The Thesis Committee for Sriram Narayanamoorthy

Certifies that this is the approved version of the following thesis:

On Accommodating Spatial Dependence in Bicycle and Pedestrian Injury Counts by Severity Level

APPROVED BY

SUPERVISING COMMITTEE:

Supervisor: _____

Chandra R. Bhat

_____

Stephen D. Boyles

On Accommodating Spatial Dependence in Bicycle and Pedestrian Injury Counts by Severity Level

by

Sriram Narayanamoorthy, B.Tech

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

December 2012

In loving memory of my grandmother

N. N. Rugmani

Acknowledgments

First and foremost, I must express appreciation to Dr. Chandra Bhat, with whom it has been a pleasure and a privilege to work. I thank him for his insights and advice as also for the stimulus and motivation to dig deep into research. I am grateful to him for showing me how to make sense of mathematical models and for instilling in me the notion that writing should be simple, direct, logical, and tight. His invaluable ideas and the high standards that he sets for himself and his students form the backbone of this thesis.

Thanks also to Dr. Stephen Boyles - not only for the valuable inputs on this thesis but also for the highly engaging and immensely useful courses. Thank you to the faculty at UT Austin who have made my stay here as a graduate student a memorable and intellectually satisfying experience. Special thanks to Dr. Jason Abrevaya, Dr. David Kendrick and Dr. Stephen Donald - the way they orchestrate classroom learning of Economics and Econometrics is pedagogy at its finest. Thanks to all the administrative staff, especially Lisa Macias who provides immeasurable assistance to the students in her care.

To all my friends and colleagues at UT Austin - I have relished the inimitable sense of camaraderie and fellowship. In particular, Rajesh for the delightful conversations, intense debates and for the inspiration as a friend and a mentor. Jassu, for his culinary skills. Chrissy, for all the conversations and the fun. Ravi, for providing endless entertainment and much needed distractions. Palvinder and Subodh, who kept me company during those long hours in basement. Tarun for his refreshing approach to research. Raghu, for all his help and the long football discussions. Marisol, for the flexibility she afforded me on the projects. Gaurav, for the debugging sessions over the web. Moby, for many a rides around Austin. Prasad, for his assistance in finding a fourth roommate on four different occasions and for sharing with me his hilarious

adventures. Thanks much to Chinmoy, Rudra, Harish, Yoganand, Ranjith and Aria.

Finally, thanks to my amazing family, who has been extremely supportive of me throughout my life. My deepest appreciation for my parents and my sister who have been unwavering in their love and encouragement and I owe all of my success and achievements to them.

This thesis is dedicated to the memory my grandmother who was my closest confidant.

On Accommodating Spatial Dependence in Bicycle and Pedestrian Injury Counts by Severity Level

Sriram Narayanamoorthy, MSE

The University of Texas at Austin, 2012

Supervisor: Chandra R. Bhat

This thesis proposes a new spatial multivariate count model to jointly analyze the traffic crash-related counts of pedestrians and bicyclists by injury severity. The modeling framework is applied to predict injury counts at a Census tract level, based on crash data from Manhattan, New York. The results highlight the need to use a multivariate modeling system for the analysis of injury counts by road-user type and injury severity level, while also accommodating spatial dependence effects in injury counts.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The continued dependence of individuals on motorized automobiles for transportation, along with rapid population growth, has led to increasing traffic congestion in most urban areas in the U.S. (see Schrank *et al.*, 2011). While several strategies are being considered to alleviate the increasing urban traffic congestion, many metropolitan planning organizations (MPOs) have started to invest in non-motorized mode infrastructure to promote the use of walking and bicycling modes (Pucher *et al.*, 1999, Metropolitan Transportation Commission, 2009, Southern California Association of Governments, 2012). In addition to reducing traffic congestion, the promotion of these transportation modes can also offer ancillary benefits to society in terms of improved health, better air quality, energy independence, and enhanced quality of life (see Pucher *et al.*, 2010 and Gotschi and Mills, 2008). However, even as MPOs look to the promotion of non-motorized modes of travel, it is illustrative to note that, according to the 2009 National Household Travel Survey (NHTS), non-motorized modes accounted for only 11.9% of all weekday trips, and 0.9% of total weekday person travel mileage. On the other hand, many cities in Europe and other nations boast substantially higher non-motorized shares in terms of trips and mileage (Bassett *et al.*, 2008).

The higher non-motorized mode shares in Europe and other nations may be attributable to many factors, including higher built environment density, expensive gas and auto ownership costs, and better land-use mix. But another important factor in travel mode choice decisions is safety from traffic crashes. In fact, studies have now established that safety from traffic crashes is a key determinant of a person's mode choice decision (see

Winters *et al.*, 2010 and Sener *et al.*, 2009). In this context, Beck *et al.* (2007) have found that, relative to passenger vehicle occupants, bicyclists and pedestrians in the U.S. are 2.3 and 1.5 times, respectively, more likely to be fatally injured on a given trip. In cross-country comparisons, Pucher and Dijkstra (2003) found that, after controlling for travel exposure in terms of mileage, U.S. pedestrians (bicyclists) are about 3 times (2 times) as likely to get killed in traffic accidents as German pedestrians (bicyclists) and over 6 times (3 times) as likely to be killed as Dutch pedestrians (bicyclists). In another more recent study at a metropolitan area level (rather than a national level that can mask risk variation within countries), McAndrews (2011) observed that the risk of a fatal traffic crash injury for pedestrians in San Francisco is 4.1 times higher than for pedestrians in Stockholm, while the corresponding figure is 1.7 for bicyclists. Overall, these studies clearly reveal the under-performance of the U.S. in terms of pedestrian and bicyclist safety relative to other advanced economies. At an absolute level, about 4280 pedestrians and 618 bicyclists were killed in traffic accidents in the year 2010 in the U.S., constituting 15% of all fatalities that year (National Highway Traffic Safety Administration or NHTSA, 2012) while non-motorized mode mileage comprises only 0.9% of total travel mileage.

To summarize, the promotion of non-motorized modes of transportation should involve, as one essential element, an understanding of the risk factors associated with pedestrians and bicyclist-related injuries. This can allow the identification of high risk crash environmental settings and inform the design of appropriate transportation policy countermeasures. Accordingly, there have been several efforts in the past that focus on modeling the frequency of non-motorized crashes as a function of relevant built environment and socio-economic indicators. In this research, we contribute to this literature by formulating a multivariate model to jointly analyze, at a "neighborhood" level, the count of pedestrians and bicyclists involved in traffic crashes by injury severity sustained. While a "neighborhood" may be defined variously as a roadway

2

street segment (see Kim *et al.*, 2007), or an intersection (Carter and Council, 2007), or a geographic area (a Census block, or a Census block group, or a Census tract; see, for example, Wier *et al.*, 2009 and Siddiqui *et al.*, 2012), the spatial unit we use to characterize a "neighborhood" is the Census tract. We do so because the more disaggregate spatial units (roadway street segment, intersection, Census block, and Census block group) can routinely experience zero pedestrian and bicyclist-related crashes for multiple years at a stretch, which reduces the variability of the count variables across such disaggregate spatial units and decreases our ability to tease out the risk factors associated with pedestrian and bicyclist crash involvement. The use of the more aggregate Census tract level avoids these problems, while also representing a reasonably homogeneous 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.[1]

Two important issues are of significance in the current research. First, the reason for our emphasis on the count of pedestrians and bicyclists injured by *severity level* is to acknowledge that accident costs vary substantially by severity level (see Wang *et al.*, 2011). 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 and bicyclist injuries. Second, the multivariate model proposed in this study recognizes many econometric issues at once: (a) It acknowledges the count nature of the number of injuries, (b) it conveniently addresses excess zeros (or any other excess count value

---

[1]Note also that the count variable used in our model corresponds to the number of pedestrian and bicyclist injuries by 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 or bicyclist 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.

for that matter) within a multivariate count setting, (c) it accommodates the potential presence of unobserved Census tract factors that can lead to dependence, within the Census tract, in the risk propensities for the different road-user type-injury severity combinations (road-user, in our analysis, may be pedestrians or bicyclists), and (d) it considers spatial dependence effects across Census tracts that are likely to be present because of the spatial nature of the analysis.

The rest of this thesis is structured as follows. Chapter 2 presents an overview of the relevant earlier literature and positions the current study. Chapter 3 presents the model structure and estimation procedure. Chapter 4 describes the study area, data source and important sample characteristics. Chapter 5 presents the empirical estimation results and their implications for reducing non-motorized user injury severity in crashes. Finally, Chapter 6 concludes the thesis.

Chapter 2

Literature Review

Several methodological challenges arise when modeling crash frequency-related data (see Lord and Mannering, 2010 for a good review). The focus of the current study is on addressing two specific methodological challenges that lead to the proposed spatial multivariate count model.

2.1 Modeling Count Data by Type

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 and bicyclists involved in traffic crashes by injury severity level. This leads to a multivariate count system within each Census tract because of the presence of unobserved Census tract factors that (1) influence the risk propensity for a specific injury severity level across both pedestrian and bicyclist injuries (for instance, motorists within a certain Census tract may have an unaccommodating attitude toward sharing the road with non-motorists, which may increase the risk of fatal injuries for both pedestrians and bicyclists - for future reference, we will label such unobserved factors as type a unobserved factors), (2) intrinsically increase or decrease the propensity for pedestrian injuries across all 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), (3) intrinsically increase or decrease the propensity for bicyclist injuries across all injury levels (for example, discontinuous bicycle paths in a Census tract may lead to a generic increase in risk propensity for bicyclists that permeates across

all injury levels; we will label the unobserved factors corresponding to (2) and (3) as type b unobserved factors), and (4) impact the overall propensity of non-motorized injuries (for instance, because of a generally high propensity to use non-motorized modes in a Census tract; we will label the unobserved factors corresponding to (4) as type c unobserved factors).

There have been several efforts in the literature to formulate and estimate multivariate count data models. *One common approach* has been to use multivariate versions of the Poisson or negative binomial discrete distributions (see, for example, Ladrón de Guevara *et al.*, 2004, Buck *et al.*, 2009, and Bermúdez and Karlis, 2011 for applications of these methods). Such multivariate count models have the advantage of a closed form, but they become cumbersome as the number of correlated counts increases and they also can accommodate only a positive correlation in the counts. *A second common approach* is to use a mixing structure, in which one or more (typically) normally distributed random terms are introduced in the parameterization of the expected value of the discrete distribution (so that the expected value is not only a function of exogenous variables, but also includes one or more additive random terms within the exponentiation). If the same error term enters in the means of multiple count variables, this generates correlation (see Chib and Winkelmann, 2001, Lee *et al.*, 2006, Park and Lord, 2007, Aguero-Valverde and Jovanis, 2009, and El-Basyouny and Sayed, 2009 for examples of such an approach). A similar, but slightly different mixing approach, has been used recently by Chiou and Fu (2012), who developed a multinomial-generalized Poisson model for the joint analysis of crash frequency and injury severity. Essentially, they use a generalized Poisson distribution for the marginal probability of total counts, and a conditional multinomial distribution for the conditional frequency of each severity level. The total count and injury severity level sub-models are then stitched together by allowing common unobserved components in a manner similar to the mixing approach for different counts. The advantage of all these mixing approaches is that they permit

both positive and negative dependency between the counts by severity type. However, it is difficult in these mixing approaches to account for excess zeros through use of techniques such as zero inflation and hurdle count models (see Lee *et al.* 2006, Alfò and Maruotti, 2010, Herriges *et al.*, 2008). Furthermore, these mixing approaches require rather cumbersome and time consuming simulation approaches for estimation (see Müller and Czado, 2005, Aguero-Valverde and Jovanis, 2006, and Ver Hoef and Jansen, 2007 for discussions). A *third approach* is to analyze crash rates (for example, number of crashes per 100 million vehicle miles of travel) by injury severity level, which translates the dependent variable vector from a multivariate count to a multivariate continuous variable. To address the preponderance of zero values, Anastasopoulos *et al.* (2012) developed a multivariate Tobit-regression model to analyze crash rates by injury severity level. However, the likelihood estimation approach again becomes cumbersome and presents a computational challenge when there are many tobit regressions in the multivariate set-up, such as when focusing on counts by injury severity as well as road-user type (pedestrian versus bicyclist injury).

