A New Spatial and Flexible Multivariate Random-Coefficients Model for the Analysis of Pedestrian Injury Counts by Severity Level

Chandra R. Bhat (corresponding author)
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Tel: 1-512-471-4535; Email: bhat@mail.utexas.edu
and
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Sebastian Astroza
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Tel: 1-512-471-4535, Email: sastroza@utexas.edu

Patrícia S. Lavieri
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712, USA
Tel: 1-512-471-4535; Email: laviepa@gmail.com

ABSTRACT

We propose in this paper a spatial random coefficients flexible multivariate count model to examine, at the spatial level of a census tract, the number of pedestrian injuries by injury severity level. Our model, unlike many other macro-level pedestrian injury studies in the literature, explicitly acknowledges that risk factors for different types of pedestrian injuries can be very different, as well as accounts for unobserved heterogeneity in the risk factor effects. We also recognize the multivariate nature of the injury counts by injury severity level within each census tract (as opposed to independently modeling the count of pedestrian injuries by severity level). In concrete methodological terms, our model: (a) allows a full covariance matrix for the random coefficients (constant heterogeneity, or CH, and slope heterogeneity, or SH, effects) characterizing spatial heterogeneity for each count category, (b) addresses excess zeros (or any other excess count value for that matter) within a multivariate count setting in a simple and elegant fashion, while recognizing multivariateness engendered through covariances in both the CH and SH effects, (c) accommodates spatial dependency through a spatial autoregressive lag structure, allowing for varying spatial autoregressive parameters across count categories, and (d) captures spatial drift effects through the spatial structure on the constants and the slope heterogeneity effects. To our knowledge, this is the first time that such a general spatial multivariate model has been formulated. For estimation, we use a composite marginal likelihood (CML) inference approach that is simple to implement and is based on evaluating lower-dimensional marginal probability expressions.

The data for our analysis is drawn from a 2009 pedestrian crash database from the Manhattan region of New York City. Several groups of census tract-based risk factors are considered in the empirical analysis based on earlier research, including (1) socio-demographic characteristics, (2) land-use and road network characteristics, (3) activity intensity characteristics, and (4) commute mode shares and transit supply characteristics. The empirical analysis sheds light on both engineering as well as behavioral countermeasures to reduce the number of pedestrian-vehicle crashes by severity of these crashes.

Keywords: Multivariate count model, spatial dependence, unobserved heterogeneity, composite marginal likelihood estimation, pedestrian injuries in traffic crashes.

1 INTRODUCTION

Walking and bicycling are two active transportation modes that can contribute in important ways to, among other things, lower traffic congestion levels, energy independence, reduced mobile-source emissions, improved public health, and vibrant social cohesion opportunities (see Wier et al., 2009). Indeed, there is increasing recognition among transportation planners, social scientists, urban design specialists, as well as public health professionals that investments in non-motorized facilities, and carefully choreographed educational campaigns to promote walking and bicycling, can be key ingredients of a broader public policy strategy to engender a happier public and a better quality of life (Rasciute et al., 2010).

Between the non-motorized modes of walking and bicycling, the former may be viewed as the most natural form of transportation (at least for most individuals) in that it does not entail any non-human mobility assistance. In fact, almost all individuals are pedestrians for at least a small part of each of their travel journeys. However, the proportion of trips in developed countries that are completely undertaken by foot is a very small fraction of total trips. For example, according to the most recent National Household Travel Survey (NHTS) conducted in 2009 in the United States, trips by the walk mode accounted for only 10.4% of all weekday trips, and 0.74% of total weekday person travel mileage. While there are many reasons for the relative lack of preference to travel by foot (including low land use mix diversity, unconducive built environment factors and weather conditions, and long trip distances), one important reason provided by individuals in surveys as a substantial impediment to the choice of the walk mode of travel (even for short-distance trips) is the perception that it is unsafe from the perspective of traffic crashes (see, for example, Kamargianni et al., 2015 and Weinstein-Agarwal, 2008). Unfortunately, this perception is not unfounded. According to the latest traffic safety data from the National Highway Traffic Safety Administration (NHTSA), in 2015, 5,376 pedestrians lost their lives and another 70,000 pedestrians sustained injuries in traffic crashes in the US (NHTSA, 2016a). That is, on average, a pedestrian was killed every 98 minutes and injured every 7.5 minutes in traffic crashes in the US. More importantly, while the total number of roadway crash fatalities in the US fell from 43,510 in 2005 to 32,675 in 2014 (a 24.9% drop), the total number of pedestrian fatalities remained virtually the same at 4,892 in 2005 and 4,910 in 2014 (NHTSA, 2016b). Further, between 2014 and 2015, while overall fatalities climbed by 7.2% (from 32,744 to 35,092), pedestrian fatalities rose much faster by 9.5% (from 4,910 to 5,376; 5,376 is the highest number of pedestrians killed in road

crashes in any year since 1996). Additionally, the percentage of pedestrian fatalities as a fraction of total fatalities has seen a steady up climb over the years, from 11% in 2005 to 18% in 2014. A similar situation exists in many other developed countries. For example, in Australia, pedestrians comprise 17% of all serious transportation-related injuries and 13% of all road fatalities, according to the Bureau of Infrastructure, Transportation, and Regional Economics (BITRE, 2013). Indeed, pedestrians are often referred to as "vulnerable road users" because of their over-representation in the pool of those fatally injured in traffic crashes. Of course, this is not surprising because, in a crash, pedestrians have little to no protection relative to other road users.

Clearly, efforts to promote walking need to be coordinated with strategies that enhance safety for the vulnerable road-user group of pedestrians. This, in turn, necessitates an understanding of the risk factors associated with pedestrian injuries in the context of traffic crashes, to allow the identification of high risk crash environmental settings and inform the design of appropriate transportation policy countermeasures. In the literature, such analyses have been undertaken through the development of pedestrian crash and injury prediction models. Such models are generally developed at either the micro-level or the macro-level location unit. The micro-level models use a roadway street segment or an intersection as the location unit of analysis, with the aim of identifying relatively shorter-term engineering solutions (such as geometric design improvements or traffic signal control re-configurations). The macro-level models, on the other hand, use a more aggregate "neighborhood" level location unit of analysis with the aim of identifying relatively longer-term planning and behavioral modification solutions (such as more equitably channeling resources for pedestrian facility investments if inequities are identified, or land use design reconfigurations, or targeting specific demographic groups with information campaigns).

In this paper, we contribute to the pedestrian crash literature by formulating a macro-level multivariate model to jointly analyze the count of pedestrians involved in traffic crashes by each of multiple injury severity levels. The reader will note that, for each injury severity level, the count variable used in the analysis corresponds to the number of pedestrian injuries of that injury severity level within a census tract, not the number of crashes within a census tract by the most severe level of injury incurred by a pedestrian in the crash (the latter approach would not appropriately consider situations where multiple non-motorized individuals are injured, and to different levels, in a single

crash).[1] The spatial unit used in our analysis to characterize a "neighborhood" is the census tract, which represents a reasonably homogenous spatial unit of an urban area (see Delmelle et al., 2011). Besides, the census directly provides socio-economic data at the level of the census tract, facilitating analysis at this spatial scale.

The analysis in this paper, unlike many other macro-level pedestrian injury studies in the literature (see, for example, Moudon et al., 2011, Wier et al., 2009 and Cai et al., 2016), explicitly acknowledges the need to model pedestrian injuries by injury severity level. This is because the risk factors for different types of pedestrian injuries can be very different, as already established by Narayanamoorthy et al. (2013) and Amoh-Gyimah et al. (2016). An understanding of these variations is critical to the identification and prioritization of planning, educational, and enforcement safety countermeasure efforts, particularly because the financial and other costs of crashes vary substantially based on the nature and extent of injuries sustained (see Wang et al., 2011 and Blincoe et al., 2015). For example, a tract with four pedestrian fatalities over a given time period should be considered more hazardous than a tract where four pedestrians are injured in a non-incapacitating manner over the same time period. In terms of site ranking for improvement or effective informational campaign strategies, it is important to identify the risk factors of the first tract that make it particularly vulnerable to fatal pedestrian injuries.

Even as analysts need to recognize the differential risk factors for different pedestrian injury severity levels, it is also important to recognize the multivariate nature of the injury counts by injury severity level within each census tract (as opposed to independently modeling the count of pedestrian injuries by severity level; see, for example, the univariate count models by severity level in Amoh-Gyimah et al., 2016). In particular, there may be unobserved census tract factors that (1) intrinsically impact pedestrian injuries in specific ways across injury levels (for example, the absence of sidewalks in a census tract may lead to a general increase in risk propensity for pedestrians across all injury levels), and (2) moderate the effect of an exogenous variable on the risk for different injury levels (for example, the absence of sidewalks may increase the impact of an exposure proxy variable such as population density on the risk for all injury severity levels). For each census tract, the first effect above generates a covariance across the intrinsic risks of

[1] Crash data include information on the individuals who are hurt and the level of injury sustained by each individual (typically in such categories as no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury). At an aggregate level of a census tract, one can then obtain, over a specific time period, the number of pedestrians involved in traffic crashes by injury severity level.

different injury levels (cross injury severity level risk covariance due to unobserved intrinsic tract-specific factors that lead to constant heterogeneity or CH across tracts), while the second effect generates a covariance across the effects of an exogenous variable on different injury levels (cross injury severity level risk covariance due to unobserved tract-specific factors that moderate the effect of an exogenous variable, leading to slope heterogeneity or SH across tracts). Of course, by definition, these effects correspond to unobserved factors, and one can only speculate on what these unobserved factors may be. The important point is that the analyst should acknowledge and test the potential presence of such effects, leading to the need for a multivariate count model system for pedestrian injuries by injury severity level. In the crash literature, multivariateness is almost exclusively accommodated through cross injury risk covariance due to CH (see, for example, Huang et al., 2017); we are not aware of multivariateness generated by cross injury risk covariance due to SH being considered.

Another important issue in the modeling of crashes is to acknowledge unobserved location-based heterogeneity effects (in our case, dependency in the census tract-based spatial heterogeneity effects; see Mannering et al., 2016). This is very closely related to the need for a multivariate system as discussed in the previous paragraph. Indeed, as discussed earlier, we generate multivariateness through the CH and SH effects, which immediately imply unobserved census tract heterogeneity (or spatial heterogeneity) in the risks (a model with both CH and SH effects is generally referred to as a random coefficients or random parameters model). But the multivariate specification by itself does not accommodate, for a given injury severity level, possible covariance between pairs of the CH and SH effects. For instance, it is possible that in some census tracts there is a greater tendency of jaywalking (unobserved factor) and this leads to say an increase in the risk of injuries in the "possible" injury category (positive CH-based effect). Then, in areas close to subway stations this jaywalking tendency may become even more pronounced and increase even more the risk propensity of possible injuries (a positive SH-based effect). In such a case, there would be a positive covariance between the constant effect and the effect of the number of subway stations for the risk of "possible" injuries. In this context of unobserved heterogeneity within multivariate specifications, a couple of relevant studies in the crash literature are Barua et al. (2016) and Anastasopoulos (2016). The model proposed here is more general in that it allows covariance in the intrinsic risk and the effects of variables on risk for each and all injury severity levels. In the two earlier multivariate studies just identified, the covariance matrix across

parameters for each injury severity level is assumed to have off-diagonal elements of zero, as is also the case in almost all other random parameters models in the crash literature, including in the recent univariate models of Xu and Huang (2015) and Amoh Gyimah et al. (2016).[2] Additionally, in a multivariate context, Narayanamoorthy et al. (2013) and Huang et al. (2017) do not accommodate SH effects in the coefficients in their multivariate model, only the CH effects.

The rest of this paper is structured as follows. Section 2 provides an overview of the method adopted in the current study, including a discussion of how spatial dependence is incorporated (spatial dependence is an issue separate from the multivariateness and spatial heterogeneity issues discussed in this section). Section 3 presents the model structure and estimation procedure. Section 4 discusses the empirical application, including data description, empirical estimation results, and implications for reducing pedestrian injury severity in roadway crashes. Finally, Section 5 concludes the paper.

2 THE CURRENT PAPER

In our multivariate analysis, we recast the traditional count model as a special case of a generalized ordered-response (GOR) model in which the count is viewed as a result of a latent risk propensity that gets mapped into the observed count outcomes through thresholds that are themselves functions of exogenous variables (see Castro et al., 2012 and Bhat et al., 2014a). Doing so allows the multivariate linkage across count categories to be easily generated through the latent risk propensity, and excess probability masses (such as excess zero values) are easily handled without the need for zero-inflated and hurdle-count type devices that get very cumbersome in multivariate count settings. As importantly, our approach to recast count models as GOR models also enables us to accommodate spatial dependence effects in a rich manner.

2.1 Spatial Dependency

An important consideration in injury count modeling relates to spatial dependency. Spatial dependency is important to recognize because injury occurrence locations, by nature, are location-based. Thus, it is not difficult to think of reasons why the risk of one injury severity level at one

[2] Two recent studies do accommodate covariance across random parameters in a flexible latent class setting; see Buddhavarapu et al. (2016) and Heydari et al. (2017). The first study is, however, a univariate model, and the second, like Anastasopoulos (2016), does not consider spatial dependence as we do in the current paper.

location (one census tract) will affect the risk of the same injury severity level at another proximal segment. Such a specification implies that both observed as well as unobserved variables (that impact risk for a specific pedestrian injury severity level in crashes) affect the injury count of the same severity level at proximally located segments. For example, it is certainly possible that motorists in census tracts with a substantial fraction of local/residential roads in the roadway system (say an observed variable in the analysis) generally are more attuned to driving in a pedestrian-heavy environment with people of different age groups, and this general motorist experience not only reduces the count of severe pedestrian injuries in the census tract, but also has a "spatial spillover" effect on the count of severe pedestrian injuries at proximally located census tracts as these pedestrian-friendly drivers travel close to their residences and traverse neighboring census tracts. In addition, there may be common unobserved (to the analyst) location factors (such as the absence of continuous pedestrian walkways or pedestrian signals at adjacent census tracts) that may lead to a "spatial correlation" effect in the error terms of the injury risk propensity at proximally located tracts. Ignoring such spatial dependencies will, in general, result in inconsistent and inefficient parameter estimation.

