposure factors, population density has a direct impact on total CBG crash counts. Higher population densities often imply more road users - motorists, cyclists, and pedestrians, thus increasing the likelihood of crashes, as suggested by Lee et al. (2017), Yasmin and Eluru (2018), and others.

With respect to sociodemographic variables, the proportion of white non-Hispanic individuals is positively related to crash counts. This association may be influenced by the number of vehicles per household, trip patterns, or other unobserved factors and requires further investigation. Notably, the proportion of high-income households displays a negative relationship with crash counts (coefficient -0.083). This may be due to factors such as safer vehicles, better road awareness, or adherence to traffic rules in such demographics (Lee et al., 2017).

Lastly, the crime rate is also a significant determinant of crash counts, evidenced by a strong positive relationship. This could be attributed to the fact that areas with high crime rates might generally have lower levels of law enforcement, resulting in more traffic violations and hence higher crash rates.

5.3.2. Linking parameter

The results indicate a significant and positive linkage parameter ϑ, providing strong evidence that disaggregate crash-specific factors indeed influence the total crash count at the aggregate level.

5.4. Measures of Fit

The log-likelihood at convergence of the linked model proposed in this paper is -4641.26. The corresponding log-likelihood of an unlinked model (that is, with ϑ=0) is -4823.44. A nested likelihood ratio test between the two models yields a value of 364.36, which is much higher than the critical chi-squared value with one degree of freedom at any reasonable level of significance, clearly rejecting the unlinked model is favor of our proposed linked model. For completeness, a naïve model with only the alternative specific constants in the injury severity MNP model (with independent and identically distributed error terms across alternatives), and with only the constant

in the γ vector and the 0|1 threshold shifter term in the crash count model, has a convergent log-likelihood value of -4868.44.

Following the procedure discussed in Section 3.3, we are also able to evaluate the model fit of our proposed model at the disaggregate level using an average probability of correct prediction (APCP) statistic. We present the APCP values only for the marginal univariate crash count distributions by injury severity, because there are too many multivariate crash count combinations across injury severities. These marginal APCP values for the observed counts by injury severity (averaged across all CBGs) are as follows: no injury -- 0.453, possible injury -- 0.534, non-incapacitating injury -- 0.560, and severe/fatal injury -- 0.870. In addition to the marginal APCP at the disaggregate CBG-level, at the aggregate level, we design a heuristic diagnostic check of model fit by computing the predicted number of CBGs for the crash count values of 0, 1, and 2+ in each injury severity level category. To evaluate the performance of the model proposed here, we compute the absolute percentage error (APE) statistic for each count value (of 0, 1, and 2+) for each injury severity level (as the difference between the predicted and observed values of the number of CBGs in each count category of each injury severity state as a percentage of the corresponding observed values), and then compute a mean weighted APE value across the count values (of 0, 1, and 2+) using the observed number for each count value as the weight for that count value. The results are presented in Table 4. The predicted values are closely aligned with the observed values with only a 4.54% weighted absolute percentage error.

5.5. Elasticity Effects and Implications

The coefficients in Table 3 provide the exogenous variable effects on the disaggregate propensities of the injury severity levels as well as the crash count model; however, they do not directly provide a sense of the direction/magnitude effects of each variable on these outcomes in terms of their impact on the overall shares and count. Therefore, we compute the "pseudo" elasticity effects of the exogenous variables to characterize the impact of each variable. For each of the binary category variables used for the disaggregate crash event state analysis, we first predict the average share of each disaggregate injury severity category in the sample for the "base" level (which is typically the "0" for a binary variable), and then predict the average shares for the "treatment" level (which is typically the "1" for a binary variable) for the entire sample. The average "pseudo" elasticity effect is then reported as the difference between the "treatment" and the "base" shares as a

percentage of the "base" share. For the crash count level variables which are primarily in the continuous forms of fractions, densities, or proportions, we assume reasonable continuous values for the "base" and "treatment" levels to compute the elasticity effects.

Table 5 provides the "pseudo" elasticity effects for our proposed model for selected exogenous variables. The numbers in the table may be interpreted as the percentage change in the shares of each injury severity category and total crashes due to a change in the exogenous variable. For example, the first percentage numeric entry of -65% in the table indicates that women are 65% less likely to be involved in a crash resulting in "No Injury" relative to men. Other numerical entries in Table 5 may be interpreted in a similar manner.