Another important point is that the approaches discussed above to accommodate multivariate counts or multivariate crash rates are already so computationally difficult that extending the approaches to accommodate spatial dependency structures becomes impractical, if not literally infeasible. This is evident from the fact that none of the multivariate models of counts discussed above accommodate spatial dependence.

2.2   Spatial Dependency Effects

Spatial dependency is important to recognize because of the mapping of crash locations to spatial units of analysis, such as Census tracts in the current study. In particular, observed factors at a particular Census tract, such as population density (a proxy measure of non-motorized travel exposure) or retail intensity, may not only have an impact on non-motorized injury risk at that Census tract, but may also have a spatial

spillover effect on non-motorized risk propensity in proximally located Census tracts. Similarly, there may be common unobserved (to the analyst) location factors such as discontinuous pedestrian and bicycle paths that can generate a "spatial correlation" effect in the stochastic terms of risk propensity at proximally located Census tracts (Lord and Mannering, 2010). Ignoring such spatial dependencies will, in general, result in inconsistent and inefficient parameter estimation in non-linear models (see LeSage and Pace, 2009 and Franzese and Hays, 2008). On a substantive note, such inconsistent estimates can lead to misinformed policy actions and countermeasures, and a sub-optimal allocation of scarce resources, to reduce the frequency and severity of traffic crash-attributable pedestrian and bicyclist injuries.

In the spatial analysis literature, the two workhorse specifications to capture spatial dependencies are the spatial lag and the spatial error specifications (Anselin, 1988). The spatial lag specification allows spatial dependence through both spatial spillover effects as well as spatial error correlation effects. The spatial error specification, on the other hand, assumes that spatial dependence is only due to spatial error correlation effects and not due to spatial spillover effects. The spatial error specification is somewhat simpler in formulation and estimation than the spatial lag model. While these spatial specifications have been used primarily in the case of a continuous dependent variable, the past decade has seen increasing use of these spatial specifications for non-linear discrete choice models. The specifications are similar to the linear models, except that they are now applied to the latent continuous propensity variables underlying the observed discrete variable. However, the spatial lag and spatial error specifications saw little use in the context of count models until Castro, Paleti, and Bhat (2012) (CPB for short in the rest of this thesis), who showed that even count models can be recast in the form of an underlying latent continuous variable framework (so that the spatial specifications can again be applied to the latent continuous propensity variables). Before CPB, a common approach was to map the count variable into

an approximate continuous variable (typically also applying a log-transformation to ensure positive predictions, and sometimes also normalizing by an exposure measure to obtain crash rates or taking ratios of different types of crashes), and then apply well-established estimation methods developed for continuous models. Examples of such efforts in the safety literature include LaScala *et al.* (2000), Quddus (2008), Ha and Thill (2011), and Delmelle *et al.* (2011). While useful, these efforts may be viewed as approximations, since they generate "continuous" variables from underlying count data. Especially as the focus shifts from modeling total crashes to total crashes by injury severity type and/or road-user type, the count data will show less variation (and a preponderance of zero counts), rendering the approximation in the translation to a continuous variable more inappropriate. It is therefore, no surprise, that none of the studies listed above that use this "continuous" transformation method consider crashes by type, instead focusing on total crashes. Another alternative approach to incorporate spatial dependency in count models in the past has been to use a conditional autoregressive (CAR) or 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 estimated using Bayesian hierarchical methods. Examples of such efforts include Miaou and Song, 2005, Aguero-Valverde and Jovanis, 2006, 2010, Mitra, 2009, Wang *et al.*, 2011, Siddiqui *et al.*, 2012). Unfortunately, this approach (which is essentially a mixing approach of the type discussed in the previous section, except with the mixing undertaken over space) can be difficult as the number of spatial units increases, and extending the approach to modeling crashes by type is extremely challenging (if not impractical). This is because the Bayesian MCMC methods used in the approach are particularly cumbersome for high dimensional problems, owing to the non-standard form of the conditional posterior distributions of model unknowns in the count case. As a result, obtaining realizations from the joint posterior distribution requires the use of the combination of numerical optimization methods combined

with time-consuming and multiple Metropolis-Hastings steps (see Herriges *et al.*, 2008 and Chib and Winkelmann, 2001). Besides, this approach considers spatial error correlation effects, but not spatial spillover effects.

Overall, earlier efforts in the safety literature that incorporate spatial dependence focus almost exclusively on total counts (Miaou and Song, 2005 is the only exception). Further, all of these studies (except LaScala *et al.*, 2000) adopt a spatial error structure, because estimating a spatial lag dependency structure is not straightforward in the methods used (LaScala *et al.* simply treat crash counts as a continuous variable to avoid methodological complexities associated with modeling spatial dependencies using non-linear count modeling frameworks). But spatial lag dependency captures, in an efficient and intuitive manner, both spatial spillover and spatial error correlation effects, and is much easier to justify when accommodating spatial dependency (see McMillen, 2010 and Sidharthan and Bhat, 2012 for discussions).

2.3 The Current Research

In the current study, we recognize and retain the count nature of the number of pedestrian and bicyclist injuries by injury severity level. In doing so, we address the multivariate nature of the counts within a Census tract. In addition, we also simultaneously recognize spatial lag dependency effects across Census tracts. To our knowledge, this study is the first to develop such a spatial multivariate count model in the literature. The approach we use is based on recasting the basic count model as a special case of a generalized ordered-response (GOR) model, as proposed by CPB. Then, the spatial lag specification is imposed on the underlying latent variable determining the counts. The GOR recasting also immediately enables the accommodation of excess zeros (or excess counts of any value) in a straightforward manner within the spatial multivariate count setting. 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 CML estimator is consistent and asymptotically normally distributed, under the usual regularity conditions that make the maximum likelihood estimator consistent and asymptotically normally distributed.

The proposed model is applied to examine, at the spatial level of a Census tract, the number of pedestrian and bicyclist injuries by injury severity level. The data for the analysis is drawn from a bicyclist and pedestrian crash database maintained by New York City (see Chapter 4 for details on how the data was assembled). Several groups of Census tract-based risk factors are considered in our analysis based on earlier research, including (1) socio-demographic characteristics (such as population density, proportions of the population by age, income, and race), (2) land-use and road network characteristics (such as proportion of retail and commercial land-use, 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).

Chapter 3

Methodology

3.1 Model Formulation

Let q ($q = 1, 2,\ldots, Q$), j ($j = 1, 2,\ldots, J$), and s ($s = 1, 2,\ldots, S$) be indices for observation units (Census tracts in our analysis), type of non-motorized user injured (pedestrian or bicyclist), and injury severity level sustained by the non-motorized road-user, respectively, where Q is the total number of observation units in the sample, J is the total number of types of non-motorist road-users ($J{=}2$ in our empirical analysis, with $j{=}1$ representing pedestrians and $j{=}2$ representing bicyclists), and S is the number of injury severity levels.[1] Let m_{qjs} be the observed count of road-users of type j injured at severity level s within the qth observational unit over a predefined time period (we considered a time period of one year for the empirical analysis in this research; note also that m_{qjs} may take a value in the range from 0 to ∞). Next define a latent risk propensity for injury at severity level s for road-user type j in observation unit q as y^*_{qjs}. Then, consider the following structure for y^*_{qjs} in the GOR representation for count models (see CPB):

$$y^*_{qjs} = \delta \sum_{q'=1}^{Q} w_{qq'} y^*_{q'js} + \boldsymbol{b}'_{js}\boldsymbol{x}_q + \omega_{qs} + u_{qj} + v_q + \varepsilon_{qjs}$$

$$y_{qjs} = m_{qjs} \text{ if } \Psi_{qjs,m_{qjs}-1} < y^*_{qjs} < \Psi_{qjs,m_{qjs}},$$

(3.1)

Where $w_{qq'}$ is the usual distance-based spatial weight corresponding to spatial units

---

[1]The number of severity levels may vary across different non-motorized road-user types. However, for notation simplicity, we assume the same number of severity levels across both pedestrian and bicyclists.

q and q' (with $w_{qq} = 0$ and $\sum_{q'} w_{qq'} = 1$) for each (and all) q, $\delta(0 < \delta < 1)$ is the spatial autoregressive parameter, $\boldsymbol{x}_q$ is a (K x 1) column vector of exogenous variables (excluding a constant), and $\boldsymbol{b}_{js}$ is a corresponding (K x 1) column vector capturing the effects of the exogenous vector $\boldsymbol{x}_q$ on the latent risk propensity y_{qjs}^*.[2] The error terms in Equation (3.1) are as follows: (1) the ω_{qs} captures unobserved spatial unit-specific factors that affect the propensity of injury of severity level s for all road-users (bicyclists and pedestrians; these are the type a factors discussed in Section 2.1); ω_{qs} is assumed to be a realization from a univariate normal distribution with variance π_s^2, (2) the u_{qj} term captures unobserved spatial unit-specific factors that impact the propensity of injury for road-user type j (corresponding to the type b factors in Section 2.1); u_{qj} is assumed to be a realization from a univariate normal distribution with variance τ_j^2, (3) the ν_q term captures unobserved factors specific to spatial unit q that impact the overall propensity of non-motorized injuries (corresponding to the type c factors in Section 2.1); ν_q is assumed to be a realization from a univariate normal distribution with variance σ^2, and (4) the ε_{qjs} captures unobserved spatial unit-specific factors that influence the propensity of injuries of type s for road-user type j; this term is assumed to be independent and identically standard normal distributed across road-user types, severity levels, and spatial units.[3]

The thresholds in Equation (3.1) takes the form:

$$\Psi_{qjs,m_{qjs}} = \Phi^{-1}\left(e^{-\lambda_{qjs}} \sum_{l=0}^{m_{qjs}} \frac{\lambda_{qjs}^l}{l!}\right) + \alpha_{js,m_{qjs}}, \; \lambda_{qjs} = e^{\gamma'_{js}z_q} \tag{3.2}$$

$$\alpha_0 = 0, \; \alpha_{js,m_{qjs}} = \alpha_{js,L_{js}} \; \text{if} \; m_{qjs} > L_{js},$$

where Φ^{-1} is the inverse function of the univariate cumulative standard normal,

---

[2]Some explanatory variables may not be important for specific road-user and/or severity levels. This situation is accommodated within our notation system by letting the corresponding elements in the vector $\boldsymbol{b}_{js}$ be equal to zero.

[3]The scale of the ε_{qjs} term is normalized to one for identification.

13

$\Psi_{qjs,-1} = -\infty \; \forall \; q$, j and s. (this restriction is needed for identification, given the parameterization of the thresholds; see CPB), z_q is a vector of exogenous variables (including a constant) associated with observation unit q (there can be common variables in z_q and x_q), γ_{js} is a corresponding coefficient vector to be estimated for road-user type j and severity level s, and L_{js} 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 $\alpha_{js,m_{qjs}}$ should be such that the ordering condition on the thresholds $(-\infty < \Psi_{js,0} < \Psi_{js,1} < \Psi_{js,2} < \dots)$ is satisfied. The presence of the $\alpha_{js,m_{qjs}}$ term provides flexibility to accommodate high or low probability masses for specific count outcomes without the need for cumbersome treatment using hurdle or zero-inflated mechanisms.

The GOR framework for count models, as just discussed, not only provides important computational benefits to accommodate statistical, econometric, and spatial considerations, but may also be motivated from an intuitive standpoint for count data in a manner similar to that for ordinal data. For example, in our empirical context, consider the count of pedestrian fatalities (the following discussion is applicable to all road-user type-injury severity level combinations, but we focus on pedestrian fatalities 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 propensity y_{q14}^* of a pedestrian (j=1) in Census tract q being involved in a crash leading to death (s=4, since the pedestrian injury severity categories in our empirical analysis are "possible" injury, "non-incapacitating injury", "incapacitating injury", and "fatal" injury). This "long-term" propensity may be impacted by such Census tract-specific variables as population density (a higher population density can be viewed as a surrogate measure of pedestrian street exposure, as well as high traffic levels, leading to higher pedestrian fatalities) and commute mode share of pedestrians (for similar reasons as the effects of population density). These variables would then get manifested in the 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 pedestrian being fatally injured in a crash at any given *instant of time* for a given long-term crash propensity y^*_{q14}. 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 fatally injured. 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 (due to higher pedestrian activity and exposure in and around areas with high levels of commercial or residential development). Further, the GOR framework in Equation (3.1) accommodates spatial dependency in counts through spatial lag ("spillover") effects and spatial correlation 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 pedestrian injury occurs at any given time (and, therefore, not the threshold elements that affect the mapping of the latent propensity to the observed count outcome). Our expectation is that Census tract factors associated with socio-demographic, land-use and road network, activity intensity, and commute mode share variables are all likely to affect "long-term" latent injury risk propensity and will have a bearing on the "spillover" effects at other neighboring tracts. On the other hand, land-use variables are likely to also load (more so than other variables) on the thresholds that affect the translation of the risk propensity to observed injury outcomes.