In the multivariate count data analysis literature in general, and the multivariate crash count data analysis literature in particular, the most common approach to introduce spatial dependence is based on using a conditional autoregressive (CAR) (that is, a joint prior on a spatial random effect) term that is introduced multiplicatively in exponential form in the parameterization of the expected value of the discrete distribution for the count variable. The resulting model is typically estimated using Bayesian hierarchical methods (see Heydari et al., 2017 and Buddhavarapu et al., 2016 for recent examples). We also refer the reader to Narayanamoorthy et al. (2013), Mannering and Bhat (2014), and Barua et al. (2015) for a review and details. Unfortunately, the CAR random-effect approach considers only spatial error correlation effects, but completely ignores spatial spillover effects. That is, a change in a variable affecting the dependent count variable will not affect the dependent count variable in a neighboring tract in the CAR approach. In this regard, the CAR approach is akin to the spatial autoregressive error (SAR) structure used commonly in the spatial econometrics literature. As indicated by Beck et al. (2006), McMillen (2010) and Bhat (2015a), the SAR (and by extension, the CAR) structure necessitate the rather illogical position that space matters in the error process but not in the effects of exogenous variables. The implication is that if a new independent variable is added to a spatial

error model "so that we move it from the error to the substantive portion of the model" (Beck et al., 2006), the variable magically ceases to have a spatial impact on neighboring observations. On the other hand, the spatial lag specification, in reduced form (see next section), allows symmetry in spatial dependence through both spatial spillover effects as well as spatial error correlation effects. Overall, we submit that, on pure theoretical and logical grounds, spatial dependency in crash models should be developed using the spatial lag formulation we use here (or its variants), and so we will not empirically test the spatial lag structure with spatial error structures (even though the SAR structure is in fact a little simpler to estimate in our inference approach).

There is yet one other reason to adopt the spatial lag structure, when combined with unobserved (spatial) heterogeneity in the effects of exogenous variables (that is, random coefficients), as discussed in detail by Bhat (2015a). Specifically, because of the spatial nature of injury occurrence locations, "spatial drift" effects are very likely wherein SH effects themselves should be correlated over tracts based on spatial proximity (see Bradlow et al., 2005 and Bhat, 2015a for a discussion of the spatial drift phenomena). Thus, for example, consider an unobserved tract variable that corresponds to motorist friendliness levels toward pedestrians (MFTP). It is certainly plausible that there is a proximity-based spatial pattern (across tracts) in this underlying MFTP attitude because of social interactions. If this MFTP attitude reduces pedestrian crash risk particularly on local residential roads (because of the high motorist-pedestrian interactions on such roads), then the extent of deviation (from the norm) in the effect of a "proportion of local residential roads" variable in a tract on pedestrian injury risk would once again be correlated with the corresponding deviation in other tracts based on spatial proximity. We accommodate such correlations because we allow unobserved SH effects, which when combined with the spatial lag structure, imply spatial drift effects (on the other hand, SH effects, when combined with a SAR or CAR structure, do not engender such drift effects because the SAR and CAR structures act solely upon the CH effect). This is one other strong reason to prefer the spatial lag structure over other spatial structures.[3]

[3]As indicated by Bhat (2015a), in the spatial literature, the so-called "spatial drift" effects have typically been incorporated using the geographically weighted regression (GWR) approach of Brunsdon et al. (1998). For recent applications of GWR in the crash literature, see Xu and Huang (2015) and Amoh-Gyimah et al. (2017). The GWR approach, however, is relatively exploratory in nature compared to our approach. Besides, the GWR approach essentially allows spatial variations in the regression coefficients based on observed exogenous variables, by allowing variable coefficients at a point to be a function of exogenous attributes at that point and exogenous attributes at

2.2 Summary Overview of Paper

The spatial random coefficients flexible multivariate count model proposed in this paper recognizes many econometric issues at once: (a) It allows a full covariance matrix for the random coefficients (CH and SH effects) characterizing spatial heterogeneity for each count category, (b) It addresses excess zeros (or any other excess count value for that matter) within a multivariate count setting in a simple and elegant fashion, while recognizing multivariateness engendered through covariances in both the CH and SH effects, (c) It accommodates spatial dependency through a spatial autoregressive lag structure, allowing for varying spatial autoregressive parameters across count categories, and (d) It captures spatial drift effects through the spatial structure on the CH and SH effects. To our knowledge, this is the first time in the crash literature, as well as the broader econometric literature, that such a general spatial multivariate model has been formulated. The likelihood function for the resulting model is analytically intractable, and simulation approaches are of little use. To overcome this issue, we use a composite marginal likelihood (CML) inference approach that is simple to implement and is based on evaluating lower-dimensional marginal probability expressions.

The proposed model is applied to examine, at the spatial level of a census tract, the number of pedestrian injuries by injury severity level. In this empirical context, an appropriate exposure measure of crash risk within a census tract would be the number of pedestrian miles of travel and motorized vehicle miles of travel. But, because of the difficulty in constructing such measures accurately, we use surrogate exposure measures such as population density, income, land-use,

---

proximally located points. That is, the GWR engenders spatial heterogeneity due to varying exogenous attributes over space, while also capturing spatial dependence through the recognition of exogenous attributes within a certain proximal space. But it fundamentally and completely ignores the presence of unobserved location attributes that impact the effects of exogenous variables, and also ignores the spatial dependence in this unobserved location heterogeneity. Our approach is conceptually more general in that it allows spatial heterogeneity, as well as spatial dependence, in the effects of both observed as well as unobserved factors. Additionally, even within the context of observed variable effects, our approach enables the direct and immediate disentangling of the effect of a variable at a point in space (direct parameter effect) from the effects of the corresponding variable values at proximal points in spaces (spatial spillover or indirect effects), while the GWR essentially combines the two into a single effect. One final issue regarding terminology, because the term "spatial drift" is used in different ways in the literature. We will specifically use the term "spatial drift" in the rest of this paper to refer to the spatial dependence pattern among unobserved effects of variables, while reserving the term "spatial spillover" to refer to the spatial dependence caused by the effects of observed variables characterizing proximally located spatial units. In this terminology, the GWR accommodates "spatial spillover" (even though it combines this effect with the direct parameter effect), while the GWR actually ignores the "spatial drift" effect. The CAR and SAR structures ignore both the spatial spillover as well as spatial drift effects.

road-network characteristics, and activity intensity characteristics. As discussed in several earlier studies (see, for example, Huang et al., 2014; Aguero-Valverde et al., 2006), this approach has the advantage that exposure is internalized, and so it is possible to identify census tracts that are likely to have a high number of pedestrian injuries based purely on the readily available census tract demographic factors and built environment characteristics. The data for our analysis is drawn from a 2009 pedestrian crash database from the Manhattan region of New York City. Several groups of census tract-based risk factors are considered in the empirical analysis based on earlier research, including (1) socio-demographic characteristics (such as population density, proportions of the population by age, income, and race/ethnicity), (2) land-use and road network characteristics (such as proportion of retail and commercial land-use, and proportion of roads by functional type), (3) activity intensity characteristics (such as retail intensity, and number of schools and universities), and (4) commute mode shares and transit supply characteristics (such as shares of commute trips by mode and number of bus stops).

3 METHODOLOGY

3.1 Count Model Recasting as a Generalized Ordered-Response Model

Let q (q = 1, 2, …, Q) be the index for census tracts and let j (j = 1, 2, …, J) be the index for injury severity, where Q is the total number of observation units (census tracts) in the sample, and J is the number of injury severity levels (J=4 in our later empirical analysis, corresponding to the pedestrian injury severity categories of "possible" injury (j=1), "non-incapacitating injury" (j=2), "incapacitating injury" (j=3), and "fatal" injury (j=4)). Let y_{qj} be the index for the count of injury severity j at the census tract q, and let m_{qj} be the actual observed count of the injury severity j at the census tract q over a predefined time period (we considered a time period of one year for the empirical analysis in this paper; note also that m_{qj} may take a value in the range from 0 to ∞). Next, define a latent risk propensity for the injury severity j at tract q as y_{qj}^{*}. Then, consider the following structure for y_{qj}^{*} in the GOR representation for count models (see Castro, Paleti, and Bhat (CPB), 2013):

$$y_{qj}^{*} = \delta_j \sum_{q'=1}^{Q} w_{qq'} y_{q'j}^{*} + \boldsymbol{\beta}_{qj}' \tilde{\boldsymbol{x}}_q \qquad y_{qj} = m_{qj} \text{ if } \psi_{qj,m_{qj}-1} < y_{qj}^{*} < \psi_{qj,m_{qj}}, \tag{1}$$

where $w_{qq'}$ is the usual distance-based spatial weight corresponding to tracts q and q' (with $w_{qq}=0$ and $\sum_{q'} w_{qq'}=1$) for each (and all) q, δ_j $(0<\delta_j<1)$ is the spatial autoregressive parameter for injury severity level j, $\widetilde{\boldsymbol{x}}_q$ is a $(K\times 1)$ column vector of exogenous variables (including a constant; $\widetilde{\boldsymbol{x}}_q=(1,\widetilde{x}_{q2},\widetilde{x}_{q3},\ldots,\widetilde{x}_{qK})'$)), and $\boldsymbol{\beta}_{qj}$ is a corresponding $(K\times 1)$ column vector capturing the effects of the exogenous vector $\widetilde{\boldsymbol{x}}_q$ on the latent risk propensity y^*_{qj}: $\boldsymbol{\beta}_{qj}=(\beta_{qj1},\beta_{qj2},\beta_{qj3},...,\beta_{qjK})'$.[4]

The thresholds in Equation (1) take the following form:

$$\psi_{qj,y_{qj}}=\Phi^{-1}\left(e^{-\lambda_{qj}}\sum_{l=0}^{y_{qj}}\frac{\lambda_{qj}^{l}}{l!}\right)+\alpha_{j,y_{qj}},\ \lambda_{qj}=e^{\gamma_j z_q},\alpha_{j,0}=0\ \forall\ j,\ \alpha_{j,y_{qj}}=\alpha_{j,L_j}\ \text{if}\ y_{qj}>L_j, \tag{2}$$

where Φ^{-1} is the inverse function of the univariate cumulative standard normal, $\psi_{qj,-1}=-\infty\,\forall\, q$ and j (this restriction is needed for identification, given the parameterization of the thresholds; see CPB, 2012), $\boldsymbol{z}_q$ is a vector of exogenous variables (including a constant) associated with observation unit q (there can be common variables in $\boldsymbol{z}_q$ and $\widetilde{\boldsymbol{x}}_q$), $\boldsymbol{\gamma}_j$ is a corresponding coefficient vector to be estimated for injury severity j, and L_j is an appropriate count level that may be determined based on the empirical context under consideration and empirical testing. Of course, as in the typical ordered-response framework, the values of $\alpha_{j,y_{qj}}$ should be such that the ordering condition on the thresholds $(-\infty<\psi_{qj,0}<\psi_{qj,1}<\psi_{qj,2}<....)$ is satisfied. While this can be guaranteed using a reparameterization (of the type suggested in Greene and Hensher, 2010, page 109 and Eluru et al., 2008), the ascending nature of the first component of the threshold and its size relative to the $\alpha_{j,y_{qj}}$ values guaranteed the ordering conditions on the overall threshold values. This is a result we have also observed in several other applications of our recasting of the count model (similar to the lack of a need to explicitly constrain the thresholds in a simple ordered-response model). At the same time, the presence of these $\alpha_{j,y_{qj}}$ terms provide flexibility to accommodate high or low probability masses for specific count outcomes without the

[4] Some explanatory variables may not be important for a specific injury type j. This situation is accommodated within our notation system by letting the corresponding elements in the vector $\boldsymbol{\beta}_{qj}$ be equal to zero.

need for using hurdle or zero-inflated mechanisms that can become cumbersome when dealing with multivariate counts. The reader can note that, for each injury severity j, if the $\alpha_{j,y_{qj}}$ terms are all identically set to zero, all elements of the vector $\boldsymbol{\beta}_{qj}$ except the one on the constant are also set to a fixed value of zero (note that this structure implies no unobserved heterogeneity in coefficients), and the constant parameter element in $\boldsymbol{\beta}_{qj}$ is replaced with a standard normally distributed error term (that is, $\beta_{qj1} \sim N(0,1)$), the result is the traditional Poisson count model for each crash type (see CPB, 2012).[5]

The framework above provides useful computational benefits to accommodate statistical, econometric, and spatial considerations, while also having an intuitive interpretation (see CPB, 2012). For example, in our empirical context, consider the count of incapacitating pedestrian injuries (the following discussion applies to all severity levels, and we pick the category of incapacitating injuries simply for illustration). The interpretation of the GOR framework is that there is a latent "long-term" (and constant over a certain time period) risk y^*_{q3} of incapacitating injuries at census tract q, which is influenced by such tract-specific variables as, say, intensity of retail activity, and commercial and residential land-use (due to higher pedestrian activity and exposure in and around areas with high levels of these developments relative to open areas). These

[5] We also attempted to estimate models that used a threshold structure in Equation (2) that adds a parameter that serves the same purpose as the dispersion parameter (in a traditional negative binomial count or NBC model) in the spirit of capturing additional overall unobserved heterogeneity. This more general structure, when combined with the many assumptions just stated (including imposing the restriction of no unobserved heterogeneity in the $\boldsymbol{\beta}_{qj}$ coefficients on exogenous variables), would collapse to a traditional NBC model for each crash type (see Bhat, 2015b and Bhat et al., 2014b, 2016 for such formulations for the threshold). However, in our empirical results, this additional parameter in the threshold became very large for each (and all) crash types and resulted in estimation instability as soon as we incorporated unobserved heterogeneity in the effects of exogenous variables through the $\boldsymbol{\beta}_{qj}$ coefficients (when the parameter tends to infinity in our formulation, the threshold specification collapses to that used in Equation (2); in the more simple traditional context, this would be the same as the NBC model collapsing to the Poisson model). The implication is that our random coefficients specification for $\boldsymbol{\beta}_{qj}$ already captures unobserved heterogeneity that otherwise would manifest itself in the additional parameter in the generalized variant of Equation (2). Thus, we have chosen to present the model structure with the simpler notation of Equation (2) to streamline the presentation and focus on other substantial methodological enhancements. However, this should not detract from the flexibility of the model presented here, which can potentially include another dispersion-related term in Equation (2). At the same time, our results here also suggest that the traditional negative binomial model specification used in many empirical contexts may actually be a mis-specification because it may simply be capturing the ignored heterogeneity in the effects of exogenous variables and thrusting all these relevant sensitivity variations to exogenous variables into a single composite "misleading" heterogeneity term. Also to be noted is that, in addition to associating unobserved heterogeneity to individual exogenous variables, our model framework has a separate mechanism (through the $\alpha_{j,y_{qj}}$ parameters) to accommodate "spikes" at specific count values.

variables would then get manifested in the $\widetilde{\boldsymbol{x}}_q$ vector. On the other hand, there may be some specific census tract characteristics (embedded in $\boldsymbol{z}_q$) that may dictate the likelihood of a crash occurring at any given *instant of time* for a given long-term crash risk y^*_{q3}. For instance, a high proportion of commercial or residential land-use in a tract may lead to higher levels of distraction and/or pre-occupation among drivers around these land-uses (relative to around open and recreational land-uses). In this situation, the effect of the high proportion of commercial or residential land-use is to increase the "instantaneous" likelihood of a crash resulting in a pedestrian being incapacitated. This risk-to-outcome translation effect (which we will also refer to as the "threshold" effect) is relatively localized, and separate and different from the effects that these same variables may have to increase the long-term risk propensity of pedestrian injuries. Further, the GOR framework in Equation (1) accommodates spatial dependency in counts through spatial lag ("spillover") effects in the "long-term" latent crash propensity, not through the elements that affect the localized and "instantaneous" translation of the propensity to whether or not a crash occurs at any given time (and, therefore, not the threshold elements that affect the mapping of the latent risk to the observed crash count outcome).