The following observations can be drawn from the results. First, the proposed model framework allows for the influence of disaggregate level variables on crash count and severity distribution (see the top panel of Table 5). In the unlinked model, none of the crash-specific variables in the top panel of Table 5 would have any effect on total crashes. Second, women are less likely to be represented in the pool of no-injury crashes relative to men, but are also less likely to be involved in vehicular crashes in the overall. As just mentioned, a multivariate count model or an unlinked model would not have picked up the effect of gender on total CBG crash count. Our empirical results suggest the need for gender-focused measures (such as perhaps reducing aggressive driving or targeted defensive driving courses) to reduce crashes. Third, "Driving under the influence" (DUI) increases overall crash risk by 7%, but increases the risk of severe injury crashes by 70%. This finding supports the need for stringent DUI enforcement and awareness programs to reduce severe crashes as well as the total crash risk. Fourth, similar to DUI effects, rainy conditions also increase overall crash risk only marginally (by 4%), but increase severe crash risk by 25%. This points to the necessity of targeted interventions during adverse weather conditions, such as enhanced road surface treatments and driver awareness campaigns. Fifth, among CBG attributes affecting crash counts (see the bottom panel of Table 5), road design attributes such as the proportion of freeway miles, principal arterial miles, and minor arterial miles have a modest absolute influence on the total crash count with a 2%-6% impact on the crash propensity. Population density and crime rate variables also yield similar figures. This finding can guide infrastructure development and improvement plans, ensuring that they are balanced with other safety measures. Sixth, however, an increase in the proportion of high-income households in a CBG is associated with a notable 6% decrease in crash propensity. This highlights a complex

interplay between socioeconomic factors and road safety, suggesting that interventions should be tailored to address specific community needs and characteristics. Specifically, there is a clear need to continue efforts to examine and address urban crash safety issues through the lens of equity, given the disparity in transportation infrastructure facilities between low-income/minority and high income/majority white neighborhoods (see, for example, Haddad et al. (2023), and Yu et al. (2022)). Overall, the results highlight the kinds of policy insights that may be drawn from the integrated model framework, particularly because variables that are crash-specific also factor in the prediction of total crash counts.

6. CONCLUSIONS

Crash frequency analysis is an integral component of safety research, traditionally undertaken by aggregating crash data over specific spatial and temporal scales. Given that the determinants of crash counts can significantly vary by crash type and severity, many studies have focused on examining specific crash types, such as road user type, injury type, impact type, or vehicle type and number. Despite these advancements, most studies often rely on univariate count models, independently estimating different crash types. While some researchers have implemented multivariate count models in crash analysis, these approaches often come with computational challenges and potential discrepancies in reflecting the ground reality between disaggregate and aggregate crash outcomes.

In response to these challenges, we have proposed an integrated parametric framework for multivariate crash count data, based on linking a univariate count model for the total count of crashes across all possible crash type states with a discrete choice model for crash event state given a crash. This structure allows us to incorporate the highest event state risk propensity from the event state model as an explanatory variable in the total crash count model, creating a more realistic representation of crash propensities and total crash count intensity. Our proposed framework employs a Generalized Ordered Response Probit (GORP) model at the aggregate level and a multinomial probit model (MNP) at the disaggregate level to examine crash severity.

We applied this approach in a demonstration exercise examining motor vehicle crashes in Census Block Groups (CBGs) in Austin, Texas, based on four injury severity levels. The data for this analysis was sourced from the Texas Department of Transportation crash incident files. The model estimation results are augmented with a host of disaggregate and aggregate data fit measures

and “pseudo” elasticity analysis. Overall, the results underscore the utility of the proposed model in determining the critical disaggregate level factors contributing to total crashes and crash severity increases. This new framework has potential implications for enhancing the precision and predictive power of crash frequency models, thereby potentially informing more effective safety measures.

ACKNOWLEDGEMENTS

This research was partially supported by the Center for Understanding Future Travel Behavior and Demand (TBD), a National University Transportation Center sponsored by the US Department of Transportation under grant 69A3552344815 and 69A3552348320. The authors are grateful to Lisa Macias for her help in formatting this document. Two anonymous reviewers provided useful comments on an earlier version of this paper.