The GOR framework of Equations (3.1) and (3.2) nests several other count modeling structures. Thus, if the variances of all the error terms in Equation (3.1), except the

ε_{qjs} term, are set to zero, and if $\boldsymbol{b}_{js} = 0 \ \forall \ j, s$, if $\delta = 0$ and if $\alpha_{js,m_{qjs}} = 0$ for all values of m_{qjs} for all j and s, the result is independent Poisson count models for each road-user type and severity level (see CPB). If the restriction that the variance of each of the error terms $(\omega_{qs}, u_{qj}, v_q)$ is zero is relaxed, one gets a specific version of a mixed multivariate Poisson model. In addition, if the restriction that $\delta = 0$ is relaxed, the result is a spatial mixed multivariate Poisson model that still does not provide adequate flexibility in treating situations with excess zeros. On the other hand, if the variances of all the error terms in Equation (3.1), except the ε_{qjs} term, are set to zero, and if $\delta = 0$, the result is independent flexible count models for each road-user type and severity level. If the restriction that the variance of each of the error terms $(\omega_{qs}, u_{qj}, v_q)$ is zero is relaxed, one gets the joint flexible count (JFC) model for injuries of all combinations of road-user type and severity level. In addition, if the restriction that $\delta = 0$ is lifted, the result is our proposed spatial joint flexible count (SJFC) model that allows spatial dependency across all Census tracts.

3.2   Model Estimation

To proceed forward, we first write the equation system in (3.1) compactly. To do so, define the following $(S \times 1)$ vectors of vertically stacked propensities, count outcome indices, observed count outcomes, and combined error terms $\left[\eta_{qj}^* \ (= v_q + u_{qj} + \omega_{qs} + \varepsilon_{qjs}) \right]$; $\boldsymbol{y}_{qj}^* = (y_{qj1}^*, \ y_{qj2}^*, \ldots, \ y_{qjS}^*)'$, $\boldsymbol{y}_{qj} = (y_{qj1}, \ y_{qj2}, \ldots, \ y_{qjS})'$, $\boldsymbol{m}_{qj} = (m_{qj1}, \ m_{qj2}, \ldots, \ m_{qjS})'$, and $\boldsymbol{\eta}_{qj}^* = (\eta_{qj1}^*, \ \eta_{qj2}^*, \ldots, \ \eta_{qjS}^*)'$.

Also, define additional vectors and matrices:

$$
\left.
\begin{aligned}
\boldsymbol{y}_q^* &= [(\boldsymbol{y}_{q1}^*)', (\boldsymbol{y}_{q2}^*)', (\boldsymbol{y}_{q3}^*)', \ldots, (\boldsymbol{y}_{qJ}^*)']', \quad \boldsymbol{m}_q = (\boldsymbol{m}_{q1}', \boldsymbol{m}_{q2}', \boldsymbol{m}_{q3}', \ldots, \boldsymbol{m}_{qJ}')' \\
\boldsymbol{y}_q &= [(\boldsymbol{y}_{q1})', (\boldsymbol{y}_{q2})', (\boldsymbol{y}_{q3})', \ldots, (\boldsymbol{y}_{qJ})']', \quad \boldsymbol{\eta}_q = (\boldsymbol{\eta}_{q1}', \ \boldsymbol{\eta}_{q2}', \ \boldsymbol{\eta}_{q3}', \ldots, \ \boldsymbol{\eta}_{qJ}')'
\end{aligned}
\right\} (JS \times 1 \text{ vectors})
$$

$$
\left.
\begin{aligned}
\boldsymbol{y}^* &= [(\boldsymbol{y}_1^*)', (\boldsymbol{y}_2^*)', (\boldsymbol{y}_3^*)', \ldots, (\boldsymbol{y}_Q^*)']', \quad \boldsymbol{m} = (\boldsymbol{m}_1', \boldsymbol{m}_2', \boldsymbol{m}_3', \ldots, \boldsymbol{m}_Q')' \\
\boldsymbol{y} &= [(\boldsymbol{y}_1)', (\boldsymbol{y}_2)', (\boldsymbol{y}_3)', \ldots, (\boldsymbol{y}_Q)']', \quad \boldsymbol{\eta} = (\boldsymbol{\eta}_1', \ \boldsymbol{\eta}_2', \ \boldsymbol{\eta}_3', \ldots, \ \boldsymbol{\eta}_Q')'
\end{aligned}
\right\} (QJS \times 1 \text{ vectors})
$$

$\boldsymbol{b}_j = (\boldsymbol{b}'_{j1}, \ \boldsymbol{b}'_{j2}, \ \boldsymbol{b}'_{j3}, \dots, \ \boldsymbol{b}'_{jS})'$ (SK x 1 vector), $\boldsymbol{b} = (\boldsymbol{b}'_1, \ \boldsymbol{b}'_2, \ \boldsymbol{b}'_3, \dots, \ \boldsymbol{b}'_J)'$ (JSK x 1 vector), $\tilde{\boldsymbol{x}}_q = \boldsymbol{I}_{JS} \otimes \boldsymbol{x}'_q$ (JS x JSK matrix; $\boldsymbol{I}_{JS}$ is an identity matrix of size JS) and $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}'_1, \tilde{\boldsymbol{x}}'_2, \tilde{\boldsymbol{x}}'_3, \dots \tilde{\boldsymbol{x}}'_Q)'$ (QJS x JSK matrix). Collect all the weights $w_{qq'}$ into a row-normalized spatial weight matrix $\boldsymbol{W}$. With these definitions, Equation (3.1) may be re-written as:

$$y^* = \delta(\boldsymbol{W} \otimes \boldsymbol{I}_{JS})y^* + \tilde{\boldsymbol{x}}\boldsymbol{b} + \boldsymbol{\eta}, \tag{3.3}$$

After further matrix manipulation to write $\boldsymbol{y}^*$ in reduced form, we obtain:

$$\boldsymbol{y}^* = \boldsymbol{C}\tilde{\boldsymbol{x}}\boldsymbol{b} + \boldsymbol{C}\boldsymbol{\eta}, \text{ where } \boldsymbol{C} = [\boldsymbol{I}_{QJS} - \delta(\boldsymbol{W} \otimes \boldsymbol{I}_{JS})]^{-1}. \tag{3.4}$$

The expected value and variance of $\boldsymbol{y}^*$ may be obtained from the above equation after developing the covariance matrix for the error vector $\boldsymbol{\eta}$. To do so, note that the error vector $\boldsymbol{\eta}$ is distributed multivariate normal with a mean vector of zero and covariance matrix $\boldsymbol{I}_Q \otimes \boldsymbol{\Lambda}$ (of size QSJ x QSJ), where $\boldsymbol{\Lambda}$ is the covariance matrix implied by the common error components in the elements of the error vector $\boldsymbol{\eta}$. Finally, we obtain $\boldsymbol{y}^* \sim MVN_{QT}(\boldsymbol{B}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{B} = \boldsymbol{C}\tilde{\boldsymbol{x}}\boldsymbol{b} \text{ and } \boldsymbol{\Sigma} = \boldsymbol{C}\left[\boldsymbol{I}_Q \otimes \boldsymbol{\Lambda}\right]\boldsymbol{C}'. \tag{3.5}$$

The parameter vector to be estimated in the model is $\boldsymbol{\theta} = (\boldsymbol{b}', \delta, \boldsymbol{\gamma}', \boldsymbol{\alpha}', \boldsymbol{\mu}')'$ where $\boldsymbol{\alpha}$ is a column vector obtained by the vertical stacking of the $\alpha_{js,r}$ ($r = 0, 1, 2, \dots L_{qs}$) parameters across severity levels and road-user types, and $\boldsymbol{\mu}$ is a column vector obtained by vertically stacking the elements $\sigma, \tau_1, \tau_2, \dots, \tau_J, \pi_1, \pi_2, \dots,$ and π_S. The likelihood function for the model is:

$$L(\boldsymbol{\theta}) = P(\boldsymbol{y} = \boldsymbol{m}) = \int_{D_{y*}} \phi_{QJS}(\boldsymbol{y}^*|\boldsymbol{B}, \boldsymbol{\Sigma})d\boldsymbol{y}^*, \tag{3.6}$$

17

where

$$D_{y^*} = \{\boldsymbol{y}^* : \Psi_{(qjs,m_{qjs}-1)} < y^*_{qjs} < \Psi_{qjs,m_{qjs}}, \forall q = 1,2,\ldots,Q, j = 1,2,\ldots,J, s = 1,2,\ldots,S\}$$

and $\phi_{QJS}(\cdot|\boldsymbol{B},\boldsymbol{\Sigma})$ is the multivariate normal density function of dimension QJS (with mean $\boldsymbol{B}$ and covariance matrix $\boldsymbol{\Sigma}$), $\boldsymbol{m}$ is a QJS x 1 - vector of observed count outcomes. 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 QSJ. 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, J, and S (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 count observations across all combinations of spatial units, road-user types, and severity levels. To write this function, define the following vectors:

$$\boldsymbol{\varphi}_{qj} = (\Psi_{qj1,m_{qj1}-1}, \Psi_{qj2,m_{qj2}-1}, \ldots, \Psi_{qjS,m_{qjS}-1}), \ \boldsymbol{\varphi}_q = (\boldsymbol{\varphi}_{q1}, \boldsymbol{\varphi}_{q2}, \ldots, \boldsymbol{\varphi}_{qJ}), \text{ and } \boldsymbol{\varphi} = (\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_Q)$$

and $\boldsymbol{\vartheta}_{qj} = (\Psi_{qj1,m_{qj1}}, \Psi_{qj2,m_{qj2}}, \ldots, \Psi_{qjS,m_{qjS}}), \boldsymbol{\vartheta}_q = (\boldsymbol{\vartheta}_{q1}, \boldsymbol{\vartheta}_{q2}, \ldots, \boldsymbol{\vartheta}_{qJ})$ and $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \ldots, \boldsymbol{\vartheta}_Q)$.

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

$$
\begin{aligned}
L_{CML}(\boldsymbol{\theta}) &= \left(\prod_{g=1}^{QSJ-1} \prod_{g'=g+1}^{QSJ} P\left([\boldsymbol{y}]_g = [\boldsymbol{m}]_g, [\boldsymbol{y}]_{g'} = [\boldsymbol{m}]_{g'}\right) \right) \\
&= \left(\prod_{g=1}^{QSJ-1} \prod_{g'=g+1}^{QSJ} \begin{bmatrix} \Phi_2(\tilde{\vartheta}_g, \tilde{\vartheta}_{g'}, \boldsymbol{\nu}_{gg'}) - \Phi_2(\tilde{\vartheta}_g, \tilde{\varphi}_{g'}, \boldsymbol{\nu}_{gg'}) \\ -\Phi_2(\tilde{\varphi}_g, \tilde{\vartheta}_{g'}, \boldsymbol{\nu}_{gg'}) + \Phi_2(\tilde{\varphi}_g, \tilde{\varphi}_{g'}, \boldsymbol{\nu}_{gg'}) \end{bmatrix} \right),
\end{aligned}
\tag{3.7}
$$

where, $\tilde{\vartheta}_g = \dfrac{[\boldsymbol{\vartheta}]_g - [\boldsymbol{B}]_g}{\sqrt{[\boldsymbol{\Sigma}]_{gg}}}, \tilde{\varphi}_g = \dfrac{[\boldsymbol{\varphi}]_g - [\boldsymbol{B}]_g}{\sqrt{[\boldsymbol{\Sigma}]_{gg}}}$, $\boldsymbol{\nu}_{gg'} = \dfrac{[\boldsymbol{\Sigma}]_{gg'}}{\sqrt{[\boldsymbol{\Sigma}]_{gg}}\sqrt{[\boldsymbol{\Sigma}]_{g'g'}}}.$

In the above expression, $[\boldsymbol{\vartheta}]_g$ represents the gth element of the column vector $\boldsymbol{\vartheta}$, and similarly for other vectors. $[\boldsymbol{\Sigma}]_{gg'}$ represents the gg^{th} element of the matrix $\boldsymbol{\Sigma}$. The

CML estimator is obtained by maximizing the logarithm of the function in Equation (3.7).