3.2 Model Formulation and Estimation

To proceed with the model formulation, we assume that the vector $\boldsymbol{\beta}_{qj}$ is a realization from a multivariate normally distributed vector for each injury severity level *j*. That is, $\boldsymbol{\beta}_{qj} \sim MVN_K(\boldsymbol{b}_j, \boldsymbol{\Omega}_j)$. For each injury level *j*, the first element of $\boldsymbol{b}_j$ is set to zero and the first diagonal element of $\boldsymbol{\Omega}_j$ is normalized to one (these aid in econometric stability and also reflect the lack of strict cardinality in the underlying latent propensity; see CPB, 2012). For future reference, let $\overline{\boldsymbol{\Omega}}_j$ be a column vector obtained by stacking the unique elements of the symmetric matrix $\boldsymbol{\Omega}_j$. By allowing a multivariate distribution for $\boldsymbol{\beta}_{qj}$, we allow for a full covariance matrix for the random coefficients for each count category. It is now well established that ignoring such variations when present will lead to inconsistent and biased parameters estimates in count models (see Bhat et al., 2014a; Mannering et al., 2016). Of course, the discussion thus far has only focused on heterogeneity in intrinsic risk and the effect of exogenous variables across census tracts for each injury level *j*. However, the multivariateness across injury levels also must be recognized. In

our model, we accommodate this multivariateness not only through common unobserved factors that may impact overall intrinsic risks across injury levels, but also through unobserved factors simultaneously moderating the effects of a variable *k* across risks of different injury levels. For this, we assume that $\overline{\boldsymbol{\beta}}_{qk} = (\boldsymbol{\beta}_{q1k}, \boldsymbol{\beta}_{q2k}, \boldsymbol{\beta}_{q3k}, ..., \boldsymbol{\beta}_{qJk})' \sim MVN_J(\mathbf{0}_J, \boldsymbol{\Xi}_k)$. Of course, because the heterogeneity across census tracts in the overall intrinsic risk as well as the impacts of exogenous variables is already considered in $\boldsymbol{\Omega}_j$, all the diagonal elements of $\boldsymbol{\Xi}_k$ for each *k* have to be normalized to zero. This is admittedly unconventional notation in that there cannot be a covariance matrix with zero diagonal entries and non-zero non-diagonal entries because the resulting matrix is not positive definite; but the reader will note that we do not directly estimate $\boldsymbol{\Xi}_k$, but include it later as an element of a larger covariance matrix to be estimated. For notational purposes, it is helpful to structure things as we have. For future reference, let $\overline{\boldsymbol{\Xi}}_k$ represent a column vector obtained by stacking the unique cross-diagonal elements of the symmetric matrix $\boldsymbol{\Xi}_k$.

To write Equation (5) compactly, we define several vectors and matrices. Let $\boldsymbol{y}_q^* = (y_{q1}^*, y_{q2}^*, ..., y_{qJ}^*)'$, $\boldsymbol{m}_q = (m_{q1}, m_{q2}, ..., m_{qJ})'$, $\boldsymbol{y}_q = (y_{q1}, y_{q2}, ..., y_{qJ})'$, $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_J)'$ ($J\times1$) vectors, $\boldsymbol{y}^* = [(\boldsymbol{y}_1^*)', (\boldsymbol{y}_2^*)', ..., (\boldsymbol{y}_Q^*)']'$, $\boldsymbol{m} = [(\boldsymbol{m}_1)', (\boldsymbol{m}_2)', ..., (\boldsymbol{m}_Q)']'$, $\boldsymbol{y} = [(\boldsymbol{y}_1)', (\boldsymbol{y}_2)', (\boldsymbol{y}_3)', ..., (\boldsymbol{y}_Q)']'$, $\widetilde{\boldsymbol{\delta}} = \mathbf{1}_Q \otimes \boldsymbol{\delta}$ ($QJ\times1$ vectors; $\mathbf{1}_Q$ is a column vector of size *Q* with all elements equal to 1, and '$\otimes$' is the Kronecker product), $\boldsymbol{x}_q = \mathbf{IDEN}_J \otimes \widetilde{\boldsymbol{x}}_q$ ($J\times JK$ matrix) ($\mathbf{IDEN}_J$ is a square identity matrix of size *J*), $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2', ..., \boldsymbol{x}_Q')'$ ($QJ\times JK$ matrix), $\boldsymbol{\beta}_q = (\boldsymbol{\beta}_{q1}', \boldsymbol{\beta}_{q2}', ..., \boldsymbol{\beta}_{qJ}')'$ ($JK\times1$ vector), and $\boldsymbol{b} = (\boldsymbol{b}_1', \boldsymbol{b}_2', ..., \boldsymbol{b}_J')'$ ($JK\times1$ vector). Next, write $\boldsymbol{\beta}_q = \boldsymbol{b} + \breve{\boldsymbol{\beta}}_q$. Based on the covariance patterns assumed across the coefficient elements, we may write $\breve{\boldsymbol{\beta}}_q \sim MVN_{JK}(\boldsymbol{b}, \boldsymbol{\Delta})$, where $\boldsymbol{\Delta}$ is a $JK\times JK$ covariance matrix that may be written as follows:

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Omega}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Omega}_J \end{bmatrix} + \sum_{k=1}^{K} \left(\boldsymbol{\Xi}_k \otimes \begin{bmatrix} 1(k=1) & 0 & 0 & 0 \\ 0 & 1(k=2) & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1(k=K) \end{bmatrix} \right), \tag{3}$$

where the notation $1(k=1)$ takes a value of 1 if $k=1$ and 0 otherwise, and similarly for all other values of k. Also, let $\breve{\boldsymbol{\beta}} = (\breve{\boldsymbol{\beta}}_1', \breve{\boldsymbol{\beta}}_2', ..., \breve{\boldsymbol{\beta}}_Q')'$ ($QJK \times 1$ vector), where $\breve{\boldsymbol{\beta}} = MVN_{QJK}(\mathbf{0}_{QJK}, \boldsymbol{IDEN}_Q \otimes \boldsymbol{\Delta})$. Define the following matrix:

$$\breve{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{x}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{x}_Q \end{bmatrix} \quad (QJ \times QJK \text{ matrix}) \tag{4}$$

Collect all the weights $w_{qq'}$ into a $Q \times Q$ row-normalized spatial weight matrix $\boldsymbol{W}$. Let $\breve{\boldsymbol{\delta}} = \widetilde{\boldsymbol{\delta}}.*(\boldsymbol{W} \otimes \boldsymbol{IDEN}_J)$ ($QJ \times QJ$ matrix), where the operation $\boldsymbol{H} = \boldsymbol{M}.*\boldsymbol{N}$ is used to refer to the element by element product of a vector $\boldsymbol{M}$ and a matrix $\boldsymbol{N}$, *i.e.*, $\boldsymbol{H}_{i,j} = \boldsymbol{M}_{i,j} * \boldsymbol{N}_{i,j}$. Define $\mathbf{C} = \left[\boldsymbol{IDEN}_{QJ} - \breve{\boldsymbol{\delta}}\right]^{-1}$ ($QJ \times QJ$ matrix). With all these definitions, Equation (1) may be re-written in a reduced and compact form as:

$$\boldsymbol{y}^* = \mathbf{C}(\boldsymbol{xb} + \breve{\boldsymbol{x}}\breve{\boldsymbol{\beta}}). \tag{5}$$

In the above compact notation, the spatial spillover effects originate from the **C** matrix being applied to the first $\mathbf{C}\boldsymbol{xb}$ term that includes the observed variable vector, and the spatial error correlation effects and the spatial drift effects originate from the **C** matrix being applied to the second $\breve{\boldsymbol{x}}\breve{\boldsymbol{\beta}}$ term (note that the $\widetilde{\boldsymbol{x}}_q$ vector embedded in $\breve{\boldsymbol{x}}$ includes a constant, giving rise to spatial correlation effects associated with the CH effect; the non-constant variables in the $\widetilde{\boldsymbol{x}}_q$ vector embedded in $\breve{\boldsymbol{x}}$ then lead to the spatial drift effects associated with the SH effect).[6]

The expected value and variance of $\boldsymbol{y}^*$ may be obtained from the above equation as $\boldsymbol{y}^* \sim MVN_{QJ}(\boldsymbol{B}, \boldsymbol{\Psi})$, where $\boldsymbol{B} = \mathbf{C}\boldsymbol{xb}$ and $\boldsymbol{\Psi} = \mathbf{C}\breve{\boldsymbol{x}}\left[\boldsymbol{IDEN}_Q \otimes \boldsymbol{\Delta}\right]\breve{\boldsymbol{x}}'\mathbf{C}'$. The parameter vector to be estimated in the model is $\boldsymbol{\theta} = (\boldsymbol{b}', \boldsymbol{\delta}', \boldsymbol{\gamma}', \boldsymbol{\alpha}', \overline{\boldsymbol{\Omega}}', \overline{\boldsymbol{\Xi}}')'$, where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2', ..., \boldsymbol{\gamma}_J')'$, $\boldsymbol{\alpha}$ is a column

[6] The traditional SAR model (and an analog to the CAR model) is obtained when the **C** matrix is applied only to the second term, as in $\boldsymbol{y}^* = \boldsymbol{xb} + \mathbf{C}\breve{\boldsymbol{x}}\breve{\boldsymbol{\beta}}$, combined with the $\breve{\boldsymbol{x}}$ matrix including only the constant term in each $\widetilde{\boldsymbol{x}}_q$ vector. The traditional random parameters model accommodating spatial heterogeneity is obtained when there is no spatial dependence assumed, as in $\boldsymbol{y}^* = \boldsymbol{xb} + \breve{\boldsymbol{x}}\breve{\boldsymbol{\beta}}$. An analog to the traditional GWR model is obtained when the **C** matrix is applied only to the first term, as in $\boldsymbol{y}^* = \mathbf{C}\boldsymbol{xb} + \breve{\boldsymbol{x}}\breve{\boldsymbol{\beta}}$, combined with the $\breve{\boldsymbol{x}}$ matrix including only the constant term in each $\widetilde{\boldsymbol{x}}_q$ vector.

vector obtained by vertically stacking the $\alpha_{j,y_{qj}}$ $(j=1,2,...,J; y_{qj}=1,2,...,L_j)$ parameters across all injury levels, $\overline{\boldsymbol{\Omega}}=(\overline{\boldsymbol{\Omega}}_1',\overline{\boldsymbol{\Omega}}_2',...,\overline{\boldsymbol{\Omega}}_J')'$, and $\overline{\boldsymbol{\Xi}}=(\overline{\boldsymbol{\Xi}}_1',\overline{\boldsymbol{\Xi}}_2',...,\overline{\boldsymbol{\Xi}}_K')'$. The likelihood function for the model is:

$$L(\boldsymbol{\theta})=P(\boldsymbol{y}=\boldsymbol{m})=\int_{D_{y^*}}\phi_{QJ}(\boldsymbol{y}^*\mid\boldsymbol{B},\boldsymbol{\Psi})d\boldsymbol{y}^*, \tag{6}$$

where $D_{y^*}=\{\boldsymbol{y}^*:\psi_{(qj,m_{qj}-1)}<y^*_{qj}<\psi_{qj,m_{qj}},\ \forall\, q=1,2,...,Q,\ j=1,2,...,J\}$ and $\phi_{QJ}(.)$ is the multivariate normal density function of dimension *QJ*. The integration domain D_{y^*} is simply the multivariate region of the elements of the $\boldsymbol{y}^*$ vector determined by the observed vector of count outcomes. The dimensionality of the rectangular integral in the likelihood function is *QJ*. Existing estimation methods including the Maximum Simulated Likelihood (MSL) method and the Bayesian Inference method become cumbersome and encounter convergence problems even for moderately sized *Q* and *J* (Bhat et al., 2010). The alternative is to use the composite marginal likelihood (CML) approach. In the current study, we use the pairwise composite marginal likelihood method based on the product of the likelihood contributions from pairs of counties across all sectors. To write this function, define threshold vectors as follows:

$$\varphi_q=\left(\psi_{q1,m_{q1}-1},\psi_{q2,m_{q2}-1},...,\psi_{qJ,m_{qJ}-1}\right)' \text{ and } \vartheta_q=\left(\psi_{q1,m_{q1}},\psi_{q2,m_{q2}},...,\psi_{qJ,m_{qJ}}\right)' \ (J\times 1 \text{ vectors}) \tag{7}$$

$$\varphi=\left(\varphi_1',\varphi_2',...,\varphi_Q'\right)' \text{ and } \vartheta=\left(\vartheta_1',\vartheta_2',...,\vartheta_Q'\right)' \ (QJ\times 1 \text{ vectors})$$

Let *g* be an index that can takes the values from 1 to *QJ*. Then,

$$\begin{aligned} L_{CML}(\boldsymbol{\theta}) &= \left(\prod_{g=1}^{QJ-1}\prod_{g'=g+1}^{QJ}P\left([\boldsymbol{y}]_g=[\boldsymbol{m}]_g,[\boldsymbol{y}]_{g'}=[\boldsymbol{m}]_{g'}\right)\right) \\ &= \left(\prod_{g=1}^{QJ-1}\prod_{g'=g+1}^{QJ}\begin{bmatrix}\Phi_2(\widetilde{\varphi}_g,\widetilde{\varphi}_{g'},\boldsymbol{v}_{gg'})-\Phi_2(\widetilde{\varphi}_g,\widetilde{\vartheta}_{g'},\boldsymbol{v}_{gg'})\\ -\Phi_2(\widetilde{\vartheta}_g,\widetilde{\varphi}_{g'},\boldsymbol{v}_{gg'})+\Phi_2(\widetilde{\vartheta}_g,\widetilde{\vartheta}_{g'},\boldsymbol{v}_{gg'})\end{bmatrix}\right), \end{aligned} \tag{8}$$

where $\widetilde{\varphi}_g=\dfrac{[\varphi]_g-[\boldsymbol{B}]_g}{\sqrt{[\boldsymbol{\Psi}]_{gg}}}$, $\widetilde{\vartheta}_g=\dfrac{[\vartheta]_g-[\boldsymbol{B}]_g}{\sqrt{[\boldsymbol{\Psi}]_{gg}}}$, $\boldsymbol{v}_{gg'}=\dfrac{[\boldsymbol{\Psi}]_{gg'}}{\sqrt{[\boldsymbol{\Psi}]_{gg}}\sqrt{[\boldsymbol{\Psi}]_{g'g'}}}$.