REFERENCES

Abdel-Aty, M., Siddiqui, C., Huang, H., Wang, X., 2011. Integrating trip and roadway characteristics to manage safety in traffic analysis zones. *Transportation Research Record* 2213, 20–28.

Adanu, E. K., Jones, S., 2017. Effects of human-centered factors on crash injury severities. *Journal of Advanced Transportation* 2017, e1208170.

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.

Anastasopoulos, P. C., Mannering, F. L., 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. *Accident Analysis and Prevention* 43(3), 1140–1147.

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

Bermúdez, L., Karlis, D., 2011. Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics* 48(2), 226–236.

Bethell, J., Rhodes, A. E., Bondy, S. J., Lou, W. Y. W., Guttmann, A., 2010. Repeat self-harm: Application of hurdle models. *The British Journal of Psychiatry* 196(3), 243–244.

Bhat, C. R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B* 37(9), 837–855.

Bhat, C. R., 2015. A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B* 79, 50–77.

Bhat, C. R., 2018. New matrix-based methods for the analytic evaluation of the multivariate cumulative normal distribution function. *Transportation Research Part B* 109, 238–256.

Bhat, C. R., Castro, M., Khan, M., 2013. A new estimation approach for the multiple discrete–continuous probit (MDCP) choice model. *Transportation Research Part B* 55, 1–22.

Bhat, C. R., Pulugurta, V., 1998. A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B* 32(1), 61–75.

Bhowmik, T., Yasmin, S., Eluru, N., 2021. 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.

Bose, D., Segui-Gomez, S., Maria, Crandall, J. R., 2011. Vulnerability of female drivers involved in motor vehicle crashes: An analysis of US population at risk. *American Journal of Public Health* 101(12), 2368–2373.

Buck, A. J., Blackstone, E. A., Hakim, S., 2009. A multivariate poisson model of consumer choice in a multi-airport region. *IBusiness* 1(2), 85–98.

Cai, Q., Lee, J., Eluru, N., Abdel-Aty, M., 2016. Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. *Accident Analysis and Prevention* 93, 14–22.

Castro, M., Paleti, R., 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.

Chapman, J., Fox, E. H., Bachman, W., Frank, L. D., 2021. Smart Location Database: Version 3.0 Technical Documentation and User Guide. U.S. Environment Protection Agency. www.epa.gov/sites/default/files/2021-06/documents/epa_sld_3.0_technicaldocumentationuserguide_may2021.pdf

Chen, C., Zhang, G., Liu, X. C., Ci, Y., Huang, H., Ma, J., Chen, Y., Guan, H., 2016. Driver injury severity outcome analysis in rural interstate highway crashes: A two-level Bayesian logistic regression interpretation. *Accident Analysis and Prevention* 97, 69–78.

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

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.

Factor, R., Williams, D. R., Kawachi, I., 2013. Social resistance framework for understanding high-risk behavior among nondominant minorities: Preliminary evidence. *American Journal of Public Health* 103(12), 2245–2251.

Fu, J., Abdel-Aty, M., Mahmoud, N., 2023. Time-specific hierarchical models for predicting crash frequency of reversible and high-occupancy vehicle lanes. *Accident Analysis and Prevention* 181, 106953.

Fu, W., Lee, J., Huang, H., 2021. How has the injury severity by gender changed after using female dummy in vehicle testing? Evidence from Florida's crash data. *Journal of Transport and Health* 21, 101073.

Gong, H., Fu, T., Sun, Y., Guo, Z., Cong, L., Hu, W., Ling, Z., 2022. Two-vehicle driver-injury severity: A multivariate random parameters logit approach. *Analytic Methods in Accident Research* 33, 100190.

Guerra, E., Dong, X., Kondo, M., 2019. Do denser neighborhoods have safer streets? Population density and traffic safety in the Philadelphia region. *Journal of Planning Education and Research* 0739456X19845043.

Haddad, A. J., Mondal, A., Bhat, C. R., Zhang, A., Liao, M. C., Macias, L. J., Kyung Lee, M., Watkins, S. C., 2023. Pedestrian crash frequency: Unpacking the effects of contributing factors and racial disparities. *Accident Analysis and Prevention* 182, 106954.