Under usual regularity assumptions, the CML estimator of $\boldsymbol{\theta}$ is consistent and asymptotically normally 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):

$$\boldsymbol{V}_{CML}(\hat{\boldsymbol{\theta}}) = [\boldsymbol{G}(\boldsymbol{\theta})]^{-1} = [\boldsymbol{H}(\boldsymbol{\theta})]^{-1}\boldsymbol{J}(\boldsymbol{\theta})[\boldsymbol{H}(\boldsymbol{\theta})]^{-1} \tag{3.8}$$

where,

$$\boldsymbol{H}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \text{ and } \boldsymbol{J}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log L_{CML}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\right]$$

The reader is referred to Bhat (2011) for complete details regarding the estimation of the matrices $\boldsymbol{H}(\boldsymbol{\theta})$ and $\boldsymbol{J}(\boldsymbol{\theta})$ in Equation (3.8) above. To ensure the constraints on the autoregressive term δ, we parameterize it as $\delta = 1/\left[1 + \exp\left(\tilde{\delta}\right)\right]$. Once estimated, the $\tilde{\delta}$ estimate can be translated back to an estimate of δ.

3.3 Model Selection

For the purpose of comparing two nested models estimated using the CML approach, one 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) for details regarding the computation of the **ADCLRT** test statistic).

Chapter 4

Study Area Description and Data

The crash data used in this research has been obtained from the CrashStat website, which is the result of a project undertaken by the New York Citys (NYC) Transportation Alternatives organization. The CrashStat website maintains geo-coded data for crashes involving bicyclists and pedestrians over several years, with the latest year being 2009. The data was compiled using crash reports from local reporting agencies, including the New York Police Department and the New York State Department of Motor Vehicles (for details on how the data was compiled and processed, please refer to the CrashStat Documentation). The crash database collected detailed information on every reported crash involving bicyclists and pedestrians in NYC, including crash factors (such as weather, road surface condition, and pedestrian/bicyclist action prior to the crash), bicyclist/pedestrian and vehicle driver details, and the details of the vehicles involved in the crash (such as vehicle age, vehicle type, and body type).

In addition to the CrashStat data, we used other data sources to obtain the land-use, demographic, and network information of the Census tracts (which is the spatial unit of analysis used in this study). Specifically, we obtained (a) the socio-demographic information from the 2010 Census data and the American Community Survey five-year estimates, (b) the land-use and road network variables from the 2009 zoning district maps and the street network map of the NYC Department of City Planning (NYC-DCP) for the Manhattan region, (c) the activity intensity variables from the tax lot details and the selected facilities and program sites data of NYC-DCP and (d) the commute mode shares and transit supply variables from the American Community Survey five-year estimates and the New York Metropolitan Transportation Council

(NYMTC) data. The 2010 TigerLine shape files were used to aggregate the data from these data sources to the Census tract level. All the geographic data processing was accomplished using ArcGIS 10.0 and the open source Geospatial Modeling Environment (see: http://www.spatialecology.com/gme/).

4.1 Sample Formation and Description

Bicycle and pedestrian crashes that occurred in the year 2009 in Manhattan constitute the sample used for the analysis in this study. The injury severity of each non-motorized road-user in a crash was recorded on a four point ordinal scale: (C) possible injury, (B) non-incapacitating injury, (A) incapacitating injury and (K) fatal injury. For our analysis, all crashes in 2009 involving non-motorized road-users within the limits of Manhattan were extracted from the CrashStat database, and were mapped to one of 285 Census tracts.[1] The counts of pedestrians and bicyclists injured per crash by severity level were next aggregated up to the Census tract level, to obtain the count of bicyclists and pedestrians injured by severity level in each of the 285 Census tracts. Across all Census tracts, the sample included a total of 2512 injured pedestrians and 845 injured bicyclists (the term "injured" as used here includes fatally injured individuals), indicating about 3 times more pedestrians injured relative to bicyclists. In the entire nation, on the other hand, the statistics for the year 2009 indicate a ratio of about 1.25 times more pedestrians injured relative to bicyclists (NHTSA, 2012). Clearly, these figures show a much higher proportion of injuries sustained by pedestrians (as a fraction of non-motorized road-users) in Manhattan relative to the nation as a whole, a reflection of the substantially higher density of pedestrian movement compared to bicyclist movement in Manhattan.

---

[1]Manhattan is divided into 288 Census tracts. However, we excluded three Census tracts from the analysis, corresponding to Liberty Island, Governors Island, and Randalls and Wards Islands. Of these, the first two tracts are primarily tourist attractions and recorded zero residential populations. Randalls and Wards Islands, which together constitute one Census tract, predominantly consist of parks and public facilities (such as the Manhattan Psychiatric Center and the Kirby Forensic Psychiatric Center), with limited public access and residential populations. Also, all these three Census tracts recorded zero bicycle and pedestrian crashes in 2009.

The distribution of the number of injured non-motorists by injury severity level (across all Census tracts in Manhattan) is presented in Table 4.1. For both pedestrians and bicyclists, the dominant injury types are "possible" and "non-incapacitating" injuries, with a lower share of "possible" injuries and higher share of "non-incapacitating" injuries for bicyclists relative to pedestrians. This is to be expected because of the speed of travel of bicyclists. In the category of "fatal" injuries, Table 4.1 reveals that there were no fatal injuries recorded amongst crash-involved bicyclists in Manhattan in 2009. However, there were 39 pedestrians killed in roadway crashes during the same period, reinforcing the higher density of pedestrian movement in Manhattan (in the nation as a whole, the number of bicyclist fatalities in roadway crashes was 15% of the number of pedestrian fatalities; see NHTSA, 2012). Overall, 1.2% of non-motorized users involved in a roadway crash were fatally injured in Manhattan, according to the CrashStat database (see Table 4.1, last column and penultimate row). In comparison, 4% of non-motorized users involved in a roadway crash were fatally injured in the nation as a whole, according to the NHTSA. The general skew toward less serious injury severity levels for both bicyclists and pedestrians in Manhattan may be attributed to high traffic congestion levels and consequent low motorized vehicle speeds. For example, according to the New York City (NYC) Department of Transportation, the speed of an average taxicab is 7.7 mph for the Midtown area of Manhattan (NYCDOT, 2010). Also, Manhattan has a high number of pedestrians and bicyclists due to its dense development. So, it is possible that a "safety in numbers" situation is at play, wherein the injury severity risk faced by pedestrians or bicyclists decreases as the number of pedestrians or cyclists increases (see Bhatia and Wier, 2011, Elvik, 2009, and Jacobson, 2003).

Table 4.1: Distribution of Number of Injured Non-Motorists by Injury Severity Level

Injury Severity	Pedestrian		Bicyclist		All Non-Motorists	
Possible injury	1700	67.7%	502	59.4%	2202	65.6%
Non-Incapacitating injury	523	20.8%	259	30.7%	782	23.3%
Incapacitating injury	250	10.0%	84	9.9%	334	9.9%
Fatal injury	39	1.5%	0	0.0%	39	1.2%
Total	**2512**	**100.0%**	**845**	**100.0%**	**3357**	**100.0%**

We next examine the sample distributions of non-motorized injuries by Census tract. The total number of non-motorized individuals injured during the year in traffic crashes per Census tract in Manhattan varied between 0 and 48, with an average of about 12 injuries per Census tract. Figures 4.1a - 4.1d present the distribution (across Census tracts) of the count of pedestrian injuries alongside that of bicyclist injuries for different injury severity levels in the study area for the year 2009. Several observations may be made from the figures. First, and as expected, there is a preponderance of Census tracts with zero count values for each road-user type-injury severity level. The percentage of tracts with zero values steadily increases from about 9% for pedestrians with possible injury (27% for bicyclists with possible injury) to 90% for pedestrians with fatal injury (100% for bicyclists with fatal injury). Further, for the possible injury severity level in particular, we also observe local spikes at non-zero count values. Such count accumulations (or inflations) in discrete probability mass are easily accommodated in our proposed model using the threshold parameters $\boldsymbol{\alpha}$. Second, it is clear from the figures that the count range and the distribution pattern of injuries across Census tracts varies substantially by road-user type as well as severity level, confirming the need to study injury counts by road-user type and severity level rather than pooling all injuries together.

Figure 4.2 is a thematic map displaying the total number of non-motorized injuries in each Census tract. While we have developed the map for each road-user type-injury severity level combination separately, the essential visual result that there is geographic clustering in count values holds for all the combinations. Thus, to economize on space, we are showing only the map for total number of pedestrian and bicyclist non-motorized injuries. The obvious spatial clustering in Figure 4.2 in the number of non-motorized injuries reinforces the notion that spatial dependency effects are likely to be at play when modeling injury counts at the Census tract level (or at any other unit of space).

Figure 4.1a: Distribution of percentage of Census tracts associated with each count of possible pedestrian injuries alongside possible bicyclist injuries

Figure 4.1b: Distribution of percentage of Census tracts associated with each count of non-incapacitating pedestrian injuries alongside non-incapacitating bicyclist injuries

Figure 4.1c: Distribution of percentage of Census tracts associated with each count of incapacitating pedestrian injuries alongside incapacitating bicyclist injuries
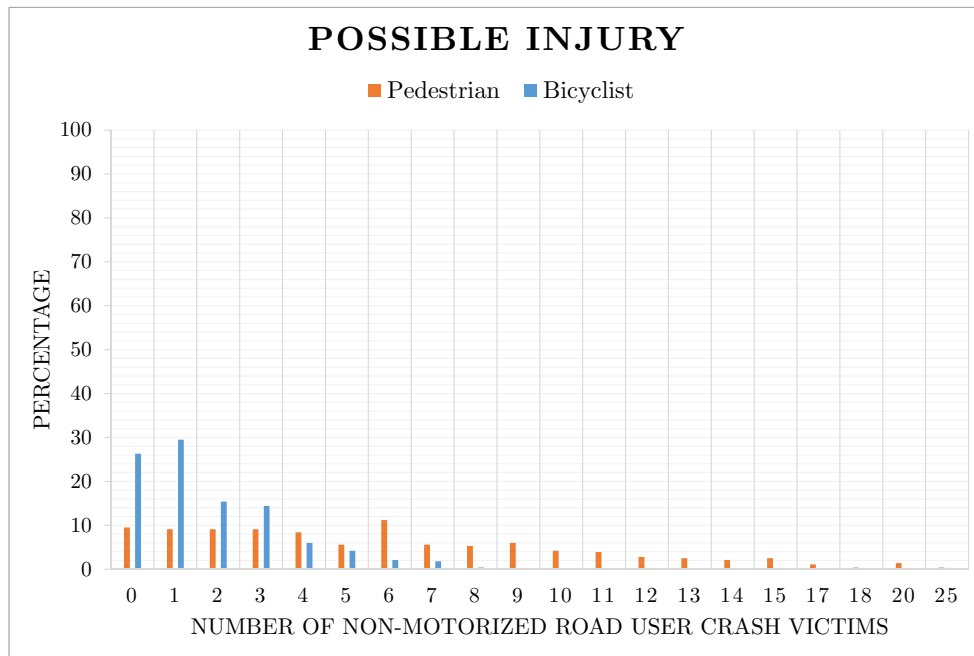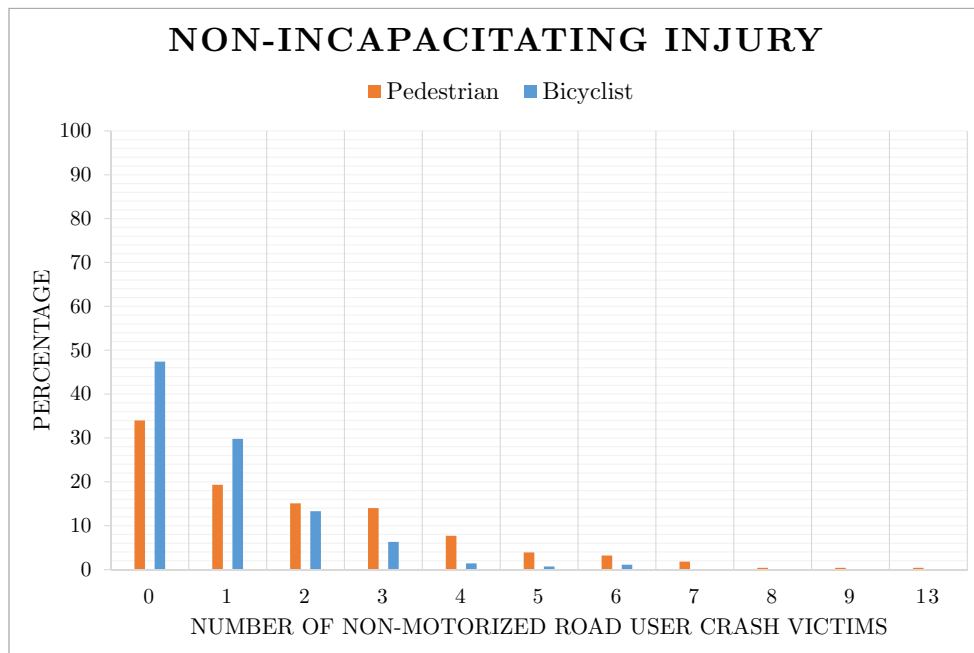