In the above expression, $[\vartheta]_g$ represents the g^{th} element of the column vector ϑ, and similarly for other vectors. $[\boldsymbol{\Psi}]_{gg'}$ represents the gg'^{th} element of the matrix $\boldsymbol{\Psi}$. The CML estimator is obtained by maximizing the logarithm of the function in Equation (8). Under usual regularity

assumptions, the CML estimator of $\boldsymbol{\theta}$ is consistent and asymptotically normal distributed with asymptotic mean $\boldsymbol{\theta}$ and covariance matrix given by the inverse of Godambe's (1960) sandwich information matrix (see Zhao and Joe, 2005, Bhat, 2014):

$$\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}}) = [\mathbf{G}(\boldsymbol{\theta})]^{-1} = [\mathbf{H}(\boldsymbol{\theta})]^{-1}\mathbf{J}(\boldsymbol{\theta})[\mathbf{H}(\boldsymbol{\theta})]^{-1},$$

$$\text{where } \mathbf{H}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 logL_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \text{ and } \mathbf{J}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial logL_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial logL_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\right]. \quad (9)$$

The reader is referred to Bhat (2014) for complete details regarding the estimation of the matrices $\mathbf{H}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ in Equation (9) above. To ensure the constraints on the autoregressive terms δ_j (j=1,2,…,J), we parameterize each of these terms as $\delta_j = 1/[1+\exp(\ddot{\delta}_j)]$. Once estimated, the $\ddot{\delta}_j$ estimate can be translated back to an estimate of δ_j.

One final important estimation issue is that the positive definiteness of the covariance matrix $\boldsymbol{\Delta}$ must be ensured. In our estimation, this is guaranteed by writing the logarithm of the pairwise-likelihood in terms of the Cholesky-decomposed elements of $\boldsymbol{\Delta}$ and maximizing with respect to these elements of the Cholesky factor. Essentially, this procedure entails passing the Cholesky elements as parameters to the optimization routine, constructing the $\boldsymbol{\Delta}$ matrix internal to the optimization routine, then computing $\boldsymbol{\Psi}$, and finally selecting the appropriate elements of the matrix for the pairwise likelihood components.

3.3 Model Comparisons

For the purpose of comparing two nested models estimated using the CML approach, the analyst can use the adjusted composite likelihood ratio test (*ADCLRT*) statistic, which is asymptotically chi-squared distributed similar to the likelihood ratio test statistic for the maximum likelihood approach. The reader is referred to Bhat (2011) and Bhat (2014) for details regarding the *ADCLRT* test statistic. In the case of the model proposed in this paper, a host of restrictive versions may be identified. While we have estimated many of these restricted versions, we will confine our presentation later in this paper to a comparison of our general model with five restricted versions that all retain unobserved heterogeneity (random coefficients). Also, when spatial dependence is considered, we will strictly retain the spatial lag structure because of the theoretical and logical

reasons (see Section 2.1) that we believe necessitate its use rather than other spatial dependence structures considered in the literature.

(1) *Independent aspatial count (IAC) model* – A set of four independent models (one for each injury severity) with all the elements of $\overline{\overline{\boldsymbol{\Xi}}}_k$ for each *k* set to zero, and δ_j for each severity category *j* set to zero. The parameter vector estimated here is $\boldsymbol{\theta} = (\boldsymbol{b}', \boldsymbol{\gamma}', \boldsymbol{\alpha}', \overline{\boldsymbol{\Omega}}')'$.

(2) *Independent constrained spatial count (ICSC) model* – A set of four independent models (one for each injury severity level) with all the elements of $\overline{\overline{\boldsymbol{\Xi}}}_k$ for each *k* set to zero, and δ_j for each severity category *j* constrained to be equal to a common parameter $\ddot{\delta}$. The parameter vector estimated here is $\boldsymbol{\theta} = (\boldsymbol{b}', \ddot{\delta}, \boldsymbol{\gamma}', \boldsymbol{\alpha}', \overline{\boldsymbol{\Omega}}')'$.

(3) *Independent unconstrained spatial count (IUSC) model* – A set of four independent models (one for each injury severity level) with all the elements of $\overline{\overline{\boldsymbol{\Xi}}}_k$ for each *k* set to zero, and δ_j for each severity category *j* estimated separately. The parameter vector estimated here is $\boldsymbol{\theta} = (\boldsymbol{b}', \boldsymbol{\delta}, \boldsymbol{\gamma}', \boldsymbol{\alpha}', \overline{\boldsymbol{\Omega}}')'$.

(4) *Multivariate aspatial count (MAC) model* – A joint count model for all severity categories together, but with δ_j for each severity category *j* set to zero. The parameter vector estimated here is $\boldsymbol{\theta} = (\boldsymbol{b}', \boldsymbol{\gamma}', \boldsymbol{\alpha}', \overline{\boldsymbol{\Omega}}', \overline{\overline{\boldsymbol{\Xi}}}')'$,

(5) *Multivariate constrained spatial count (MCSC) model* – A joint count model for all severity categories together, with the δ_j terms for each severity category *j* constrained to be equal to a common parameter $\ddot{\delta}$. The parameter vector estimated here is $\boldsymbol{\theta} = (\boldsymbol{b}', \ddot{\delta}, \boldsymbol{\gamma}', \boldsymbol{\alpha}', \overline{\boldsymbol{\Omega}}', \overline{\overline{\boldsymbol{\Xi}}}')'$.

In the terminology above, the proposed model is the multivariate unconstrained spatial count (MUSC) model.

4 EMPIRICAL APPLICATION

4.1 Data and Sample Formation

The crash data used in this paper corresponds to the pedestrian data component of the non-motorized data used in Narayanamoorthy et al. (2013). For completeness, we provide an overview of the data and sample formation procedures here, though readers may obtain a more detailed description in Narayanamoorthy et al. (2013). The crash data, compiled based on reports from multiple reporting agencies (including the New York Police Department and the New York State

Department of Motor Vehicles), is obtained from the CrashStat website, which is the result of a project undertaken by the New York City's (NYC) Transportation Alternatives organization. All pedestrian crashes in the year 2009 in the Manhattan area of New York City are extracted from this crash data. These crashes are geocoded, with an identification of each pedestrian injured and the severity level of the injury (classified into four levels; possible injury, non-incapacitating injury, incapacitating injury, and fatal injury[7]). The counts of pedestrians per crash by severity level are next aggregated up to the census tract level, to obtain the count of pedestrians injured by severity level in each of 285 census tracts of Manhattan. In addition to the CrashStat data, we used other geo-referenced data sources to obtain census tract-level (the spatial unit of analysis in the current paper) information on (a) socio-demographics, (b) land-use and road network, (c) activity intensity, and (d) commute mode shares and transit supply variables. These formed the independent variables in the analysis.[8]

4.2 Sample Descriptives of the Dependent Variable

Across all census tracts, the sample included a total of 2512 injured pedestrians split by injury severity level as follows: possible injury (1700 or 67.7%), non-incapacitating injury (523 or 20.8%), incapacitating injury (250 or 10.0%), and fatal injury (39 or 1.5%). More important for the analysis in this paper, however, is to examine the sample distributions of pedestrian injuries by census tract. The total number of pedestrians injured (across all injury severity levels) in traffic crashes during the year 2009 per census tract in Manhattan varied between 0 and 40, with an average of about 8.8 injuries per census tract. The corresponding ranges and average for the count of pedestrians injured in traffic crashes per census tract by injury severity level was as follows: possible injury (0 to 25, mean of 5.96), non-incapacitating injury (0 to 13, mean of 1.84), incapacitating injury (0 to 6, mean of 0.88), and fatal injury (0 to 4, mean of 0.14). Figure 1 presents the percentage of census tracts associated with each count of pedestrian injuries by injury severity level, though the figure aggregates counts of 10 or more into a single 10+ category to keep things

[7] An injury from a crash that results in death within 30 days of the crash is labeled as a "fatal" injury.

[8] A note here regarding the empirical analysis. As in almost all earlier pedestrian crash studies, especially those undertaken at the macro-level, we have not included pedestrian infrastructure variables in our analysis because of lack of suitable and easily available data. Thus, the empirical results and policy implications discussed in this paper need to be viewed and scrutinized with some caution, even though we have attempted to develop many proxy variables for pedestrian infrastructure quality, as just listed and discussed more later.

manageable from a presentation standpoint (note that there is no such aggregation in the actual count modeling, and so the spike of the count at 10+ for the "possible injury" category in the figure is artificial). The figure points to an increasingly higher percentage of census tracts with zero counts as one goes from the lowest injury severity level to the highest. In particular, the percentage of tracts with zero values steadily increases from about 9% for pedestrians with possible injury to 90% for pedestrians with fatal injury. For the possible injury severity level in particular, we also observe local spikes at non-zero count values. While these may be explained by some observed exogenous variables in the corresponding count model, any remaining unexplained spikes in discrete probability mass are easily accommodated in our proposed model using the threshold parameters $\boldsymbol{\alpha}$.

Figure 2 is a thematic map displaying the total number of pedestrian injuries in each census tract. While we have developed the map for each injury severity level separately, the essential visual result that there is geographic clustering in count values holds for all the severity levels. Thus, to economize on space, we are showing only the map for total number of pedestrian injuries across all injury severity levels. The spatial clustering in Figure 2 in the tract-level count of pedestrian injuries should be obvious, which supports the need to accommodate spatial dependency effects in the analysis.

4.3 Sample Descriptives for the Independent Variables

The census tract-level sociodemographic data indicates a racially diverse region. On average, the percentage of the residential population in a census tract is 48% non-Hispanic White, 15% non-Hispanic Black, 12% non-Hispanic Asian, and 23% Hispanic (including Latino and of any race), with 2% being other race/ethnicity combinations (including American Indian, mixed races (not Hispanic), and other non-Hispanic races). The corresponding percentages in the US population as a whole are 64% non-Hispanic White, 12% non-Hispanic Black, 5% non-Hispanic Asian, 16% Hispanic, and 3% other (Humes et al., 2011). A comparison clearly reveals the higher race/ethnicity diversity in the Manhattan population compared to the US population.[9] Further,

[9] Technically speaking, a more appropriate comparison with the US racial profile would be to compute the overall race proportions in the entire Manhattan population (as opposed to computing the proportions by race in each census tract in Manhattan and then taking the average of these proportions across census tracts, as we have done here). However, the intent here is to provide a general picture of the Manhattan study area, while retaining the disaggregation

there is evidence of strong racial clustering within Manhattan, with some census tracts being completely Hispanic in terms of residential population, and some tracts being dominated by White, Black or Asian populations. The remaining socio-demographic variables pertain to education levels and household income, and indicate that, on average, more than half the adult population (18 years or over) in a census tract have a Bachelor's degree or higher, and the median earnings of the households in Manhattan is $72,800. On the other hand, as per the American Community Survey data of the U.S. Census Bureau, the corresponding national statistics for the percentage of the adult population with a Bachelor's degree and the household median earnings are 27.75% and $51,914, respectively. Overall, the descriptive statistics for the socio-demographic variables in the study area indicate a more racially diverse, relatively affluent and highly educated population in Manhattan relative to the country as a whole, though there is a huge variation in the population characteristics across tracts within Manhattan.

Among the land-use and road network variables, the proportion of land-use in a specific type of development (commercial, industrial, residential, and other land uses – vacant lots, open space, recreational) is computed as the ratio of the tract land area in that specific type to the total tract land area. As one may expect, the land-use in the census tracts of Manhattan is predominantly residential (an average proportion of 0.57) and commercial (an average proportion of 0.30), with some tracts being completely invested in residential or commercial land-uses (the average proportion of industrial land-use is 0.7 and the average proportion of other land uses is 0.6). The road network variables are constructed as the ratio of the total length of a specific road type (highways, local neighborhood roads and city streets, and other road types, including alleys and driveways) in the census tract to the total length of the road carriageway in that census tract. The Manhattan census tracts have a high proportion of local neighborhood roads and city streets, with a mean proportion of 0.91. However, there is substantial variation across census tracts, with the minimum proportion being 0.22 and the maximum being 1.00.

---

by census tracts (which is the unit of spatial analysis in the current paper). Overall, however, the census tract-based mean racial proportions computed for Manhattan (as we have done) is close to the racial proportions for the entire Manhattan area population (48% non-Hispanic White, 12.9% non-Hispanic Black, 11.2% non-Hispanic Asian, 25.4% Hispanic, and 2.5% other; as per the American Community Survey data, U.S. Census Bureau). The point of this footnote also applies to some other comparisons undertaken in this section, but we will not belabor over this technicality in the rest of this section.

The activity intensity variables are included to proxy the intensity of non-motorized travel in the census tracts. The number of schools in the census tract refers to the total number of elementary, middle and high schools (both public and private) present in the tract (range of 0 to 10, with a mean of 1.81). The number of Universities is the number of post-secondary degree granting institutions in the census tract (range of 0 to 5, with a mean of 0.15). The intensity of retail activity is computed as the ratio of total floor space allocated for retail use to the total land area of the census tract.[10] The mean value of this variable is 0.18 across tracts, again though with substantial variation across tracts ranging from a low of 0 to a high of 1.62.

The commute mode share and transit supply variables reveal the high transit and walk mode shares in the region (mean transit share across tracts is 0.57 and mean walk share across tracts is 0.22). The transit supply variables are measured by the number of bus stops (varies from 0 to 60 across tracts, with a mean of eight bus stops per tract) and the number of subway stops (varies from 0 to six across tracts, with a mean of 0.5 subway stops per tract). Finally, the spatial proximity metric used in characterizing the spatial dependence between any pair of census tracts is computed as the Euclidean distance (in miles) between tract centroids. This metric has an average value of 3.78 miles, with a minimum of 0.09 miles and a maximum of 13.15 miles.