Halton, J. H., 1960. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik* 2, 84–90.

Hanchate, A. D., Paasche-Orlow, M. K., Baker, W. E., Lin, M.-Y., Banerjee, S., Feldman, J., 2019. Association of race/ethnicity with emergency department destination of emergency medical services transport. *JAMA Network Open* 2(9), e1910816.

Hanks, A., Solomon, D., Weller, C., 2018. Systematic Inequality: How America's Structural Racism Helped Create the Black-White Wealth Gap. Center for American Progress. www.americanprogress.org/article/systematic-inequality/

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

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

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.

Kim, D.-G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models. *Accident Analysis and Prevention* 39(1), 125–134.

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. *Accident Analysis and Prevention* 102, 213–226.

Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis and Prevention* 38(4), 751–766.

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.

Milton, J. C., Shankar, V. N., Mannering, F. L., 2008. Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis and Prevention* 40(1), 260–266.

Musio, M., Sauleau, E. A., Buemi, A., 2010. Bayesian semi-parametric ZIP models with space–time interactions: An application to cancer registry data. *Mathematical Medicine and Biology: A Journal of the IMA* 27(2), 181–194.

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

National Center for Statistics and Analysis, 2023. Early estimate of motor vehicle traffic fatalities for the first quarter of 2023. Crash Stats Brief Statistical Summary. Report No. DOT HS 813 482. crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813482

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

Pervaz, S., Bhowmik, T., Eluru, N., 2023. An econometric framework for integrating aggregate and disaggregate level crash analysis. *Analytic Methods in Accident Research* 39, 100280.

Ramsey, K., Bell, A., 2014. Smart location database. Washington, DC.

Regev, S., Rolison, J. J., Moutari, S., 2018. Crash risk by driver age, gender, and time of day using a new exposure methodology. *Journal of Safety Research* 66, 131–140.

Rhodes, N., Pivik, K., 2011. Age and gender differences in risky driving: The roles of positive affect and risk perception. *Accident Analysis and Prevention* 43(3), 923–931.

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.

Shin, K., Washington, S. P., 2012. Empirical Bayes method in the study of traffic safety via heterogeneous negative multinomial model. *Transportmetrica* 8(2), 131–147.

Sivakumar, A., Bhat, C., 2002. Fractional split-distribution model for statewide commodity-flow analysis. *Transportation Research Record* 1790, 80–88.

Song, X., Yin, Y., Cao, H., Zhao, S., Li, M., Yi, B., 2021. The mediating effect of driver characteristics on risky driving behaviors moderated by gender, and the classification model of driver's driving risk. *Accident Analysis and Prevention* 153, 106038.

Terza, J. V., Wilson, P. W., 1990. Analyzing frequencies of several types of events: A mixed multinomial-Poisson approach. *The Review of Economics and Statistics* 72(1), 108–115.

Wu, Q., Chen, F., Zhang, G., Liu, X. C., Wang, H., Bogus, S. M., 2014. Mixed logit model-based driver injury severity investigations in single-and multi-vehicle crashes on rural two-lane highways. *Accident Analysis and Prevention* 72, 105–115.

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

Yan, X., He, J., Zhang, C., Liu, Z., Wang, C., Qiao, B., 2021. Temporal analysis of crash severities involving male and female drivers: A random parameters approach with heterogeneity in means and variances. *Analytic Methods in Accident Research* 30, 100161.

Yamamoto, T., Hashiji, J., Shankar, V. N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accident Analysis and Prevention* 40(4), 1320–1329.

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

Yu, C.-Y., Zhu, X., Lee, C., 2022. Income and racial disparity and the role of the built environment in pedestrian injuries. *Journal of Planning Education and Research* 42(2), 136–149.

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

Ziakopoulos, A., Yannis, G., 2020. A review of spatial approaches in road safety. *Accident Analysis and Prevention* 135, 105323.

Zou, R., Yang, H., Yu, W., Yu, H., Chen, C., Zhang, G., Ma, D. T., 2023. Analyzing driver injury severity in two-vehicle rear-end crashes considering leading-following configurations based on passenger car and light truck involvement. *Accident Analysis and Prevention* 193, 107298.