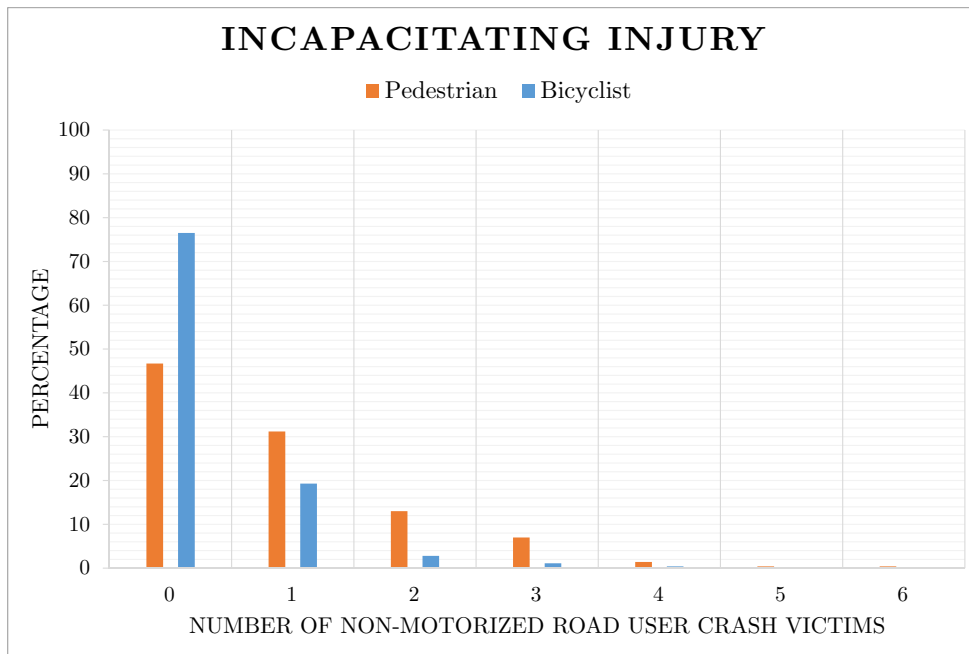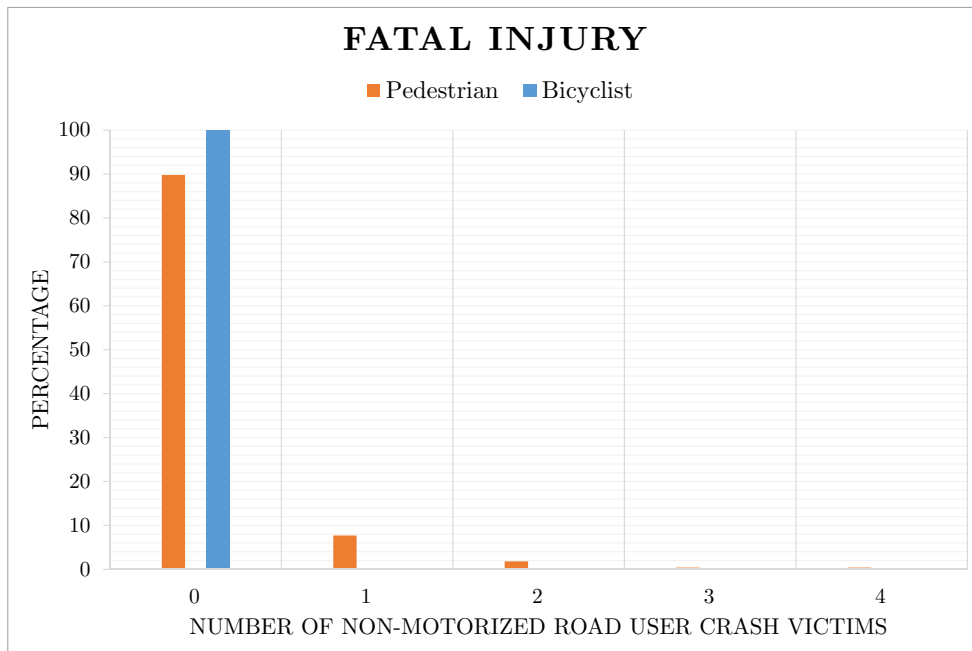**Figure 4.1d:** Distribution of percentage of Census tracts associated with each count of fatal pedestrian injuries alongside fatal bicyclist injuries
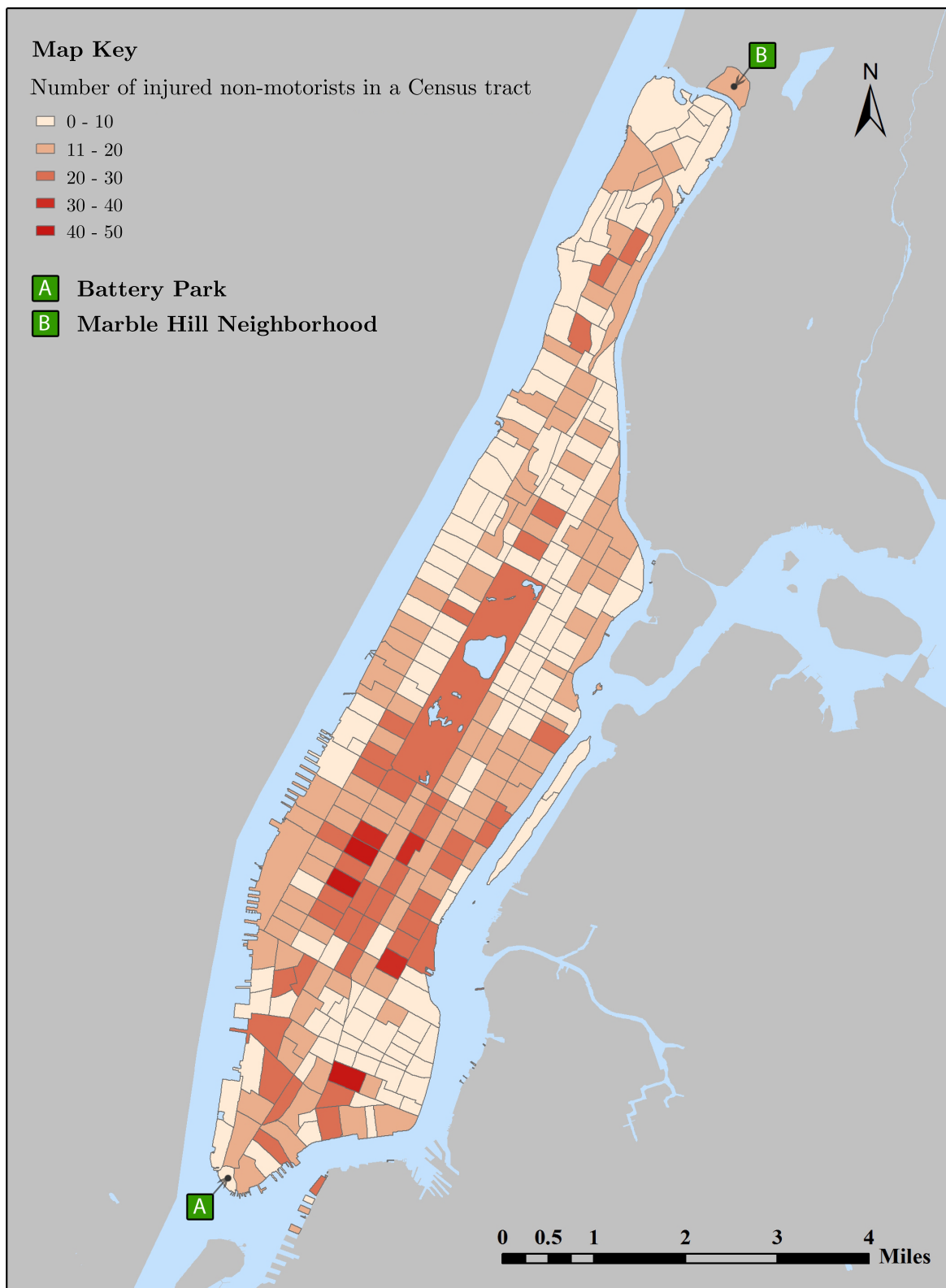
Figure 4.2: Thematic map of non-motorized injuries aggregated to Census tracts

Table 4.2 presents the sample characteristics of the 285 Census tracts.[2] The average area of a Census tract is 19.7 x 10^4 sq. meters, though Table 4.2 indicates a wide variation, which also manifests itself in the population density variable. 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.[3] However, the descriptive statistics in Table 4.2 also indicate 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 percentage of the population below poverty level varies substantially across Census tracts in Manhattan, with a mean (across Census tracts) of 18% (this is higher than the corresponding percentage in 2009 of 14.3% in the US population (DeNavas-Walt *et al.*, 2010)).[4]

---

[2] Many variables in Table 4.2 did not turn out to be statistically significant in our final empirical model; however, these variables are included in Table 4.2 to provide a sense of the variables considered in our analysis as well as for completeness in characterizing the study area.

[3] 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 by Census tracts (which is the unit of spatial analysis in the current study). 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.