4.4 Model Selection and Variable Specification

Several weight matrix specifications were considered in our empirical analysis to characterize the nature of the dynamics of the spatial lag dependence. These included (1) a contiguity specification that generates spatial dependence based on whether or not two tracts are contiguous, (2) another contiguity specification but based on shared boundary length, (3) weights based on *k*-nearest neighbors, (4) the inverse of a continuous Euclidean distance specification between census tracts, (5) the inverse of the square of the continuous distance specification, and (6) the inverse of the exponential of the continuous distance specification. For the last three continuous distance-based specifications, we also explored alternative distance bands (2 miles, 5 miles, 7.5 miles, 10 miles, and 15 miles, the last of which corresponds to considering all pairs of tracts) to select the distance band to consider for the pairing of tracts in the composite marginal likelihood (CML) estimation.

[10] Technically speaking, the net floor area in retail in a census tract can be more than the land area of the census tract (because of the vertical build-up). Thus, the retail intensity measure can be higher than 1 (the land-use measures previously discussed, however, are confined to the 0-1 range).

This is because, as discussed in detail in Bhat (2011) and Bhat (2014), a higher efficiency of the CML estimator can be achieved by lowering the number of pairings used in the CML estimation. The best estimator efficiency, based on minimizing the trace of the asymptotic covariance matrix (see Bhat, 2014), was obtained with a distance band of 5 miles for all three continuous distance specifications, which was then retained in subsequent estimations. The determination of the best weight specification was next based on the composite likelihood information criterion (CLIC) statistic, which may be used to compare the data fit of non-nested formulations (see Varin and Vidoni, 2005; Bhat, 2014). This CLIC statistic takes the form shown below:

$$\text{CLIC} = \log L_{CML}(\hat{\boldsymbol{\theta}}) - tr\left[\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}})^{-1}\right],$$

where $\hat{\boldsymbol{\theta}}$ is the estimated model parameter vector, and $\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}})$ are the "vegetable" and "bread" matrices used in the estimation of the asymptotic variance matrix $\mathbf{V}_{CML}(\hat{\boldsymbol{\theta}})$. In the current context, the weight specification that provides the highest value of the CLIC statistic is preferred over the other competing weight specifications. Our results indicated that, for all variable specifications we attempted and for all injury severity categories, the best spatial weight matrix specification was consistently the inverse of the continuous distance specification with a 5-mile distance band.

Concurrent with the weight matrix specification, we also explored several different variable specifications and functional forms of the variables. Except for a handful of exogenous count variables, all other variables are continuous. For the continuous variables, we considered the variables as is, in logarithmic form (to introduce marginally decreasing effects), in spline form (to introduce flexible and piecewise non-linear effects), as well as dummy variables representing ranges. The exogenous count variables (number of schools and universities, number of bus stops, and number of subway stops) were introduced as is. All the variables were introduced in both the latent variable and threshold specifications. Interaction effects of many of the variables were also considered. The final variable specification was based on intuitive, data fit, and statistical significance considerations, combined with a healthy dose of practical realism in the number of combinations of variables and functional forms that can be tested.

4.5 Model Estimation Results

Several sets of parameters constitute the model proposed here. In this section, we first discuss the mean effects of the exogenous variables on the long-term injury risk propensities (the $\boldsymbol{b}_j$ parameters), along with the covariance effects of these exogenous variables (the $\boldsymbol{\Omega}_j$ elements), for the count models of different injury severity levels j. We next proceed to the elements of the $\boldsymbol{\Xi}_k$ for each variable k that engender multivariateness across injury levels, and then to the effects of exogenous variables on the thresholds (the $\boldsymbol{\gamma}_j$ parameters) and the $\alpha_{j,y_{qj}}$ elements in the thresholds. Finally, we present the spatial dependency effects δ_j.

4.5.1 Long-Term Injury Risk Propensity

4.5.1.1 Socio-Demographic Variables

Table 1 provides the effects of variables and related covariance elements. As expected, the table shows a positive mean effect of the logarithm of population density (which is the main exposure measure in terms of pedestrian traffic in the current paper) for all severity levels. There also is substantial heterogeneity (across census tracts) in the effect of this variable for the two lower injury severity levels of "possible" and "non-incapacitating" counts. In fact, the results indicate that an increase in population density is actually associated with a decrease in the risk propensity of possible injuries and non-incapacitating injuries for 1% and 5% of census tracts, respectively. This wide variation in the effect of population density (including positive and negative effects) represents a mix of different factors that moderate the population density influence, including the positive effects of exposure and potential social deprivation effects, as well as negative effects that may be related to more cautious driver behavior because of awareness of pedestrians and reduced motorist driving speeds. Earlier studies that have not considered heterogeneity effects indicate a pure positive or no effect at all on some injury severity categories (such as Ha and Thill, 2011, Jermprapai and Srinivasan, 2014, and Narayanamoorthy et al., 2013) or a pure negative effect (see, for example, Pharr et al., 2013). Our results reconcile the apparent contradictory results from earlier research by allowing a range of effects of population density. Also interesting from our results is that there is no heterogeneity in the population density effect on pedestrian injury risk propensity for the more severe "incapacitating" and "fatal" injury levels, a finding also obtained by Amoh-Gyimah et al. (2016). This suggests that pedestrian exposure and any social deprivation

effects uniformly impact risk propensity for severe pedestrian injuries, and especially so for the highest severity level of fatal injuries. Note that, since all the propensity risks are normalized to the same error scale of one, the coefficients on a variable are comparable across injury severity levels in Table 1 if they both do not have unobserved heterogeneity associated with them. But, in the presence of unobserved heterogeneity (as for the population density effects on the two less severe injury levels in Table 1), elasticity effects will need to be computed to determine relative magnitude effects, which we undertake in Section 4.7.

The next two variables relate to age and education distributions. The first of these reveal that a higher proportion of young individuals (19 years of age or less) in a tract tends to decrease pedestrian risk propensity for "incapacitating" injuries. This result is not consistent with some earlier studies (see, for example, Amoh-Gyimah et al., 2016 and more extensive reviews in Stoker et al., 2015 and Rothman et al., 2016) that suggest that a combination of less developed safety cognition/navigation abilities and larger variation in actions results in an increase in the risk of pedestrian crashes among young individuals. However, the alternative explanation is that much of the children-oriented pedestrian exposure takes place in and around school zones, where low speed limits are in place and there is more awareness among drivers. Interestingly, though, we did not find statistically significant heterogeneity on the "proportion of young individuals" variable that can reconcile the differing perspectives, suggesting that the analysis contexts may matter. For instance, Amoh-Gyimah et al.'s study is based on data on the entire city of Melbourne, Australia, with relatively aggregate zones as the analysis unit, while ours is based on data from a specific part (Manhattan) of New York City with much smaller census tracts as the analysis unit. The next variable in Table 1 is an education-related variable that proxies low education levels in a tract (captured based on the proportion of adults without a high school degree in the tract). The effect of this variable indicates the higher risk of incapacitating pedestrian injuries in tracts with low education levels. This is not surprising, because, as Vaughn et al. (2011) note, a low education level (specifically less than a high school education) is positively correlated with more reckless driving that can put pedestrians in particular at risk. In addition, individuals with low levels of education seem to face more challenges in understanding traffic symbols and signs (Al-Madani and Al-Janahi, 2002).

The next two variable effects on the risk for "incapacitating" and "fatal" injury categories in the table, suggest, in general, the presence of higher exposure effects (more pedestrian

movements) and social deprivation effects in tracts with a high proportion of Hispanic population and low median household income. Similar results have been found in many other macro-level analyses of pedestrian crashes (see, for example. Chakravarthy et al., 2010, Cottrill and Thakuriah, 2010, Karsch et al., 2012, and Jerrett et al., 2016). The deprivation effects (which may be another explanation also for the higher pedestrian risk for incapacitating injuries in tracts with low education) can include factors such as the absence of sidewalks, footpaths, and appropriate traffic signage/signals in predominantly Hispanic and/or low income neighborhoods. Further, the absence of public spaces such as recreational parks and centers in such neighborhoods appears to result in the streets being treated as public spaces, leading to more pedestrian exposure to motorized traffic (see Cooper et al., 2015). Coughenour et al. (2017) also indicate that drivers appear to be racially biased and less likely to yield to a person of color compared to a white individual, when the individual is already in the roadway crossing a street. Interestingly, though, our results show no statistically significant differences in the specific risk of "possible" pedestrian injury crashes in tracts with a high proportion of Hispanics, and show a mean negative impact of the proportion of Hispanic population on the risk for "non-incapacitating" pedestrian crashes. This reduction in risk propensity may be attributable to more experience in navigating pedestrian-heavy travel environments in relatively Hispanic-dense living neighborhoods. However, there is also considerable heterogeneity in this mean effect with a negative effect of the variable for 55% of census tracts and positive effect for 45% of the census tracts. Similar substantial heterogeneity is also reflected in the effects for the "incapacitating" and "fatal" injury severity levels, suggesting a mix of experience/awareness of pedestrian movements (resulting in lower risk) and exposure/deprivation effects (resulting in higher risk) in the effect of the "proportion of Hispanic population" variable.

4.5.1.2 Land-use and Road Network Variables

Land-use and road network variables also have an impact on the long-term injury propensities. The increased risk of "incapacitating" pedestrian injuries in tracts with a higher proportion of commercial and residential land-use (relative to the industrial and other land-uses) may be traced to exposure effects, as commercial/residential areas land-uses are associated with high pedestrian movement activity and pedestrian-vehicular conflict (see also Yu and Zhu, 2016, Quistberg et al., 2015, and Rothman et al., 2016). The effect of the "proportion of local neighborhood roads and

city streets" is to reduce the number of non-incapacitating and fatal injuries, potentially because of the lower speed limits on such streets. However, the standard deviation on this variable for the "fatal" injury level also suggests that, in 32% of the census tracts, a higher proportion of local neighborhood roads and streets actually contribute to an increase in the risk of fatal injuries, perhaps again because of the unexpected nature of crashes at such locations.

4.5.1.3 Activity Intensity Variables

Among the activity intensity variables, the results generally reflect an exposure effect with an increase in the risk of non-fatal injury crashes. The only exception is the negative effect of the "number of schools" on the incapacitating injury risk propensity, which may be explained by the low speed limits (and more time for drivers to react to hazards; see Yu and Zhu, 2016) in the vicinity of schools. However, as we will note later, there is also a strong risk-to-injury translation represented in the threshold elements (discussed in Section 4.5.3) that, in totality, leads actually to an increase in pedestrian injuries in tracts with a higher number of schools.

4.5.1.4 Commute Mode Shares and Transit Supply

In the final set of variables, walk commute mode share has the expected positive effect on the risk propensity for fatal pedestrian injuries not only solely due to an exposure effect, but also potentially because of the social deprivation effects already discussed (the population segments affected by social deprivation have less of a choice to use motorized modes and have to use non-motorized modes in a potentially unsafe walking environment). Also, the "number of subways stops" tends to increase the risk propensity of crashes with "possible" injury, a reflection of subway stops being focal points of pedestrian activity either to access the transit system or one of the many activity locations in the vicinity of subway stops (Ukkusuri et al., 2012; Yu and Zhu, 2016). But because subway stops also represent locations of good pedestrian infrastructure, any crashes at these locations tend to be relatively minor. Also, consistent with differences across tracts in the quality of the pedestrian infrastructure in and around subway stations, there is a good bit of variation in the effect of this variable on risk propensity.

4.5.1.5 Non-diagonal Elements of $\boldsymbol{\Omega}_j$

The last set of entries in Table 1 correspond to the non-diagonal elements of the covariance matrix $\boldsymbol{\Omega}_j$ for each injury severity level (the diagonal elements correspond to the square of the standard deviations of variables provided along with the mean values earlier in the table). Two such covariances turned up to be statistically significant. The first is the positive covariance between the effect of the constant and the effect of the "number of subway stops" variable for the "possible injury" level. Essentially, this is indication that there are some common set of elements (say, for example, jaywalking, as discussed in the introduction section) between unobserved factors that increase the overall risk propensity of "possible injuries" and the unobserved factors that moderate the effect of the "number of subway stops" variable on the risk propensity of "possible injuries". The second covariance element is that between the effects on log of population density and intensity of retail activity for the "non-capacitating" injury level. The positive covariance here may represent unobserved factors (such as poor pedestrian infrastructure) that simultaneously increase the risk associated with the increased pedestrian activity associated with high population density locations and high retail intensity locations.

4.5.2 Multivariate Effects Across Severity Levels (the Ξ_k elements for each k)

In this section, we present the elements that engender a covariance across risk propensities of different injury severity levels within the same tract. The generic covariances in the risk propensities (that is, covariances among the constants) show, as expected, a positive and statistically significant covariance across four pairs of constants (but not between the constants of the "possible injury" level with the constants of the "incapacitating" and "fatal" injury levels).[11] These covariances likely reflect geometric, design, and other unique unobserved features within a tract that simultaneously move the injury risk up or down for all injury levels. But, in addition, the random coefficients corresponding to three variables also showed covariation in risk across specific pairs of injury severity levels: between the log of population density parameters in the "possible" and "incapacitating" injury levels (value of 0.264; t-statistic of 3.27), between the

[11] The covariance estimates, all significant at the 95% level, were as follows: 0.132 (between the constants in the "possible" and "non-incapacitating" injury categories), 0.262 (between the constants in the "non-incapacitating" and "incapacitating" injury categories), 0.177 (between the constants in the "non-incapacitating" and "fatal" injury categories), and 0.396 (between the constants in the "incapacitating" and "fatal" injury categories).

proportion of Hispanic population parameters in the "non-incapacitating" and "incapacitating" levels (0.400; t-statistic of 3.11), and between the proportion of Hispanic population parameters in the "incapacitating" and "fatal" levels (0.192; t-statistic of 2.59). These covariation effects may be attributed to, for example, factors such as the absence of sidewalks, footpaths, and appropriate traffic signage/signals associated with high density and predominantly Hispanic neighborhoods that simultaneously increase the risk for pedestrian injuries of multiple levels.

4.5.3 Threshold Parameters

The threshold parameters include, for the count model for each injury severity level, the threshold specific constants (the $\alpha_{j,y_{qj}}$ parameters) as well as the $\boldsymbol{\gamma}_j$ parameters. These are provided in Table 2. The threshold specific constants do not have any substantive interpretations but their presence provides flexibility in the count model to accommodate high or low probability masses for specific outcomes (after controlling for the effect of other exogenous variables), as discussed in Section 3.1. But the negative value for $\alpha_{1,5}$ (labeled in Table 2 simply as α_5 because all the threshold-specific constants are relevant only to the first injury level of *j*=1) has the effect of opening up the window corresponding to the count value of six, thereby increasing the probability mass for a count of six, which is consistent with the discernible spike at this count value for "possible" injury in Figure 1.