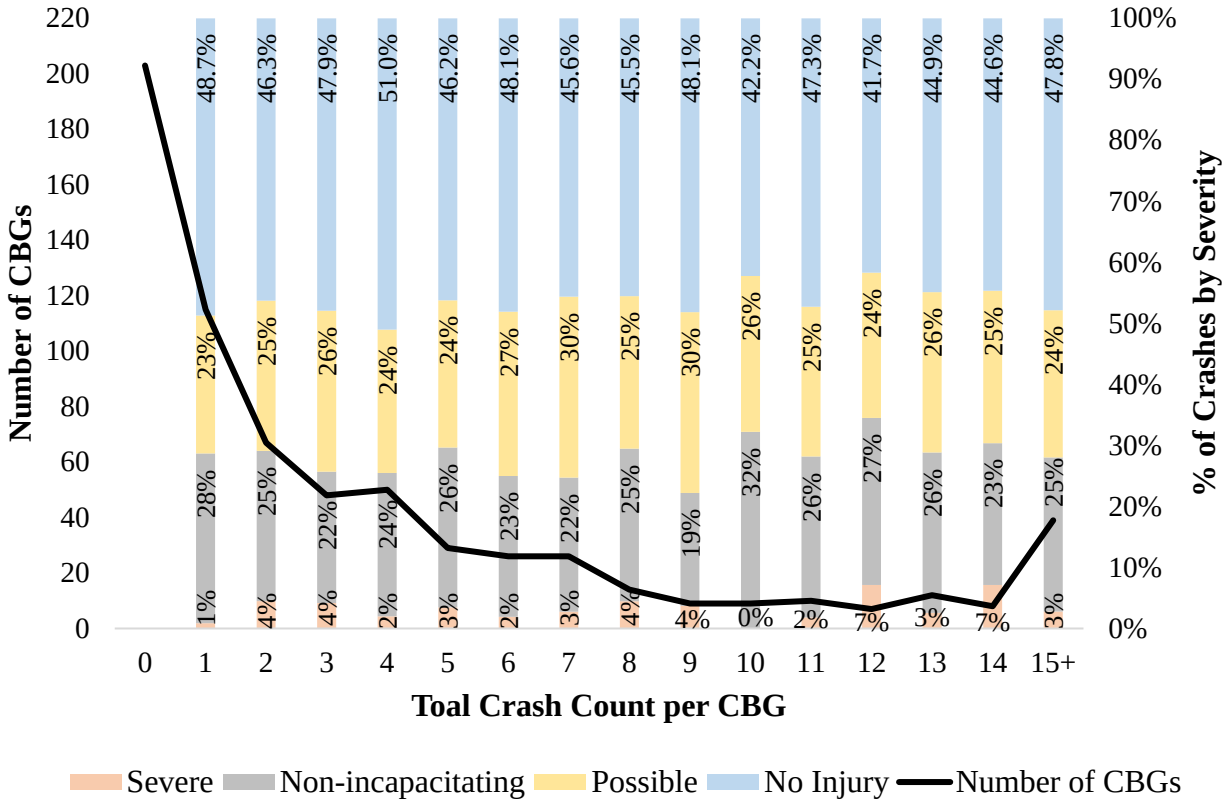

Figure 1. Distribution of crashes across CBGs and severity levels

Table 1. Summary Statistics of Exogenous Variables at the Crash Level

Variable	Percentage of observations in each injury severity level (%)			
	No Injury	Possible	Non-incapacitating	Severe
Most severely injured individual				
Gender				
Male	56.52	20.08	21.11	2.29
Female	40.15	29.29	26.61	3.95
Age				
<13 years old	2.56	53.85	41.03	2.56
13-25 years old	48.53	22.18	25.94	3.35
26-60 years old	48.08	25.11	23.91	2.90
>60 years old	36.09	25.65	32.17	6.09
Race/Ethnicity				
White	55.20	22.40	19.62	2.78
Black	38.10	31.57	26.32	4.01
Hispanic	44.26	26.00	26.45	3.29
Other	58.00	14.00	26.00	2.00
At-fault vehicle and parties				
Gender				
Male	49.93	23.48	23.13	3.46
Female	45.16	26.81	25.13	2.90
Age				
≤25 years old	51.59	25.26	19.52	3.63
26-60 years old	46.70	25.86	24.53	2.91
>60 years old	40.19	23.83	32.24	3.74
Race/Ethnicity				
White	54.46	23.5	19.13	2.91
Black	41.86	28.8	25.87	3.47
Hispanic	43.17	25.72	27.37	3.74
Other	54.54	14.55	30.91	0.00
Driver under the influence				
Yes	80.48	7.32	9.76	2.44
No	47.16	25.33	24.30	3.21
Crash Time				
Weather				
Clear	47.84	25.12	23.85	3.19
Cloud	40.30	27.65	28.17	3.88
Rain	60.73	19.63	17.81	1.83
Other	20.00	20.00	60.00	0.00
Time				
Day	46.43	26.28	24.35	2.94
Night	50.76	22.01	23.41	3.82