[4] The U.S. Census Bureau employs a set of money income thresholds that differ across family size and composition to determine poverty level. If a family's total income is less than the corresponding threshold, then that family and every individual in it is considered to be in poverty. For further details on poverty measurement the reader is referred to: `http://www.census.gov/hhes/www/poverty/about/overview/measure.html`

Table 4.2: Descriptive Statistics of Census Tracts (285 Observations)

Variable	Minimum	Maximum	Mean	Std. Dev.
Socio-Demographic Variables				
Total area in square meters (scaled by 10000)	4.13	293.99	19.70	19.67
Population density (population per sq-meter)	0.00*	0.08	0.03	0.02
Race/Ethnicity Variables				
Proportion of non-Hispanic White population	0.00	0.91	0.48	0.30
Proportion of non-Hispanic Black and African American population	0.00	0.81	0.15	0.20
Proportion of non-Hispanic Asian population	0.00	0.88	0.12	0.13
Proportion of Hispanic population	0.00	1.00	0.23	0.23
Proportion of all other non-Hispanic population	0.00	0.05	0.02	0.01
Percent below poverty level	0.00	0.54	0.18	0.13
Age Distribution				
Proportion of population aged 14 years and below	0.00	0.24	0.12	0.05
Proportion of population aged 15-19 years	0.00	0.50	0.05	0.05
Proportion of population aged 20-29 years	0.00	0.67	0.21	0.10
Proportion of population aged 30-64 years	0.15	0.82	0.49	0.07
Proportion of population aged 65 years and above	0.00	0.37	0.13	0.07
Educational Attainment Distribution				
Proportion of population 18 years and above without high school degree	0.00	0.51	0.15	0.15
Proportion of population 18 years and above with high school degree	0.00	0.44	0.14	0.09
Proportion of population 18 years and above with some college or associate's degree	0.00	0.63	0.17	0.07
Proportion of population 18 years and above with Bachelor's degree or higher	0.05	1.00	0.54	0.26
Median household income (scaled by $10,000)	0.98	23.28	7.28	4.21

* All Census tracts had a non-zero value of population density. But the value of this variable for some Census tracts is very low (of the order 0.001 or lower), and so the minimum is listed as 0.0

Table 4.2: Descriptive Statistics of Census Tracts (285 Observations) [**continued**]

Variable	Minimum	Maximum	Mean	Std. Dev.
Land-use and Road Network Variables				
Land-use type Distribution				
Proportion of commercial land-use	0.00	1.00	0.30	0.32
Proportion of industrial land-use	0.00	0.96	0.07	0.17
Proportion of residential land-use	0.00	1.00	0.57	0.35
Proportion of other land-uses (vacant lots, open space, recreational *etc.*)	0.00	0.95	0.06	0.15
Roadway Type Distribution				
Proportion of highways	0.00	0.78	0.02	0.07
Proportion of local neighborhood roads and city streets	0.22	1.00	0.91	0.14
Proportion of bicycle lanes and trails	0.00	0.40	0.03	0.07
Proportion of other road types (alleys, driveways *etc.*)	0.00	0.45	0.04	0.11
Activity Intensity Variables				
Number of schools	0.00	10.00	1.81	1.94
Number of universities	0.00	5.00	0.15	0.50
Park area in US Acres	0.00	7.06	0.06	0.44
Intensity of office activity	0.00	9.57	0.79	1.72
Intensity of retail activity	0.00	1.62	0.18	0.24
Commute Mode Shares and Transit Supply				
Mode Share Distribution				
Drive alone	0.00	0.23	0.07	0.04
Shared ride	0.00	0.17	0.02	0.02
Transit	0.00	0.91	0.57	0.15
Walk	0.00	1.00	0.22	0.15
Telecommuting	0.00	0.38	0.06	0.04
Other modes (taxicab, motorcyclist *etc.*)	0.00	0.39	0.06	0.04
Transit Supply				
Number of bus stops	0.00	60.00	8.03	5.95
Number of subway stops	0.00	6.00	0.49	0.81
Distance between Centroids of Census Tracts (miles)	0.09	13.15	3.78	2.52

The mean proportions (across Census tracts) of the population in different age groups indicates a sizeable proportion (70%) in the age group of 20 to 64 years (the percentage of this age group in the overall US population is about 64%) and a relatively high proportion (13%) of the population over 65 years (the corresponding percentage in the US population is 9%).[5] The remaining socio-demographic variables in Table 4.2 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 Bachelors degree or higher while 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 Bachelors 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 is computed as the ratio of the tract land area in that specific type to the total tract land area. The New York City zoning regulations govern these designations of permitted land-use. The statistics for the land-use variables in Table 4.2 show that 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 road network variables are constructed as the ratio of the total length of a specific road type in the Census tract to the total length of the road carriageway (including bicycle lanes and trails) in that Census tract. As can be observed from Table 4.2, the

---

[5]The zero values in Table 4.2 for the minimum value of percentage of population in certain age groups is because of the fact that some of the Census tracts in Manhattan have very small total resident populations.

Manhattan Census tracts have a very high proportion of local neighborhood roads and city streets.

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. The number of Universities is the number of post-secondary degree granting institutions in the Census tract. The park area shows substantial variation across counties. The last two variables in this category of variables, the intensity of office activity and the intensity of retail activity, are computed as the ratio of total floor space allocated for office use and retail use, respectively, to the total land area of the Census tract. This serves as a measure of the extent to which office and retail activities are concentrated in a Census tract.[6] There is clear evidence of high office activity in the Manhattan Census tracts, which is to be expected as Manhattan is the nerve center for many financial institutions. The intensity of retail activity is modest in comparison.

The commute mode share and transit supply variables toward the end of Table 4.2 reveal the high transit and walk mode shares in the region. The final statistic in Table 4.2 provides information on the Euclidean distance between centroids of Census tracts, which is used as the metric to characterize spatial proximity when constructing spatial weight matrices. The average inter-Census tract distance is 3.78 miles, with a minimum of 0.09 miles and a maximum of 13.15 miles (the maximum distance corresponds roughly to the length of the line from Battery Park at the southern tip of Manhattan to the Marble Hill neighborhood at the northernmost end of Manhattan; see Figure 4.2).

---

[6]In cities such as Manhattan, the net floor area in, for example, office activity in a Census tract can be more than the land area of the Census tract (because of the vertical build-up). Thus, the intensity measures can be higher than 1 (the land-use measures previously discussed, however, are confined to the 0-1 range).

Chapter 5

Empirical Analysis

5.1 Variables Considered

We considered all the variables listed in Table 4.2 for the analysis, and several variable specifications and functional forms for the variables, in the process of arriving at the final model specification. Many of the Census tract variables (such as age and race/ethnicity distribution, educational attainment, land-use type distribution, roadway type distribution, and commute mode share) were introduced as categorical variables. Several other Census tract variables (total area, median household income, population density, area of parks, and intensity of office and retail activity) were introduced in a continuous form (for the total area, population density and median household income, we also considered a logarithmic transformation; such a transformation could not be considered for the other continuous variables because these other variables did take the value of zero for some Census tracts. In addition, spline effects of the continuous variables as well as dummy variables created from the continuous variables were considered to introduce non-linearities. Other variables (number of schools and universities, number of bus stops, and number of subway stops) were introduced as is, in the form of exogenous count variables. All the above variables were introduced in both the latent variable and threshold specifications.

The variables retained in the final model specification are based on their statistical significance and intuitive explanatory power. Overall, the results suggest that there are substantial differences in the factors that impact the number of injured non-motorists across road-users as well as across injury severity levels.

5.2 Estimation Results

We estimated three different model formulations (1) A set of seven independent models one for each combination of non-motorized road-user type and injury severity; we will refer to this model as the *independent flexible count (FC) model*, (2) A joint model allowing for cross-correlation effects among the count variables based on the error components in Equation (3.1), but no spatial correlation; this is the *joint flexible count (JFC) model* referenced in Section 3.1, and (3) A spatial joint model allowing auto-regressive spatial dependency as well as cross-correlation effects; this is the *spatial joint flexible count (SJFC)* model referenced in Section 3.1. To keep the discussion focused and compendious, we present only the estimation results for the SFJC model here, though we will discuss data fit issues of all the three models in a later section.

Table 5.1 presents the estimation results. We first discuss, in the next section, the variable effects on the long-term (injury) risk propensity and on the thresholds (that affect the "instantaneous" translation of the propensity to the count outcome) for the pedestrian sub-component of the model. Section 5.2.2 presents the corresponding results for the bicycle sub-component of the model.[1] Section 5.2.3 summarizes the main findings and discusses implications. Section 5.2.4 presents the results of cross-correlation and spatial effects in the proposed spatial multivariate count model, and Section 5.2.5 provides data fit measures.

5.2.1 Pedestrian Injury Model Component

5.2.1.1 Long Term Injury Risk Propensity

The variable effects in Table 5.1a reveal that Census tracts with a high population density have a high risk propensity for fatal pedestrian injuries. This is a manifestation

---

[1]In these sections, the base categories for the categorical explanatory variables correspond to those not listed in the tables (for example, if the only race/ethnicity proportion variable appearing in the specification is the "proportion of Hispanic population", it implies that all the other race/ethnicity categories listed in Table 4.2 together constitute the base category).

of a pedestrian exposure effect on the street network. In particular, regions with high residential population density are known to be, in general, areas of low income and relatively good transit service, leading to a substantial fraction of walk trips. In addition to an exposure effect, this result could also be a result of a social deprivation effect due to relatively poor cross-walk and pedestrian facilities. Several earlier studies have found a similar effect of population density on total pedestrian crashes (see, for example, Cottrill and Thakuriah, 2010, Ha and Thill, 2011, Wier *et al.*, 2009). However, our study, which partitions injuries by severity level, indicates that this effect of population density is particularly disturbing, because of the loading on fatal pedestrian injuries (with no impact on the number of pedestrian injuries at lower severity levels). There is a suggestion that the quality and availability of pedestrian facilities, and more generally, access facilities, in dense urban areas have to be reviewed and evaluated carefully, both from a traffic safety standpoint and from an environmental justice standpoint (for example, Lyons *et al.*, 2008 identify the lack of secure play space, jobs-housing balance issues, little separation between dwelling units and busy streets, and hazardous routes from home to school as being mobility problems associated with the urban poor). The result above is reinforced by the next finding that tracts with a large proportion of Hispanic population appear to be particularly at risk for pedestrian injuries at all severity levels except for fatal injuries (where it has no effect) (see also Loukaitou-Sideris *et al.*, 2007 and Cottrill and Thakuriah, 2010).[2] The socio-demographic variable effects in Table 5.1a also indicate that tracts with a high proportion of the population 14 years of age or below have a lower long term risk propensity of experiencing pedestrian injuries at the non-incapacitating and fatal injury severity levels (as also observed by LaScala *et al.*, 2000). Further, a high proportion

---

[2]There was multicollinearaity among the Census tract-level socio-demographic variables of population density, proportion of minority populations, median household income, and percent below poverty level. At the end, the best specification was achieved with the first two variables in the long term risk propensity, and the median household income in the threshold effects discussed in the next section. The "percent below poverty level" variable turned out to be statistically insignificant after accommodating the other three variables.

of teenagers in the age group of 15-19 years decreases the long term risk propensity of pedestrian injuries at all severity levels except the incapacitating injury level. These effects may be related to an exposure effect, where tracts with a high share of children and teenagers generate fewer walking trips and walking mileage (presumably related to the general reluctance of parents to allow children to walk due to safety and security concerns; see Sidharthan *et al.*, 2011; McMillan, 2007 and Prezza *et al.*, 2005). Finally, within the set of socio-demographic variables, Table 5.1a reveals the strong impacts of education level on pedestrian risk propensity. While education levels have seldom been included in earlier studies (but see LaScala *et al.*, 2000), our results indicate a lower risk propensity of fatal pedestrian injuries, and "possible" pedestrian injuries in tracts with a high proportion of adults (age 18 years and above) with a Bachelor's degree or higher. These education-related effects may be capturing another dimension of exposure (for instance, individuals with low education are more likely to be blue collared field workers, who are then exposed more to roadway hazards), or may be a reflection of higher safety awareness and consciousness levels among highly educated individuals. While the reasons for the influence of education, as provided above, are admittedly speculative, they do suggest the importance of the education dimension in the 4Es of safety - engineering, enforcement, education and emergency medical services - as identified by the Federal Highway Administration (FHWA, 2006) and highlight the need for conducting educational campaigns to promote safe pedestrian and roadway practices across the region and particularly in areas with low education levels.

Among the land-use and road network variables, four variables turned out to be statistically significant (at the 0.1 level or lower). Interestingly, each of these variables had an impact on injury risk propensity for only one of the four possible injury severity levels, strongly supporting the count analysis of pedestrian injuries by injury severity level (as opposed to modeling the count of total pedestrian injuries regardless of severity

level). Table 5.1a shows a high risk propensity of non-incapacitating pedestrian injuries in tracts with a high proportion of commercial land-use, presumably a reflection of higher levels of walking in and around commercial land-uses (this is also consistent with the results of Loukaitou-Sideris *et al.*, 2007). The road network variable effects indicate the lower risk propensity for non-incapacitating pedestrian injuries in tracts with a high proportion of highways and local neighborhood roads/city streets, relative to tracts with a high proportion of other road-way types (driveways, alleys, etc.) (perhaps capturing the heightened pedestrian alertness levels on roadways with high automobile volumes), though there are no effects of these network variables on pedestrian injuries for other severity levels. The effect of the final network variable, "proportion of bicycle lanes and trails", indicates the benefits of providing exclusive non-motorized mode use facilities to reduce pedestrian injuries.