The elements of the $\boldsymbol{\gamma}_j$ vector for injury severity level *j*, as discussed in Section 3.1, translate the long-term crash risk propensity for injury level *j* into the actual occurrence of an injury of that level at any given *instant of time*. The constants in the $\boldsymbol{\gamma}_j$ vectors in Table 2 do not have any substantive interpretation, and simply adjust for the values of other exogenous variables impacting the threshold. For other variables, if an element of the $\boldsymbol{\gamma}_j$ is positive, an increase in the corresponding variable has the effect of shifting the thresholds toward the left on the crash propensity scale and reducing the probability of the zero injury count for injury level *j* (see CPB, 2012). In other words, a positive element implies that local spatial/temporal circumstances preceding a crash (and associated with the variable in question) are such that a given risk is more likely to manifest itself in injuries because of a risk "reinforcing" effect. On the other hand, a negative element implies that an increase in the corresponding variable increases the probability

of the zero injury count for injury level j. That is, a negative element implies that local spatial/temporal circumstances preceding a crash (and associated with the variable in question) are such that a given risk is less likely to manifest itself in injuries because of a "cushioning" effect. With that as a prelude, it is first interesting to note from Table 2 that the number of variables that moderate the translation of risk to injury decrease as we go from the lowest injury severity level to the highest. This is intuitive, as there can only be less of moderating effects as more serious injuries are about to happen. Second, almost all the moderating variables are built environment variables rather than socio-economic variables, because the risk-moderating variables are likely to be associated with the environment more so than socio-economics. However, the first variable under the γ vector is "median household income", and the effect of this variable suggests that, given a certain level of risk propensity for fatal injuries, tracts with higher income (relative to tracts with lower income) are more likely to actually see fatal injuries. Combined with the lower risk propensity to begin with in higher income tracts (see Table 1), the suggestion is that higher income tracts probably have better pedestrian facilities, but motorists are not used to pedestrian-motor vehicle conflicts so that when one starts to develop, they are less experienced in navigating through to avoid fatalities. The net effect of being in a higher income tract on pedestrian fatalities will be a combination of the risk propensity effect and the threshold effect, which shows a strong negative effect on fatalities (see the discussion later in Section 4.7). However, from a policy standpoint, the implication is that there would be value in education and information campaigns for motorists and pedestrians even in high income neighborhoods, even if such neighborhoods have much fewer pedestrian fatalities than low income neighborhoods. The other results indicate the risk reinforcing effects for the lower injury levels (but not for the "fatal" level) of being in dominantly commercial and residential tracts, as well as in tracts with a number of schools and high walk commute mode share, possibly due to distraction effects as well as less time (due to the density of pedestrian traffic) to avoid crashes. On the other hand, there is a cushioning effect of the risk in tracts with high proportions of secondary roads, local neighborhood roads and city streets, and minor roads, suggesting that the low speeds on these roads add a layer of safety so that both pedestrians and motorists can identify an impending collision and have time to take evasive action.

4.5.4 Spatial Dependence

The spatial dependence parameters lead to inter-relationships or spatial spillovers in the risk propensity across proximal census tracts. The dependence parameter for each injury severity level is statistically significant and positive (see last row of Table 2), indicating a clear "snowballing" effect in our empirical analysis. That is, a change in an independent variable in one tract impacts not only the long-term risk propensity in that tract in a specific direction, but also affects the long-term risk propensity in the same direction in surrounding tracts within a 5-mile distance band. Additionally, the intensity of this "snowballing" effect increases with an increase in the injury severity level. This supports our hypothesis that pedestrian injury risk has different spatial dependency patterns based on the injury severity level (Narayanamoorthy et al., 2013, on the other hand, constrained all the dependence parameters to be the same across injury levels). As we will see later, ignoring spatial dependencies or constraining these to be the same across injury severity levels provides an inaccurate picture of the effects of socio-economics and built environment variables.

In addition to spatial spillovers, the results indicate the presence of clear spatial drift in the variable coefficients; that is, a spatial pattern in the unobserved variable effects across tracts caused by the combination of unobserved heterogeneity across tracts and spatial dependence across tracts. For example, as discussed earlier in the introduction section, there is reason to believe that the unobserved factors (such as MFTP) moderating the effect of the "proportion of local neighborhoods roads and city streets" on fatal injuries will themselves be spatially correlated. Our modeling approach constitutes a simple, elegant, and intuitive structural mechanism to capture such spatial drift effects in multiple variables at once.

4.6 Measures of Data Fit

In this section, we examine the data fit of the proposed multivariate unconstrained spatial count model (MUSC) with its more restricted versions, as discussed in Section 3.3: (1) the independent aspatial count (IAC) model, (2) the independent constrained spatial count (ICSC) model, (3) the independent unconstrained spatial count (IUSC) model, (4) the multivariate aspatial count (MAC) model, and (5) the multivariate constrained spatial count (MCSC) model. These models may be tested against the MUSC model using the adjusted composite likelihood ratio test (ADCLRT) statistic, which is asymptotically chi-squared distributed (similar to the likelihood ratio test statistic

for the maximum likelihood approach[12]). In addition to testing the models using the disaggregate ADCLRT statistic, we also predict the expected count of injuries at each severity level at each census tract (the procedure is discussed in Appendix A), aggregate these expected counts at each injury severity level across all census tracts, and compare these aggregate prediction counts with the actual counts of injuries at each severity level for the entire Manhattan region using the Mean Absolute Percentage Error (MAPE) measure. This provides another intuitive way of assessing model performance.

Table 3 presents the results of the data fit comparisons. For completeness, we note that the composite log-likelihood (CLL) value of the naïve model with no variables in the injury risk propensity of each severity level and only a constant in the threshold (that is, only a constant in the z_q vector) is -2,962,001.29 (the second numeric row). The number of parameters and the CLL values for the five models (from the simplest IAC model to the proposed MUSC model) are provided in the third and fourth numeric rows, respectively, of Table 3. The fifth numeric row compares the MUSC model with its restrictive versions using ADCLRT tests, which clearly demonstrate the superior performance of the MUSC model relative to other models. While not shown in the table, an ADCLRT of the IAC model and ICSC model (the first and second models in Table 3) clearly rejects the IAC model, supporting the notion of the presence of spatial effects (spillover, error correlation, and drift effects). A similar result is obtained when comparing the MCSC model with the MAC model. Between the ICSC and IUSC models, and the MCSC and MUSC models, the unconstrained versions come out to be the clear winners, implying that it is not only important to accommodate for spatial dependencies, but also allow the dependencies to vary across injury severity levels. This is also reflected in the next numeric row that presents the estimated spatial autoregressive lag parameters for the many models. The reader will note that these parameters are set to zero for the IAC and MAC models, and constrained to be equal across the injury severity levels for the ICSC and MCSC models. Finally, the last row panel of the table compares the predicted and actual counts at each injury severity level over the entire Manhattan region. The last row indicates substantial MAPE differences not only between the IAC and ICSC (and MAC and MCSC models), but also between the ICSC and IUSC (and MCSC and MUSC)

[12] The variable specifications did not change between the many different models, and so all the models are nested versions of the MUSC model.

models. Indeed, the MUSC model rejects all other models based on not only the disaggregate ADCLRT test (as discussed earlier), but also based on the aggregate MAPE statistic.

4.7 Aggregate Elasticity Effects

The model fit comparisons in the previous section show the benefits of our proposed MUSC formulation for modeling the count of pedestrian injuries by severity level. But to examine if the MUSC model also predicts quite different effects of the risk factors, we compute the aggregate-level "elasticity effects" from the many different models. To focus the presentation and avoid clutter, we provide the elasticity results for only the three multivariate models: the MAC, the MCSC, and the MUSC, and for only the two more serious "incapacitating" and "fatal" injury levels. For each variable, the "elasticity" computed is a measure of the percentage change in total injury count (for each injury severity level) across the entire study region due to a change in the variable value in each tract (the procedure is the same as that used to compute the aggregate prediction counts in the data fit section for the entire Manhattan region). To compute the aggregate level "elasticity effect" of the socio-demographic and land use and road network variables that appear in the final specification (all of which, except median household income and population density, are in the form of proportions), we increase the proportion of each variable in each tract by 0.2. For the median household income variable and the population density variable, we increase the value by 20% in each tract. For the activity intensity variables, the count variables (number of schools and Universities) are increased by one for every tract, while the intensity of office/retail activity are inflated by 20%. In the group of commute mode share and transit supply variables, the walk commute mode share is increased by 0.2 for each tract (with a cap again at 1.00), and the number of subway stops is increased by a value of one for each tract.

Table 4 presents the elasticity effects for the three models (along with their t-statistics) for the "incapacitating" and "fatal" levels. The first two row entries in the first numeric row of the table indicate that an increase in population density by 20% in each tract would result in a 5.7% and 25.7% increase in the annual count of "incapacitating" and "fatal" pedestrian injuries, respectively, in the entire study region, according to the predictions of the MAC model. Other entries may be similarly interpreted. As a general comment, all the elasticity effects are statistically significant at the 95% level of confidence, except for the elasticity effect of population density on incapacitating injuries for the three models (these are significant only at the 90% level of

confidence). In addition, except for the population density effect on incapacitating injuries, there is a clear pattern of increasing elasticities as one goes from the MAC model to the MUSC model. This is because the aspatial model completely ignores the spatial spillover ("snowballing" effect in our case) due to spatial dependence. The MCSC model accommodates the "snowballing", but constrains this effect to be the same across all injury severity levels (and, as already discussed in Section 4.5.4, the snowballing effect actually increases with pedestrian injury severity level, as correctly reflected in the MUSC model). Overall, the MAC and MCSC models underestimate the elasticity effects relative to the MUSC model. The underestimations are particularly evident for the critical "fatal" injury severity level, because it is in this category that the "snowballing" effect is the highest. Thus, for example, according to the MUSC model, an increase in the proportion of Hispanic population by 0.2 in each tract would result in a 46.1% increase in the annual count of fatal pedestrian injuries. In the same situation, the MAC model predicts only a 21.8% increase and the MCSC model predicts only a 33.8% increase. Another way to view this is that the "direct elasticity effect" is 21.8% (from the MAC model that ignores spatial dependence), while the "indirect elasticity effect" due to spillovers (the snowballing effect) is estimated to be 12.0% (=33.8-21.8) from the MCSC model and 24.3% (=46.1-21.8) from the MUSC model. Indeed, the spillover effect is very substantial. Thus, in this case, not accounting for spatial dependence (and in a flexible manner as in our proposed MUSC model) would substantially underestimate (and ill-inform) the importance of focusing in on tracts with high Hispanic population proportions as "hot spots" for targeted countermeasures. Similar conclusions may be drawn for other variables. Another intriguing result from Table 4 is the suggestion that, while the risk propensity for "incapacitating" injuries is lower in school areas, distractions and the unexpectedness of crashes in these areas actually lead to an overall increase in pedestrian "incapacitating" injuries.

4.8 Policy Implications

Our census tract-level analysis can provide important information for pedestrian crash countermeasures and interventions. First, our analysis emphasizes the importance of identifying "hot spots" based on the socioeconomic status (SES) of tracts, as quantified by such measures as high population density, high Hispanic proportion in the population, low education levels, low income earnings, and high walk commute mode shares. Further, the aggregate elasticity effects from the previous section suggest that the differences in pedestrian injury counts across low SES

neighborhoods and other neighborhoods is much more stark than what traditional models suggest. This further reinforces the need to identify "hot spots" based on SES. Many countermeasures may be implemented, some based on engineering approaches and others on behavioral approaches. Engineering approaches may include measures such as better lighting, cross-walks, continuous footpaths, traffic signage/signals, and also good recreational/open spaces so that streets do not get treated as public spaces. Behavioral countermeasures may include campaigns that reduce reckless driving, improve pictorial symbols/signs to make their recognition and understanding easier, and promote racial tolerance so that motorists treat all pedestrians in the same way as opposed to yielding less to people of color. Of course, additional research investigating the relative reasons for our area-level analysis results of social deprivation and higher pedestrian fatalities in low SES communities would assist in determining more concrete thrust mechanisms for the countermeasures. Further, a better understanding of the precise reasons for the substantial unobserved heterogeneity in risk across different Hispanic-dense tracts would be helpful, and can inform factors and contexts that do not lead to a degradation of safety (and actually can increase safety) in minority-dense tracts. Second, the positive effects of the land-use variables (proportion of commercial and residential land-use, and intensity of retail activity) on the number of incapacitating pedestrian injuries suggest engineering countermeasures that reduce vehicular-pedestrian conflicts (through more pedestrian-friendly design rather than the typical auto-oriented design) as well as behavioral countermeasures such as espousing less distracted driving. Again, our proposed model suggests that policy analysts may be underestimating the need for such countermeasures based on an incorrect and downward-biased assessment of the magnitude of increased pedestrian injuries at high activity intensity locations. Third, there appears to be a need for information campaigns regarding driving (and walking) in and around schools that emphasizes vigilance in the face of a seemingly safe environment that may lull people into complacency. This is because the low risk propensity for serious injuries in and around schools is more than compensated by the risk-to-injury translation effect, as discussed in the previous section. This kind of disentangling between long-term risk propensity and short-term risk-to-injury translation would not have been possible if not for the modeling framework in this paper. The upshot is that the methodology proposed here indicates a continued need for implementation (or continuation) of programs such as the "Safe Routes to School" initiative. Further, our results recommend that policy makers should avoid school siting and commercial/retail land use in close proximity. Finally, our

results based on the model proposed in this paper indicate that behavioral countermeasure campaigns should not simply target areas with currently low levels of pedestrian injuries, which is the underlying concept of "hot spot" detection. For example, even though there are much fewer "fatal" pedestrian injuries in high income areas, our proposed model indicates that the risk-to-injury translation is high in these areas, perhaps because, just as around schools, drivers (and pedestrians) are not used to vehicular-pedestrian conflict and are not experienced enough to navigate those occasional instances of conflict.

5 CONCLUSIONS

In this paper, we contribute to the pedestrian crash literature by formulating a macro-level (*i.e.* census tract level as opposed to roadway street segment level) multivariate model to jointly analyze the count of pedestrians involved in traffic crashes by different injury severity levels. Specifically, we propose a comprehensive spatial random coefficients multivariate count model that recognizes many issues at once: (a) It addresses excess zeros within a multivariate count setting through a recast of the traditional count model as a special case of a generalized ordered-response model, (b) It allows variation in the effects of determinant exogenous variables because of unobserved location-specific factors, (c) It accommodates spatial dependency in the count of pedestrian injuries based on location proximity, and (d) It captures spatial drift effects through the spatial structure on constant and slope heterogeneity effects.