Table 2. Summary Statistics of Exogenous Variables at the CBG level

Variable	Definition	Mean	Std. dev.
Road/network BE features			
Intersections			
# intersections per mi^2 (in 100)	Number of intersections/ total area of CBG*100	1.003	0.651
Fraction of four or more–leg intersections	Number of for or more-leg intersections/Number of intersections in CBG	0.040	0.077
Fraction of signalized intersections	Number of traffic signals/Number of intersections in CBG	0.133	0.179
General roadways			
Total road miles in CBG - mi (in 10)	Sum of roadway centerline miles in CBG / 10	0.425	0.357
Road density mi/mi^2 (in 100)	Total road miles in CBG/ total area of CBG*100	0.234	0.227
Functional class			
Proportion of freeway miles	Total freeway miles in CBG/ total road miles in CBG	0.116	0.111
Proportion of principal arterial miles	Total principal arterial miles in CBG/ total road miles in CBG	0.039	0.099
Proportion of minor arterial miles	Total minor arterial miles in CBG/ total road miles in CBG	0.097	0.143
Proportion of major collector miles	Total major collector miles in CBG/ total road miles in CBG	0.151	0.151
Proportion of minor collector miles	Total minor collector miles in CBG/ total road miles in CBG	0.014	0.044
Proportion of local miles	Total local miles in CBG/ total road miles in CBG	0.583	0.280
Number of lanes			
Proportion of one-lane road miles	Total one-lane miles in CBG/ total road miles in CBG	0.034	0.081
Proportion of two-lane road miles	Total two-lane miles in CBG/ total road miles in CBG	0.668	0.274
Proportion of three-lane road miles	Total three-lane miles in CBG/ total road miles in CBG	0.096	0.148
Proportion of four or more-lane road miles	Total four or more-lane miles in CBG/ total road miles in CBG	0.202	0.183
Land use			
Proportion of residential land use	Total area of residential land use/ total area of the CBG	0.483	0.293
Proportion of commercial land use	Total area of commercial land use/ total area of the CBG	0.117	0.136
Proportion of office land use	Total area of office land use/ total area of the CBG	0.064	0.106
Proportion of industrial land use	Total area of industrial land use/ total area of the CBG	0.058	0.118
Proportion of civic land use	Total area of civic land use/ total area of the CBG	0.082	0.137
Proportion of open space land use	Total area of open space land use/ total area of the CBG	0.095	0.143
Proportion of utility land use	Total area of utility land use/ total area of the CBG	0.026	0.087
Proportion of undeveloped or agricultural land use	Total area of undeveloped/agricultural LU/ total area of CBG	0.075	0.124
Crash exposure factors			
Population Density - people/acre (in 10)	Gross population density (people/acre)	0.085	0.072
Modes of transport to work			
Proportion of individuals who drive to work	Proportion of workers aged 16 years and over who commute by a private vehicle in the CBG	0.876	0.16

Table 2. Summary Statistics of Exogenous Variables at the CBG level (Cond.)