The influence of the activity intensity and the "walk commute mode share" variables are all as expected, and indicate the heightened long term risk propensity for injuries of various severity levels caused by increased pedestrian activity.

5.2.1.2 Threshold Parameters

The threshold parameters include the threshold specific constants ($\alpha_{js,1}, \alpha_{js,2}, ...\alpha_{js,L_{js}}$ values), as well as the parameters associated with the $\boldsymbol{\gamma}$ vector (see Equation (3.2)). The threshold specific constants do not have any substantive interpretations. However, 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). In the pedestrian models, our analysis indicated no need for these flexibility terms for all injury severity categories except for the possible injury category (consistent with the initial observations from Figure 4.1a). The elements in the $\boldsymbol{\gamma}$ vector are presented next in Table 5.1a. The constants within the $\boldsymbol{\gamma}$ vector for the four injury severity levels do not have any particular interpretation. For the other

variables, a positive coefficient shifts all the thresholds toward the left of the injury propensity scale, which has the effect of reducing the probability of the zero injury outcome (increasing the overall probability of the non-zero outcome). A negative coefficient, on the other hand, shifts all thresholds toward the right of the injury propensity scale, which has the effect of increasing the probability of the zero injury outcome (decreasing the overall probability of the non-zero outcome; see CPB). The results in Table 5.1a indicate that high median household income Census tracts tend to have a higher observed level of non-zero pedestrian non-incapacitating injuries than other Census tracts, for the same level of long-term risk propensity of such injuries, an observation that needs more research to tease out the precise relationship between income levels and pedestrian injuries by severity level. High proportions of commercial, industrial, and residential land-uses (relative to open and recreational land-uses) in a tract also lead to an increase in non-zero count values for incapacitating pedestrian injuries, perhaps for the reasons identified in Section 3.1. Finally, the effects of the remaining variables reflect the higher likelihood of non-zero "non-incapacitating" injuries in tracts with many schools, a reduction in incapacitating injuries in tracts with a high transit commute mode share (perhaps due to the consequent reduction of motorized vehicle trips), and an increase in the count of non-zero "possible" injuries in tracts with a high walk commute mode share.

Table 5.1a: Model Estimation Results for Pedestrian Injuries
(Weight Matrix: inverse of distance, Distance Band: 5 miles)

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								
Population density (logarithmic)							0.152	8.45
Proportion of Hispanic population	1.552	14.21	1.788	9.15	0.745	4.26		
Proportion of population aged 14 and below			-4.283	-4.93			-1.956	-12.47
Proportion of population between ages 15-19	-0.035	-12.26	-0.044	-8.38			-0.083	-9.36
Proportion of population 18 years and above with Bachelor's degree or higher	-2.782	-12.04					-2.271	-21.79
Land-use and Road Network Variables								
Proportion of commercial land-use			1.099	7.78				
Proportion of highways			-3.305	-6.03				
Proportion local neighborhood roads and city streets			-0.944	-4.99				
Proportion of bicycle lanes and trails					-2.741	-3.28		
Activity Intensity Variable								
Intensity of office activity					0.136	3.28		
Number of schools	1.143	11.12						
Number of universities			0.181	2.87				
Commute Mode Shares and Transit Supply								
Walk commute mode share			1.169	3.50			2.275	26.79

Table 5.1a: Model Estimation Results for Pedestrian Injuries
(Weight Matrix: inverse of distance, Distance Band: 5 miles) [**continued**]

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	-0.075	-4.87						
α_5	-0.245	-11.66						
α_{11}	-0.255	-9.13						
γ *Vector*								
Constant	1.238	20.19	0.990	4.26	-1.223	-3.88	1.043	7.28
Socio-Demographic Variables								
Median household income			0.030	5.12				
Land-use and Road Network Variables								
Proportion of commercial land-use	0.781	22.04			1.073	3.11		
Proportion of industrial land-use					1.668	4.35		
Proportion of residential land-use					1.112	3.41		
Activity Intensity Variable								
Number of Schools			0.421	5.93				
Commute Behaviors and Transit Supply								
Transit commute mode share					-0.972	-2.88		
Walk commute mode share	1.065	13.23						

40

5.2.2 Bicyclist Injury Model Component

For the bicyclist injury component of the model system, only three severity levels are considered: possible injury, non-incapacitating injury, and incapacitating injury. This is because, as discussed in Section 4.1, there were no bicyclist fatalities in any of the Census tracts in Manhattan in the year 2009.

5.2.2.1 Long Term Injury Risk Propensity

Among the socio-demographic variables, Census tracts with a high proportion of teenage populations aged 15 to 19 years of age have a low long term risk propensity for non-incapacitating and incapacitating injuries. This is similar to the result found for the case of pedestrian injuries. This reduction in non-incapacitating and incapacitating injuries may be attributable to the New York State law that requires NYC bicyclists under 13 years of age to wear a state approved helmet (Lee *et al.*, 2005, Kim *et al.*, 2007). Because of the helmet law enforcement at a young age, it is possible that teenage bicyclists continue to use a helmet and bicycle more safely. However, a more in-depth causal analysis needs to be undertaken before a definitive connection can be drawn between helmet use and the fewer number of bicyclist injuries.

The effects of the land-use and road network variables in Table 5.1b reveal a high risk propensity of non-incapacitating bicyclist injuries in tracts with a high proportion of commercial and industrial land-use, likely attributable to the higher levels of bicycling in and around commercial and industrial land-uses. Also, the presence of bicycling lanes and trails greatly decreases the long-term risk propensity of incapacitating bicyclist injuries. This is intuitive, because of the resulting separation of motorized and bicycle traffic. Interestingly, however, the presence of bicycling lanes and trails does not affect the risk propensity for injuries at other severity levels. The results also indicate the exposure-related positive effects of the number of schools, office intensity, and park area in the tract on the long term risk propensity for "possible" injuries.

In the group of the commute mode share and transit supply variables, there is a heightened long term risk propensity for non-incapacitating bicyclist injuries in tracts with a high walk commute mode share, presumably caused by generally higher bicyclist activity in zones with high walk commute mode share (the bicycling commute mode share, which would have been a more direct measure, was almost zero in the Census tracts in Manhattan; however, the walk commute mode share can be viewed as a surrogate measure of overall bicycling activity). Finally, Census tracts with a high percentage of workers who telecommute have a high risk propensity for "possible" and "non-incapacitating" bicyclist injuries. There is some evidence in the literature that telecommuting generates new short distance non-motorized trips during the middle parts of the day and in the evening (Andreev *et al.*, 2010). Such non-motorized trips would lead to an exposure-triggered higher bicyclist risk propensity.

5.2.2.2 Threshold Parameters

Among the effects of the land-use and road network variables, two turned out to be statistically significant. High proportions of commercial and industrial land-uses, and high retail intensity, in a Census tract lead to an increase in non-zero count values for the "possible" bicyclist injury category.

Table 5.1b: Model Estimation Results for Bicyclist Injuries
(Weight Matrix: inverse of distance, Distance Band: 5 miles)

Injury Severity	Possible		Non-Incapacitating		Incapacitating	
Parameter	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Long Term Risk Propensity						
Socio-Demographic Variables						
Proportion of population between ages 15-19			-0.044	-4.87	-0.023	-1.81
Land-use and Road Network Variables						
Proportion of commercial land-use			0.591	2.19		
Proportion of industrial land-use			0.872	3.25		
Proportion of bike lanes and trails					-4.228	-3.48
Activity Intensity Variable						
Number of schools	1.515	3.74				
Intensity of office activity	0.128	2.50				
Park area in US acres	0.655	18.85				
Commute Behaviors and Transit Supply						
Walk Commute mode share			1.148	1.73		
Telecommuting share	7.074	6.77	2.755	1.74		

Table 5.1b: Model Estimation Results for Bicyclist Injuries
(Weight Matrix: inverse of distance, Distance Band: 5 miles) [**continued**]

Injury Severity	Possible		Non-Incapacitating		Incapacitating	
Parameter	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Threshold Parameters						
Threshold Specific Constants						
α_2					-0.219	-3.79
α_3					-0.433	-6.08
α_4			-0.486	-4.78		
α_5			-0.877	-4.79		
$\boldsymbol{\gamma}$ *Vector*						
Constant	-1.693	-26.79	-1.239	-2.81	-0.908	-3.99
Land-use and Road Network Variables						
Proportion of commercial land-use	1.121	5.21				
Proportion of industrial land-use	1.901	8.59				
Activity Intensity Variable						
Retail Intensity					0.507	2.59

5.2.3   A Summary of Results and Implications

The results in the previous few sections provide several important general planning insights. First, socio-demographics appear to be much more of an influencing factor for the count of pedestrian injuries of all severity levels than for the count of bicyclist injuries. This is intuitive, since socio-demographics may be viewed, in part, as being proxy measures of exposure. In this context, pedestrian travel is generally dictated by the lack of availability of other modes of travel (which is related to demographics), while bicycle travel is more associated with a choice-based decision mechanism wherein bicycling is pursued for exercise and recreation (Xing *et al.*, 2010, Coogan *et al.*, 2007) Overall, Census tracts with a high population density, high proportion of Hispanic residents, high proportion of the population over 19 years of age, and with low education levels are particularly vulnerable to pedestrian injuries. As indicated earlier, this could be an exposure result, but could also be related to discrimination across neighborhoods in the level of non-motorized mode facility planning and investment. There is a clear need to continue to emphasize environmental justice considerations in traffic engineering and project planning/prioritization. Second, as anticipated in Section 3.1, the results for both pedestrian and bicyclist injuries indicate the particularly strong influence of land-use variables through the threshold effects, reinforcing the notion that distraction and pre-occupation among motorized drivers around commercial, industrial, and residential land-uses (relative to open and recreational land-uses) are issues of concern. At the same time, Census tracts with high built-up commercial and industrial land-use have a high long-term risk propensity of non-motorized injury (due to an exposure effect caused by higher pedestrian and bicycling activity). A similar situation applies to Census tracts with high office and retail intensity. Overall, there appears to be a situation of "dangerous convergence" where distraction and pre-occupation combine with high non-motorized mode activity, suggesting the institution of information campaigns (and enforcement mechanisms) to ensure that motorized vehicle drivers, and non-motorized mode users, are particularly vigilant and avoid cell

phone use and related distraction activities in densely built-up areas. Third, the results unequivocally underscore the need to invest in non-motorized mode infrastructure as a precursor to any actions directed toward increasing the share of non-motorized modes for the commute. That is, transportation policy actions that attempt to increase non-motorized mode use through mixed land-use development or road pricing strategies, without concurrent investment in improved non-motorized mode facilities, are likely to be unsuccessful on three counts: (a) safety is a consideration in mode choice decisions (see Chapter 1), and there will be less traction in increasing non-motorized mode use without a clear information campaign on the safety investments being made to reduce non-motorized user safety risk, (b) any increase in non-motorized mode use in response to mixed land-use or pricing actions (notwithstanding the earlier comment) will lead to a higher count of non-motorized mode user injuries in general, and fatal pedestrian injuries in particular, if the *status quo* is maintained in terms of non-motorized mode infrastructure (as per our estimation results), and (c) those "financially-challenged" segments of the population who may turn to non-motorized modes to avoid additional financial burden (in response to actions such as road pricing, even without investment in non-motorized facilities) become more exposed to injury risk, reinforcing what already appears to be environmental justice problems in the planning process. On the other hand, investment in non-motorized mode facilities, such as investment in bicycle lanes and trails, when undertaken in concert with other demand management actions, addresses the three obstacles just identified. More generally, our results underscore the need to carefully consider safety issues when exploring demand management actions, even those demand management actions that may appear to be innocuous from a safety standpoint. For example, our estimation results suggest an increase in bicyclist injuries as the telecommuting share increases. Finally, the presence of schools and universities increases the long term risk propensity of injuries, even though limited to only the less severe injury categories, emphasizing the need for the continuation of federal programs such as the Safe Routes to School program (U.S. Department of

Transportation or USDOT, 2005).