The proposed model is applied to examine, at the spatial level of a census tract, all pedestrian crashes in the year 2009 in the Manhattan area of New York City. The results reinforce the need for studying pedestrian injuries by severity level, and accommodating unobserved heterogeneity, multivariateness, and spatial dependence. In particular, the determinants of different levels of injury severity do vary, as do the intensity of exogenous variables. As identified in the introduction section of this paper, the economic and societal cost of crashes vary substantially based on the nature and extent of injuries sustained, and so it is imperative to consider injury counts by severity level. The empirical analysis sheds light on possible engineering as well as behavioral countermeasures to reduce the number of pedestrian-vehicle crashes and the severity of these crashes. Of course, with any macro-level quantitative analysis of the type in this paper, estimated relationships provide more of correlational rather than causal connections. Future research needs to consider and combine additional micro-level qualitative and quantitative analyses to identify

precise causation pathways and countermeasures. The inclusion of better pedestrian infrastructure variables, whether at a macro-level or a micro-level, is also important in this regard, and calls for more concerted efforts to capture this type of data.

ACKNOWLEDGMENTS

This research was partially supported by the U.S. Department of Transportation through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center. The first author would like to acknowledge support from a Humboldt Research Award from the Alexander von Humboldt Foundation, Germany. The authors are grateful to Lisa Macias for her help in formatting this document, and to two anonymous referees who provided useful comments on an earlier version of the paper.

REFERENCES

Aguero-Valvarde, J., Jovanis P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania, *Accident Analysis & Prevention* 38(3), 618-625.

Amoh-Gyimah, R., Saberi, M., Sarvi, M., 2016. Macroscopic modeling of pedestrian and bicycle crashes: A cross-comparison of estimation methods. *Accident Analysis & Prevention* 93, 147-159.

Amoh-Gyimah, R., Saberi, M., Sarvi, M., 2017. The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. *Analytic Methods in Accident Research* 13, 28-51.

Al-Madani, H., Al-Janahi, A. R., 2002. Role of drivers' personal characteristics in understanding traffic sign symbols. *Accident Analysis & Prevention* 34(2), 185-196.

Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. *Analytic Methods in Accident Research* 11, 17-32.

Barua, S., El-Basyouny, K., Islam, M.T., 2015. Effects of spatial correlation in random parameters collision count-data models. *Analytic Methods in Accident Research* 5, 28-42.

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

Beck, N., Gleditsch, K.S., Beardsley, K., 2006. Space is more than geography: Using spatial econometrics in the study of political economy. *International Studies Quarterly* 50(1), 27-44.

BITRE (The Bureau of Infrastructure, Transport and Regional Economics), 2013. Road Deaths Australia, 2012 Statistical Summary. Canberra ACT.

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

Bhat, C.R., 2014. The Composite Marginal Likelihood (CML) inference approach with applications to discrete and mixed dependent variable models. *Foundations and Trends (R) in Econometrics* 7(1), 1-117.

Bhat, C.R., 2015a. A new spatial (social) interaction discrete choice model accommodating for unobserved effects due to endogenous network formation. *Transportation* 42(5), 879-914.

Bhat, C.R. 2015b. 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., Sener, I.N., Eluru, N., 2010. A flexible spatially dependent discrete choice model: formulation and application to teenagers' weekday recreational activity participation. *Transportation Research Part B* 44(8), 903-921.

Bhat, C.R., Paleti, R., Singh, P., 2014a. A spatial multivariate count model for firm location decisions, *Journal of Regional Science* 54(3), 462-502.

Bhat, C.R., Born, K., Sidharthan, R., and Bhat. P.C., 2014b. A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. *Analytic Methods in Accident Research* 1, 53-71.

Bhat, C.R., Pinjari, A.R., Dubey, S.K., Hamdi, A.S., 2016. On accommodating spatial interactions in a generalized heterogeneous data model (GHDM) of mixed types of dependent variables. *Transportation Research Part B* 94, 240-263.

Blincoe, L. J., Miller, T. R., Zaloshnja, E., Lawrence, B. A., 2015. The economic and societal impact of motor vehicle crashes, 2010 (Revised), Report No. DOT HS 812 013, National Highway Traffic Safety Administration, Washington, DC.

Bradlow, E.T., Bronnenberg, B., Russell, G.J., Arora, N., Bell, D.R., Duvvuri, S.D., Hofstede, F.T., Sismeiro, C., Thomadsen, R., Yang, S., 2005. Spatial models in marketing. *Marketing Letters* 16(3), 267-278.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(3), 431-443.

Buddhavarapu, P., Scott, J.G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. *Transportation Research Part B* 91, 492-510.

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 & 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.

Castro, M., Paleti, R., Bhat, C., 2013. A spatial generalized ordered response model to examine highway crash injury severity. *Accident Analysis & Prevention* 52, 188-203.

Chakravarthy, B., Anderson, C.L., Ludlow, J., Lotfipour, S., Vaca, F.E., 2010. The relationship of pedestrian injuries to socioeconomic characteristics in a large Southern California County. *Traffic Injury Prevention* 11(5), 508-513.

Cooper, J. F., Wilder, T. R., Lankina, E., Geyer, J. A., Ragland, D. R., 2015. Traffic safety among Latino populations in California: Current status and policy recommendations. *Safe Transportation Research & Education Center*.

Cottrill, C.D., Thakuriah, P.V., 2010. Evaluating pedestrian crashes in areas with high low-income or minority populations. *Accident Analysis & Prevention* 42(6), 1718-1728.

Coughenour, C., Clark, S., Singh, A., Claw, E., Abelar, J., Huebner, J., 2017. Examining racial bias as a potential factor in pedestrian crashes. *Accident Analysis & Prevention* 98, 96-100.

Delmelle, E.C., Thill, J.-C., Ha, H.-H., 2011. Spatial epidemiologic analysis of relative collision risk factors among urban bicyclists and pedestrians. *Transportation* 39(2), 433-448.

Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis & Prevention* 40(3), 1033-1054.

Godambe, V.P., 1960. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* 31(4), 1208-1211.

Greene, W.H., Hensher, D.A., 2010. *Modeling ordered choices: A primer*. Cambridge University Press.

Ha, H.H., Thill, J.C., 2011. Analysis of traffic hazard intensity: A spatial epidemiology case study of urban pedestrians. *Computers, Environment and Urban Systems*, 35(3), 230-240.

Heydari, S., Fu, L., Miranda-Moreno, L.F., Joseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. *Analytic Methods in Accident Research* 13, 16-27.

Huang, H., Abdel-Aty M., Darwiche A., 2014. County-level crash risk analysis in Florida. *Transportation Research Record: Journal of the Transportation Research Board* 2148(4), 27-37.

Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. *Analytic Methods in Accident Research* 14, 10-21.

Humes, K., Jones, N.A., Ramirez, R.R., 2011. Overview of race and Hispanic origin, 2010. US Department of Commerce, Economics and Statistics Administration, US Census Bureau.

Jermprapai, K., Srinivasan, S., 2014. Planning-level model for assessing pedestrian safety. *Transportation Research Record: Journal of the Transportation Research Board* 2464, 109-117.

Jerrett, M., Su, J.G., Macleod, K.E., Hanning, C., Houston, D., Wolch, J., 2016. Safe routes to play? Pedestrian and bicyclist crashes near parks in Los Angeles. *Environmental Research* 151, 742-755.

Kamargianni, M., Dubey, S.K., Polydoropoulou, A., Bhat, C.R., 2015. Investigating the subjective and objective factors influencing teenagers' school travel mode choice - An integrated choice and latent variable model. *Transportation Research Part A* 78, 473-488.

Karsch, H.M., Hedlund, J.H., Tison, J., Leaf, W.A., 2012. Review of Studies on Pedestrian and Bicyclist Safety. National Highway Traffic Safety Administration, Washington, DC, 1991–2007, Report No. DOT HS 811 614.

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

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.

McMillen, D.P., 2010. Issues in spatial data analysis. *Journal of Regional Science* 50(1), 119-141.

Moudon, A., Lin, L., Jiao, J., Hurvitz, P., Reeves, P., 2011. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County Accident Washington. *Accident Analysis & Prevention* 43(1), 11–24.

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.

NHTSA (National Highway Traffic Safety Administration), 2016a. Traffic Safety Facts Research Note - 2015 Motor Vehicle Crashes: Overview. Report No. DOT HS 812 318, National Center for Statistics and Analysis, U.S. Department of Transportation, Washington, DC. Available at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812318.

NHTSA (National Highway Traffic Safety Administration), 2016b. Traffic Safety Facts 2014 Data - Pedestrians. Report No. DOT HS 812 270, National Center for Statistics and Analysis, U.S. Department of Transportation, Washington, DC. Available at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812270.

Pharr, J., Coughenour, C., Bungum, T., 2013. Environmental, human and socioeconomic characteristics of pedestrian injury and death in Las Vegas, NV. *International Journal of Sciences* 2(10), 31-39.

Quistberg, D.A., Koepsell, T.D., Johnston, B.D., Boyle, L.N., Miranda, J.J., Ebel, B.E., 2015. Bus stops and pedestrian-motor vehicle collisions in Lima, Peru: A matched case-control study. *Injury Prevention* 21(e1), e15-e22.

Rasciute, S., Downward, P., 2010. Health or happiness? What is the impact of physical activity on the individual? *Kyklos* 63(2), 256-270.

Rothman, L., Howard, A., Buliung, R., Richmond, S.A., Macarthur, C., Macpherson, A., 2016. 72 dangerous student passenger drop-off, pedestrian behaviours and the built environment near schools. *Injury Prevention* 22(2), A28-A28.

Stoker, P., Garfinkel-Castro, A., Khayesi, M., Odero, W., Mwangi, M.N., Peden, M., Ewing, R., 2015. Pedestrian safety and the built environment: A review of the risk factors. *CPL Bibliography* 30(4), 377-392.

Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J., 2012. The role of built environment on pedestrian crash frequency. *Safety Science* 50(4), 1141-1151.

Varin, C., Vidoni, P., 2005. A note on composite likelihood inference and model selection. *Biometrika* 92(3), 519-528.

Vaughn, M.G., Define, R.S., DeLisi, M., Perron, B.E., Beaver, K.M., Fu, Q., Howard, M.O., 2011. Sociodemographic, behavioral, and substance use correlates of reckless driving in the United States: Findings from a national sample. *Journal of Psychiatric Research* 45(3), 347-353.

Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis & Prevention* 43(6), 1979-1990.

Weinstein-Agarwal, A., Schlossberg, M., Irvin, K., 2008. How far, by which route and why? A spatial analysis of pedestrian preference. *Journal of Urban Design* 13(1), 81-98.

Wier, M., Weintraub, J., Humphreys, E.H., Seto, E., Bhatia, R., 2009. An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accident Analysis & Prevention* 41(1), 137-145.

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention* 75, 16-25.

Yu, C.Y., Zhu, X., 2016. Planning for safe schools impacts of school siting and surrounding environments on traffic safety. *Journal of Planning Education and Research* 36(4), 476-486.

Zhao, Y., Joe, H., 2005. Composite likelihood estimation in multivariate data analysis. *The Canadian Journal of Statistics* 33(3), 335-356.

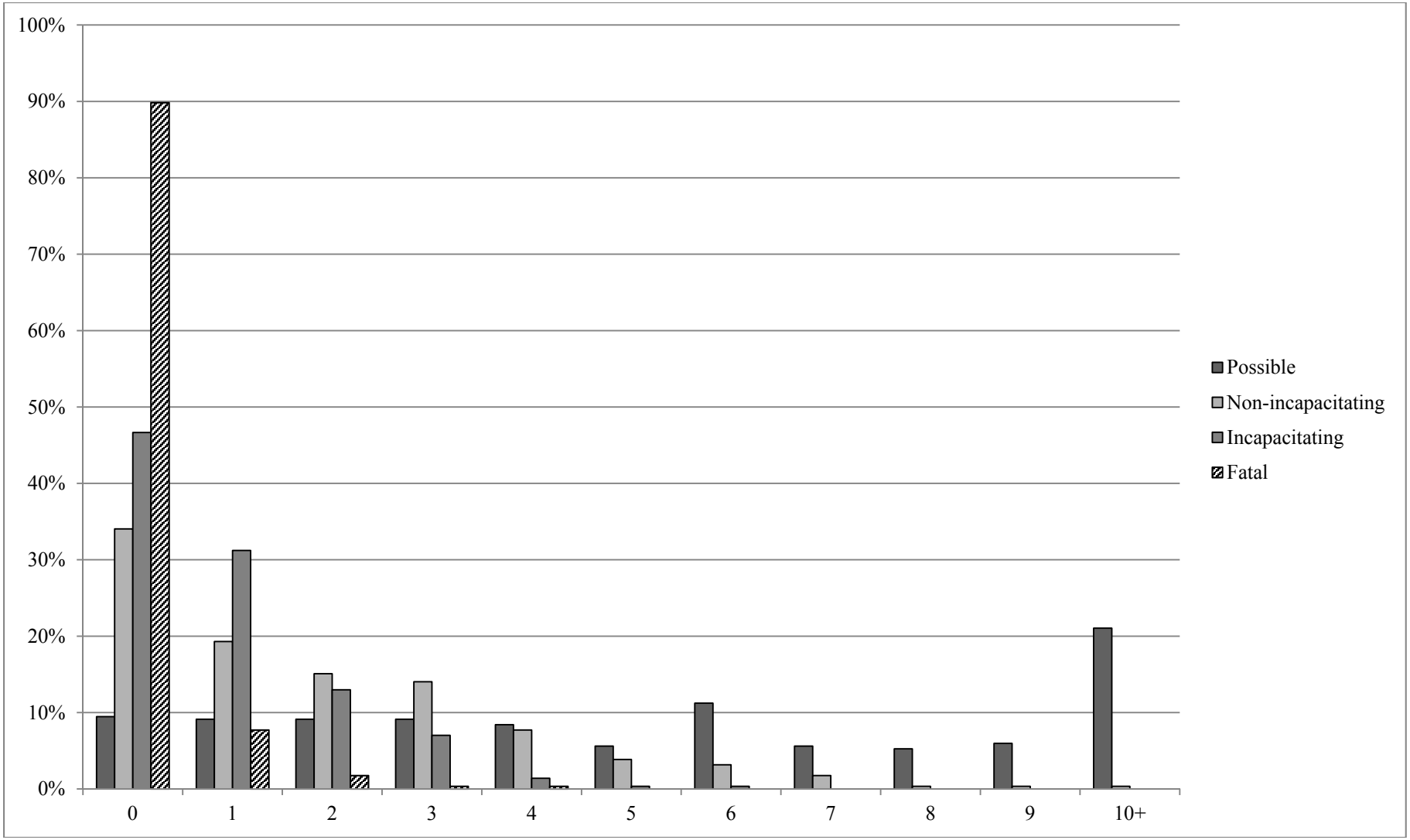

Figure 1: Distribution of Percentage of Census Tracts Associated with Each Count of Pedestrian Injuries by Severity Level