Variable	Definition	Mean	Std. dev.
Road/network BE features			
Proportion of individuals who use public transit to get to work	Proportion of workers who commute by public transit in CBG	0.034	0.05
Proportion of individuals who walk to work	Proportion of workers who commute by foot in CBG	0.044	0.09
Vehicle ownership			
Proportion of HH owning zero vehicles	Proportion of zero-car households in CBG, 2018	0.080	0.081
Proportion of HH owning one vehicle	Proportion of one-car households in CBG	0.429	0.143
Proportion of HH owning two or more vehicles	Proportion of two or more-car households in CBG	0.491	0.169
Sociodemographic variables			
Racial/ethnic composition			
Proportion of CBG population that is not Hispanic	--	0.651	0.236
Proportion of CBG population that is not Hispanic white	--	0.456	0.233
Proportion of CBG population that is not Hispanic Black	--	0.090	0.102
Proportion of CBG population that is not Hispanic other	--	0.105	0.101
Proportion of CBG population that is Hispanic	--	0.349	0.236
Proportion of CBG population that is Hispanic white	--	0.184	0.143
Proportion of CBG population that is Hispanic Black	--	0.006	0.026
Proportion of CBG population that is Hispanic other	--	0.159	0.160
Income levels			
Proportion of low-income HH	Proportion of workers earning $1250/month or less	0.196	0.046
Proportion of medium-income HH	Proportion of workers earning more than $1250/month bus less than $3333/month	0.311	0.086
Proportion of high-income HH	Proportion of workers earning $3333/month or more	0.493	0.123
Crime and traffic violations			
Crime rate – total crimes/total population (in 10)	Total police-reported crimes from 2003 till 2022 / total population in CBG*10	0.319	0.358
Proportion of cases where drivers failed to yield	Total number of failure to yield cases/total number of violations charged in CBG from 2018 till 2022	0.040	0.067
Proportion of cases where drivers were intoxicated	Total number of public intoxication cases/total number of violations charged in CBG from 2018 till 2022	0.041	0.075
Proportion of cases where drivers ran a red light	Total number of red light running cases/total number of violations charged in CBG from 2018 till 2022	0.020	0.094
Proportion of cases where drivers ran a stop sign	Total number of stop sign running cases/total number of violations charged in CBG from 2018 till 2022	0.005	0.03
Proportion of cases where drivers were speeding	Total number of speeding cases/total number of violations charged in CBG from 2018 till 2022	0.094	0.154

Table 3. Model estimation results (N = 2,959)

Exogenous Variables	Severity Level (base: No Injury)						Total count of crashes in CBG	
	Possible Injury		Non-incapacitating Injury		Severe Injury			
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Most severely injured party								
Gender (Base: Male)								
Female	0.846	5.95	0.780	3.48	0.923	1.99	na	na
Age (Base: 13-60 years old)								
<13 years old	1.854	6.60	1.755	5.58	–	–	na	na
>60 years old	0.378	2.17	0.436	2.64	0.601	1.34	na	na
Race (Base: Not Black)								
Black	0.352	2.93	0.352	1.69	0.352	1.69	na	na
At-fault Vehicle and Parties								
Driver gender (Base: Male)								
Female	-0.331	-4.49	-0.306	-2.56	-0.587	-1.40	na	na
Driver age (Base: ≤60 years old)								
>60 years old	-0.096	-1.63	–	–			na	na
Driver under the influence (Base: No)								
Yes	-0.854	-2.56	-0.743	-2.99	–	–	na	na
Crash Time								
Weather (Base: Not rain)								
Rain	-0.185	-1.79	-0.197	-2.47	–	–	na	na
Time (Base: Not night)								
Night	–	–	–	–	0.110	1.21	na	na
Crash Location (Intersection) Level Variables								
Traffic control (Base: None or other devices)								
Yield Sign	–	–	0.187	1.85	–	–	na	na
Intersecting road segments (Base: not two intersecting highway segments)								
Two intersecting highway segments	–	–	-0.121	-1.66	–	–	na	na
CBG Level Variables								
Road design								
# intersections per mi^2 (in 100)	-0.027	-1.59	–	–	–	–	0.016	2.08
Proportion of four or more–leg intersections	–	–	–	–	–	–	0.427	2.02
Proportion of signalized intersections	–	–	–	–	–	–	-0.091	-1.50

Table 3. Model estimation results (N = 2,959) (Contd.)

Exogenous Variables	Severity Level (base: No Injury)						Total count of crashes in CBG	
	Possible Injury		Non-incapacitating Injury		Severe Injury			
	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat	Coef.	t-stat
Total road miles (in 10)	–	–	–	–	–	–	0.150	2.00
Road density mi/mi^2 (in 100)	–	–	–	–	–	–	-0.080	-1.46
Proportion of freeway miles	–	–	–	–	–	–	-0.749	-2.36
Proportion of principal arterial miles	–	–	–	–	–		0.115	2.04
Proportion of minor arterial miles	–	–	–	–	–	–	0.075	1.62
Land-use (Base: Residential, commercial, office, civic and other land-use types)								
Fraction of industrial and agricultural LU	–	–	0.3297	2.21	0.864	1.55	–	–
Crash exposure factors								
Population density - people/acre (in 10)	–	–	–	–	–	–	0.109	1.98
Proportion of individuals commuting by car	–	–	0.363	2.12	–	–	–	–
Sociodemographic variables								
Proportion of white non-Hispanic individuals	–	–	–	–	–	–	0.146	1.58
Proportion of low-income HH	–	–	0.624	3.01	–	–	–	–
Proportion of high-income HH	–	–	–	–	–	–	-0.072	-2.17
Crime and traffic violations								
Crime rate – total crimes/total population (in 10)	–	–	–	–	–	–	0.090	2.29
Linking parameter	na	na	na	na	na	na	2.788	5.12
Constants	-0.534	-5.56	-0.977	-4.98	-1.998	-1.95	2.284	10.91
Threshold shifter terms								
0|1	na	na	na	na	na	na	6.192	11.53
Correlation Terms								
Possible Injury		1.00#		0.69*		-0.26*		
Non-incapacitating Injury				1.00#		-0.28*		
Severe Injury						1.00#		

* Correlation terms are statistically significant at 85% confidence level

Fixed scales (for imparting stability to the estimation routine)

Table 4. Aggregate data fit measures

Marginal count category shares			
Severity level	**Count category**	**Observed**	**Predicted**
No injury	0	294	310
	1	133	113
	2+	244	248
Possible	0	383	389
	1	135	142
	2+	153	140
Incapacitating	0	375	385
	1	157	138
	2+	139	148
Severe	0	599	590
	1	60	65
	2+	12	16
Weighted Absolute Percentage Error (WAPE)		**4.54%**	

Table 5. "Pseudo" elasticity effects

Variables	Base Level	Treatment Level	Injury Severity				Total count of crashes in CBG
			No-injury	Possible	Non-incapaci-tating	Severe	
Disaggregate Level Variables							
Most severely injured party							
Gender	Male	Female	-65%	50%	39%	237%	-13%
Race	Non-Black	Black	-16%	25%	3%	15%	-7%
At-fault Vehicle and Parties							
Driver gender	Male	Female	24%	-28%	-9%	-58%	12%
Driver age	<60 years old	>60 years old	5%	-24%	16%	4%	1%
Driver under the influence	No	Yes	55%	-58%	-50%	70%	7%
Crash Time							
Weather	No Rain	Rain	18%	-14%	-23%	25%	4%
Time of Day	Day	Night	-1%	-2%	-1%	28%	0%
Aggregate (CBG) Level Variables							
Road design							
Proportion of freeway miles	Proportion of Freeways=0.1; Principal arterials=0.1; Minor arterials=0.1; Other roads=0.7	Proportion of ***Freeways=0.5***; Principal arterials=0.1; Minor arterials=0.1; ***Other roads=0.3***	-	-	-	-	-6%
Proportion of principal arterial miles	Proportion of Freeways=0.1; Principal arterials=0.1; Minor arterials=0.1; Other roads=0.7	Proportion of Freeways=0.1; ***Principal arterials=0.5***; Minor arterials=0.1; ***Other roads=0.3***	-	-	-	-	2%
Proportion of minor arterial miles	Proportion of Freeways=0.1; Principal arterials=0.1; Minor arterials=0.1; Other roads=0.7	Proportion of Freeways=0.1; Principal arterials=0.1; ***Minor arterials=0.5***; ***Other roads=0.3***	-	-	-	-	2%
Crash exposure factors							
Population density–people/acre (in 10)	0.05	0.25	-	-	-	-	3%
Sociodemographic variables							
Proportion of high-income HH	0.2	0.5	-	-	-	-	-6%
Crime and traffic violations							
Crime rate–total crimes/total population (in 10)	0.1	0.5	-	-	-	-	3%