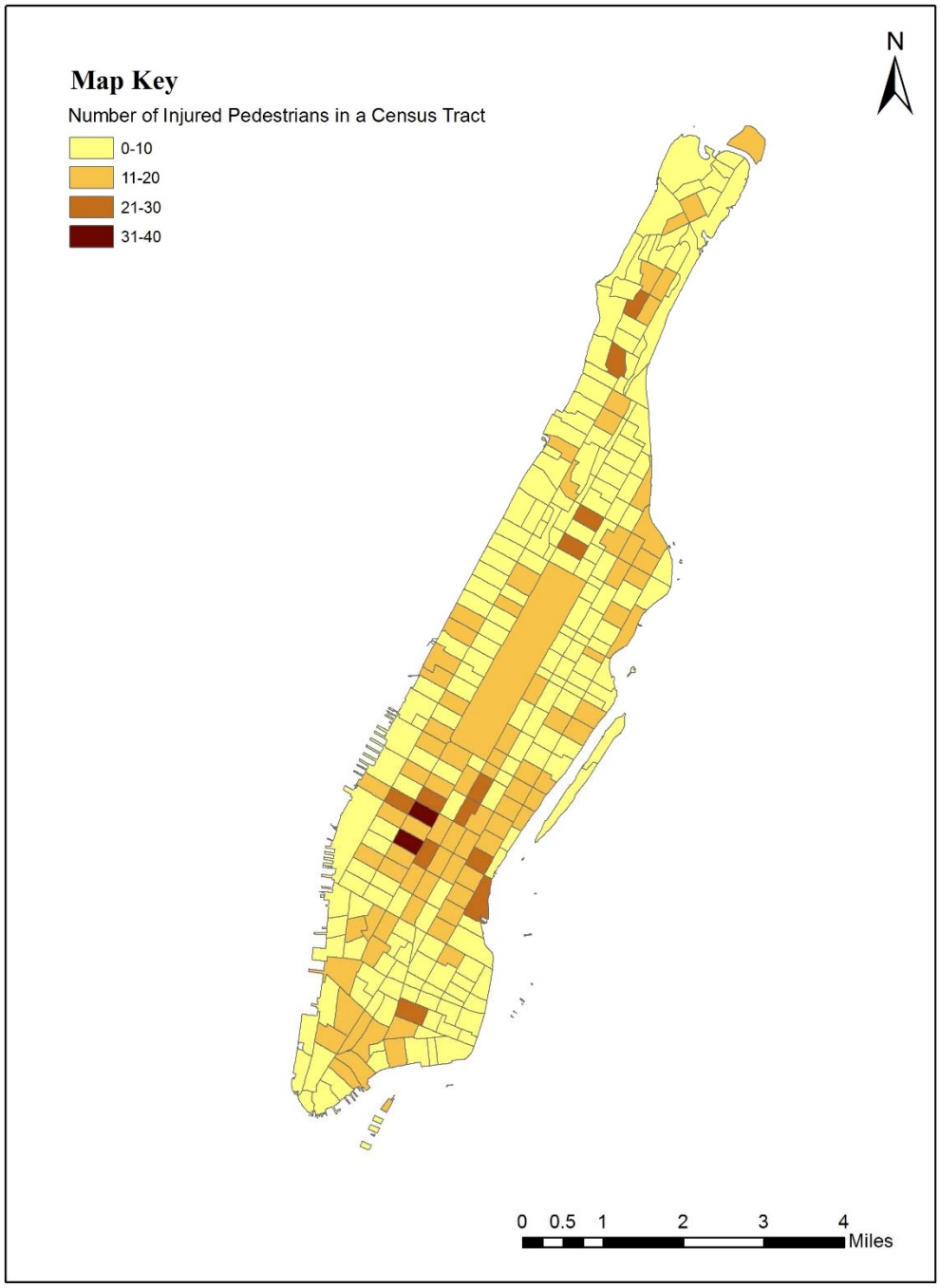

Figure 2: Thematic Map of Pedestrian Injuries in Manhattan by Census Tract

Table 1: Effects of Variables on Long-Term Risk Propensity

Injury Severity	Possible		Non-Incapacitating		Incapacitating		Fatal	
Parameters	**Estimate**	**T-Stat**	**Estimate**	**T-Stat**	**Estimate**	**T-Stat**	**Estimate**	**T-Stat**
Long term risk propensity								
Socio-Demographic Variables								
Log of Population Density								
Mean	0.550	11.23	0.650	11.49	0.058	14.35	0.287	13.21
Standard Deviation	0.217	4.34	0.398	2.43				
Proportion of population <=19 years								
Mean					-1.378	-8.81		
Standard Deviation								
Proportion of adults (>=18 years) without high school degree								
Mean					0.774	15.57		
Standard Deviation								
Proportion of Hispanic population								
Mean			-0.170	-19.73	0.175	7.44	1.249	13.33
Standard Deviation			1.247	3.21	1.085	6.51	0.855	5.24
Median household income (in $100,000)								
Mean							-1.090	-6.00
Standard Deviation								
Land-use and Road Network Variables								
Proportion of commercial land-use								
Mean					0.670	16.23		
Standard Deviation								
Proportion of residential land-use								
Mean					0.751	16.29		
Standard Deviation								
Proportion of local neighborhood roads and city streets								
Mean			-0.764	-13.46			-0.254	-7.39
Standard Deviation							0.545	3.25
Activity Intensity Variables								
Intensity of retail activity								
Mean	0.732	5.45	0.674	3.84	0.325	8.80		
Standard Deviation			0.445	3.62				
Number of schools								
Mean			0.273	11.65	-1.802	-10.81		
Standard Deviation					0.219	2.80		
Number of universities								
Mean			0.096	2.90				
Standard Deviation			0.282	5.97				
Commute Mode Shares and Transit Supply								
Walk commute mode share								
Mean							2.396	6.55
Standard Deviation								
Number of subway stops								
Mean	0.181	6.80						
Standard Deviation	0.501	3.80						
Non-diagonal elements of Ω_j								
Between the constant and the number of subway stops effect	0.128	2.16						
Between the effects of (log) population density and intensity of retail activity			0.175	2.44				

Table 2: Threshold Elements and Spatial Dependence Effects

Injury Severity	Possible		Non-Incapacitating		Incapacitating		Fatal	
Parameters	**Estimate**	**T-Stat**	**Estimate**	**T-Stat**	**Estimate**	**T-Stat**	**Estimate**	**T-Stat**
Threshold Parameters								
Threshold specific constants								
α_4	-0.103	-4.38						
α_5	-0.272	-3.29						
α_{11}	-0.298	-2.97						
γ_j ***vector for each j***								
Constant	0.546	5.28	1.321	2.56	-0.573	-4.83	-4.928	-5.07
Socio-Demographic Variables								
Median household income (in $100,000)							2.682	7.83
Land-use and Road Network Variables								
Proportion of commercial land-use	0.132	8.12	0.546	5.90				
Proportion of residential land-use	0.082	1.99						
Proportion of secondary roads	-2.771	-6.34						
Proportion of local neighborhood roads and city streets	-0.969	-7.22						
Proportion of minor roads	-1.599	-6.89						
Activity Intensity Variable								
Number of schools	1.469	9.22	0.164	3.22	2.006	6.67		
Commute Mode Shares and Transit Supply								
Walk commute mode share	0.295	2.30	0.775	2.98				
Spatial Dependence Effects								
Auto-Regressive Parameter	0.221	2.18	0.428	5.11	0.629	2.54	0.794	3.27

Table 3: Measures of Fit

Summary Statistic		IAC	ICSC	IUSC	MAC	MCSC	MUSC
Number of Observations		285 tracts					
Composite log-likelihood (CLL) at convergence of the naïve model		-2,962,001.29					
Number of parameters		51	52	55	58	59	62
Composite log-likelihood (CLL) at convergence		-1,466,637.90	-1,466,291.95	-1,465,930.95	-1,466,497.66	-1,466,150.74	-1,465,793.77
Adjusted composite likelihood ratio test (ADCLRT) between MUSC and the corresponding model		1,688.3 > Chi-Squared statistics with 11 degrees of freedom at any reasonable level of significance	996.4 > Chi-Squared statistics with 10 degrees of freedom at any reasonable level of significance	274.4> Chi-Squared statistics with 7 degrees of freedom at any reasonable level of significance	1,407.8 > Chi-Squared statistics with 4 degrees of freedom at any reasonable level of significance	713.9 > Chi-Squared statistics with 3 degrees of freedom at any reasonable level of significance	Not Applicable
Spatial correlation							
Possible injury		0.0 (fixed)	0.495 (3.98)	0.168 (2.23)	0.0 (fixed)	0.511 (4.25)	0.221 (2.18)
Non-incapacitating injury		0.0 (fixed)	0.495 (3.98)	0.430 (5.73)	0.0 (fixed)	0.511 (4.25)	0.428 (5.11)
Incapacitating injury		0.0 (fixed)	0.495 (3.98)	0.503 (3.12)	0.0 (fixed)	0.511 (4.25)	0.629 (2.54)
Fatal injury		0.0 (fixed)	0.495 (3.98)	0.731 (4.28)	0.0 (fixed)	0.511 (4.25)	0.794 (3.27)
Injury severity level	**Actual count**	**Predicted count**					
Possible injury	1,700	1,930.1	1,853.4	1,837.3	1,922.3	1,847.2	1,827.0
Non-incapacitating	523	579.2	562.2	550.3	578.1	559.6	542.4
Incapacitating	250	287.1	270.1	269.6	279.5	273.0	267.2
Fatal	39	52.0	48.3	44.8	50.7	46.9	43.7
Mean Absolute Percentage Error (MAPE)		13.01%	9.92%	7.77%	12.99%	8.99%	6.92%

Table 4: Aggregate-Level Elasticity Effects for the Two Most Severe Injury Levels (t-stats in parenthesis)

Variable	Multivariate Aspatial Count (MAC) Model		Multivariate Constrained Spatial Count (MCSC) Model		Multivariate Unconstrained Spatial Count (MUSC) Model	
	Incapac.	Fatal	Incapac.	Fatal	Incapac.	Fatal
Socio-Demographic Variables						
Log of Population Density	5.7 (1.84)	25.7 (5.15)	6.5 (1.81)	33.0 (5.29)	5.8 (1.86)	44.3 (6.70)
Proportion of population ≤19 years	-6.8 (-6.72)	0.0	-7.8 (-7.24)	0.0	-10.7 (-7.30)	0.0
Proportion of population ≥18 years without high school degree	8.2 (8.07)	0.0	10.8 (8.31)	0.0	15.3 (9.10)	0.0
Proportion of Hispanic population	9.8 (5.20)	21.8 (7.10)	13.9 (5.37)	33.8 (6.91)	17.6 (5.35)	46.1 (7.23)
Median household income (in $100,000)	0.0	-40.3 (-9.32)	0.0	-48.9 (-9.81)	0.0	-57.7 (-10.53)
Land-use and Road Network Variables						
Proportion of commercial land-use	10.9 (5.17)	0.0	12.9 (5.27)	0.0	19.1 (6.01)	0.0
Proportion of residential land-use	12.1 (5.25)	0.0	15.1 (6.05)	0.0	18.2 (6.26)	0.0
Proportion of local neighborhood roads and city streets	0.0	-6.0 (-7.28)	0.0	-9.2 (-7.22)	0.0	-11.4 (-6.79)
Activity Intensity Variables						
Intensity of retail activity	16.6 (11.24)	0.0	19.2 (11.81)	0.0	21.3 (12.24)	0.0
Number of schools	5.6 (3.27)	0.0	5.9 (3.10)	0.0	6.5 (3.23)	0.0
Commute Mode Shares and Transit Supply						
Walk commute mode share	0.0	30.2 (9.65)	0.0	34.1 (10.91)	0.0	39.7 (9.58)

APPENDIX A: Procedure to Predict the Expected Count Values for each Census Tract

The expected value of injury count in census tract q for injury severity level j may be written as:

$$E(y_{qj}) = \sum_{k=0}^{\infty} P(y_{qj} = k) \cdot k, \tag{A.1}$$

where $P(y_{qj} = k)$ is the probability of occurrence of k injuries at injury severity level j in census tract q. Although the summation in the equation above extends until infinity in our count model, we consider counts only up to k = 25 in our prediction procedure (this value represents the maximum count of injuries across census tracts and across injury severity levels in the estimation sample, corresponding to the possible injury severity level for pedestrian injuries). Beyond the count value of 25, the probabilities are very close to zero and hence do not have any significant impact on the predicted value. The expected value in Equation (A.1) for a tract is a function of the exogenous variables for all Q census tracts for the spatial models (but only the exogenous variables for that tract for the aspatial models), as well as a function of the variable vector $\boldsymbol{z}_q$ embedded in the thresholds in Equation (2).

For all the models, we adopt the same approach to develop an estimate of $P(y_{qsj} = k)$, even though, for some simpler models, a more customized version may be used. Specifically, we simulate the $QJ \times 1$ – vector $\boldsymbol{y}^*$ (as $\boldsymbol{y}^* \sim MVN_{QJ}(\boldsymbol{B}, \boldsymbol{\Psi})$, where $\boldsymbol{B} = \mathbf{C}\breve{\boldsymbol{x}}\boldsymbol{b}$ and $\boldsymbol{\Psi} = \mathbf{C}\breve{\boldsymbol{x}}\left[\boldsymbol{IDEN}_Q \otimes \boldsymbol{\Delta}\right]\breve{\boldsymbol{x}}'\mathbf{C}'$) five hundred times using the estimated values of $\breve{\boldsymbol{\delta}}$, $\boldsymbol{b}$, and the covariance matrix $\boldsymbol{\Delta}$ of $\breve{\boldsymbol{\beta}}$. Subsequently, we compare each of the 500 draws of the q^{th} element of $\boldsymbol{y}^*$ with the corresponding thresholds (estimated from Equation (2), after obtaining estimates of the $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ vectors), and assign the count value for each of the 500 draws based on this comparison. The share of each count prediction is taken across the 500 draws to estimate $P(y_{qj} = k)$.[13]

[13] The predictions were not sensitive to the number of draws beyond about 400 draws, and so we settled on 500 draws.