5.2.4 Cross-Correlation and Spatial Effects

Table 5.1c provides the estimates of the cross-correlation and spatial parameters. Among the parameters π_s of the error terms ω_{qs} (see Sections 2.1 and 3.1), only π_1 turned out to be statistically significant, suggesting the presence of Census tract-specific unobserved factors that impact the long-term risk propensity of the "possible injury" severity level for both pedestrians and bicyclists. In the set of τ_s parameters, only τ_2 of error term u_{q2} appears in the final specification, indicating Census tract-specific unobserved factors impacting pedestrian injury risk at all severity levels. The standard deviation σ of the error term ν_q is positive and statistically significant, reflecting the presence of common Census tract-specific unobserved factors that affect the risk propensity for all types of injuries at all severity levels. Overall, the results demonstrate the importance of considering a multivariate count modeling approach rather than estimating independent and univariate count models for each road-user type-injury severity level combination.

The spatial autoregressive parameter δ in the final spatial lag formulation is also highly statistically significant, with a positive value of 0.486. This result supports the hypothesis that the number of non-motorized injuries in a Census tract is not just a function of its characteristics, but is also influenced by the observed factors (such as retail intensity, land-use type, and road network characteristics) and unobserved factors (such as county regulations, unobserved design features, and driving attitudes of the people in the neighborhood) of spatially proximate Census tracts. As we will demonstrate in Section 5.3, ignoring these spatial effects can substantially bias the estimated effects of exogenous variables on the count of injuries.

Table 5.1c: SJFC Model: Additional Parameters and Summary Statistics
(Weight Matrix: inverse of distance, Distance Band: 5 miles)

	Estimate	t-stat
Cross model correlation		
π_1 – S.E. of error linked with "possible injury" injuries in a Census tract	1.041	21.31
τ_2 – S.E. of error linked with pedestrian injuries in a Census tract	0.421	2.45
σ – S.E. of error linked with individual Census tract	0.597	5.67
δ – (*spatial correlation parameter*)	0.486	8.01
Number of observations	285	
Number of parameters estimated	59	
Log-composite likelihood at convergence	-1694396.25	

5.2.5 Model Selection and Data Fit

The spatial joint flexible count model (SJFC) is superior to both the joint flexible count model (JFC) model and the independent flexible count model (FC), as should be evident from the statistically significant spatial lag autoregressive parameter and other error components in Table 5.1c. This may also be observed by computing the adjusted composite likelihood ratio test (**ADCLRT**) for testing the SJFC model with the other two models. To do so, we first provide the composite log-likelihood (CLL) values for the three models: the SJFC model (CLL value of -1694396.25 with 59 parameters), the JFC model (CLL value of -1695251.25 with 58 parameters), and the FC model (CLL value of -1749546.60 with 55 parameter). The **ADCLRT** statistic for the comparison between the SJFC and JFC models is 7.07, which is greater than the critical chi-squared value corresponding to one degree of freedom even at the 0.01 level of significance.[3] Similarly, the **ADCLRT** statistic for the comparison between the SJFC model and the FC model is 427.29, which is higher than the critical chi-squared table value corresponding to four degrees of freedom at any reasonable level of significance. For completeness, the **ADCLRT** value for the test between the JFC and FC models is 404.16, which is higher than the corresponding chi-squared table value with three degrees of freedom at any reasonable significance level.

5.3 Aggregate Elasticity Effects

The estimated model parameters in Table 5.1, and discussed in Section 5.2, do not directly provide the magnitude of impact of variables on injury frequency. In this section, we compute the aggregate-level "elasticity effects" from the SJFC models for selected variables (we focus only on the SJFC model, and only on selected variables, to focus the presentation and conserve on space). The variables selected are based

---

[3] Unlike in the case of maximum likelihood, the CLRT statistic computed as twice the difference of the CLL values is not chi-squared distributed. An adjustment needs to be made to the CLRT statistic to obtain the **ADCLRT** statistic that is chi-squared distributed (see Bhat, 2011 and Pace *et al.*, 2011).

on the discussion in Section 5.2.3, and include the following: (1) population density, (2) proportion of Hispanic population, (3) proportion of commercial land-use, and (4) proportion of bicycle lanes and trails. For each variable, the "elasticity" computed is a measure of the percentage change in total injury count (for each road-user type-injury severity level combination) across the entire study region (see Appendix for details). To compute the aggregate level "elasticity effect" of population density, we increase the population density of each tract by 20%. For the remaining variables, we increase the proportion by 0.2 for each Census tract.[4]

The elasticity effects for the SJFC model (along with their t-statistics) are presented in Table 5.2. The first entry in the second row of the table indicates that an increase in the proportion of the Hispanic population by 0.2 in a tract would, on average, result in about a 24.9% increase in the tract in the annual count of "possible injury" pedestrian injuries, while the second entry in the same row indicates a 54.3% increase in the annual count of "non-incapacitating" pedestrian injuries. Other entries may be similarly interpreted. The results indicate the statistical significance of all the implied elasticity effects. Further, three other important observations may be made. <u>First</u>, it is obvious that each variable can have quite different impacts on the counts of injuries based on road-user type and injury severity level, highlighting the potential pitfalls of using an aggregated total non-motorized injury count as the dependent variable. <u>Second</u>, the elasticity effects combine the effects of variables on both the long-term risk propensity as well as the threshold parameters. Thus, the effect of commercial land-use on the expected number of "possible" and "incapacitating" pedestrian injuries originates from the threshold effect, while its effect on the expected number of "non-incapacitating"

---

[4]Strictly speaking, we should modify other proportions within appropriate groups of variables. For example, the sum of all land-use proportions after increasing the proportion of commercial land-use by 0.2 should continue to remain at 1.0 for each tract; this may be easily achieved by drawing away from each non-commercial land-use in direct proportion to the current distribution of each non-commercial land-use share in the tract. However, doing so makes it difficult to isolate the impacts of the variables under study because of the changes in other variables too. So, we follow a more straightforward approach to assess the impact of each proportion variable by simply increasing its value by 0.2.

injuries originates from the long-term risk propensity effect. The implied elasticity effects of the variables (including for the proportion of commercial land-use) in Table 5.2 are consistent with the discussion in Section 5.2, but these elasticity effects are more tangible because they provide an estimate of the overall impact of each variable on the expected count. Third, we also computed the elasticity effects for the simple FC model that ignores the jointness of counts (in the number of injuries by road-user type and severity level) and spatial dependence. In general, the elasticity effects from the SJFC model are higher in magnitude than those from the FC model, a consequence of the "spillover" effects in the SJFC model that causes a spatial multiplier effect.[5] Specifically, a change in a variable in one Census tract influences the injury count in other Census tracts that then has a circular ripple impact back on the initial Census tract. The FC model ignores such spatial "multiplier" effects because it considers the injury count in one Census tract to be independent of injury counts in other Census tracts. The result can be quite different estimates of variable effects. For instance, a 0.2 increase in the proportion of bicycle lanes and trails in a tract, as per the FC model, would result in only a 38% (67%) decrease in pedestrian (bicycle) non-incapacitating injury counts. In contrast, the SJFC model in Table 5.2 indicates a 62% (89%) decrease in pedestrian (bicycle) non-incapacitating injury counts. This, and other similar results for other variables, underscores the potentially misinformed investments in crash-related injury reduction countermeasures if jointness across counts of different types and/or spatial dependencies are ignored. In the particular case of bicycle lanes and trails, the FC model underestimates the benefits that would accrue from investing in more bicycle lanes and trails.

---

[5]For the few cases where the FC model has a higher elasticity magnitude, the corresponding variable effect is through the thresholds and not through the long-term risk propensity variable that contributes to the spillover effect. Population density is the only exception, and the higher FC elasticity for this variable is because of the logarithmic transformation used for this variable.

Table 5.2a: Aggregate-Level Elasticity Effects of SJFC Model (Pedestrian)

Variable	Pedestrian							
	Possible		Non-Incapacitating		Incapacitating		Fatal	
	Elasticity	t-stat	Elasticity	t-stat	Elasticity	t-stat	Elasticity	t-stat
Population density	0.00	-	0.00	-	0.00	-	8.68	2.98
Proportion of Hispanic population	24.85	6.31	54.33	3.54	24.76	3.82	0.00	-
Proportion of commercial land-use	16.33	17.37	32.19	3.70	17.33	2.72	0.00	-
Proportion of bicycle lanes and trails	0.00	-	0.00	-	-61.66	-4.12	0.00	-

Table 5.2b: Aggregate-Level Elasticity Effects of SJFC Model (Bicyclist)

Variable	Bicyclist					
	Possible		Non-Incapacitating		Incapacitating	
	Elasticity	t-stat	Elasticity	t-stat	Elasticity	t-stat
Population density	0.00	-	0.00	-	0.00	-
Proportion of Hispanic population	0.00	-	0.00	-	0.00	-
Proportion of commercial land-use	11.83	4.07	18.92	0.53	0.00	-
Proportion of bicycle lanes and trails	0.00	-	0.00	-	-89.41	-6.97

Chapter 6

Conclusions

This thesis has proposed a new econometric approach to specify and estimate a model of non-motorized injury frequency. It is based on the recasting of count models as a special case of a generalized ordered-response (GOR) framework, which then conveniently allows for the accommodation of zero inflation, cross-correlation, and spatial dependency in spatial multivariate count model systems. A composite marginal likelihood inference approach is used to estimate the model parameters. To our knowledge, this is the first such formulation of a spatial multivariate count model in the literature.

The thesis has modeled the number of pedestrian and bicycle injuries by injury severity level in the Census tracts within Manhattan, New York. The empirical results highlight the need to (1) differentiate injury counts by road-user type as well as injury severity level, (2) use a multivariate modeling system for the analysis of injury counts by road-user type and injury severity level, rather than estimating independent univariate count models for each road-user type-injury severity level combination, and (3) accommodate a spatial lag structure to accommodate dependence effects in injury counts across space. Accommodating these important econometric considerations is not simply an esoteric scholarly issue, but has very real implications for accurately capturing variable effects, for predictive ability, and for informed decision-making.

From a substantive standpoint, Census tracts with a high population density, minority population groups, low education levels, and high built-up density are particularly vulnerable to pedestrian and bicycle injuries. This suggests a need to examine

environmental justice considerations in non-motorized mode facility provision, as well as consider information campaigns (and enforcement mechanisms) to encourage motorized vehicle drivers, and non-motorized mode users, to exercise particular caution and avoid distraction when driving in densely built-up areas. Our results also underscore the need to invest in non-motorized mode infrastructure and improve non-motorized road-user safety as a precursor to implementing travel demand management actions (such as mixed land-use development and road pricing) directed toward promoting non-auto mode use.

Appendix

Procedure to Predict the Expected Count Values for each Census Tract

The expected value of injury count in Census tract q for each road-user type j and injury severity level s may be written as:

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

where $P(y_{qjs} = k)$ is the probability of occurrence of k injuries of type j and injury severity level s 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 combinations of road-user type and injury severity level in the estimation sample, corresponding to the possible injury severity level for pedestrian injuries; see Figure 4.1a). 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) is a function of the $(QSJ \text{ x } 1)$ matrix of exogenous variables for all Q Census tracts, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ..., \boldsymbol{x}_Q)'$, as well as a function of the variable vector $\boldsymbol{z}_q$ embedded in the thresholds in Equation (3.2).

The estimate of $P(y_{qjs} = k)$ in Equation (A.1) for the FC model is obtained from Equation (3.1) in a fairly straightforward manner. For the JFC model, we need to accommodate the effects of the error covariances across different severity levels and road-user types within a Census tract, and, for the SJFC model, we also need to consider the spatial dependency effects across Census tracts. To estimate $P(y_{qjs} = k)$

in these models, we simulate the QSJ x 1 vector $\boldsymbol{y}^*$, from Equation (3.4), five hundred times using the estimated values of δ, $\boldsymbol{b}$, and the QSJ x 1 vector $\boldsymbol{\eta}$. Subsequently, we compare each of the 500 draws of the qth element of $\boldsymbol{y}^*$ with the corresponding thresholds for the qth element from Equation (3.2), 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 $P(y_{qjs} = k)$.[1]

---

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

References

Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention, 38*(3), 618-625. doi:10.1016/j.aap.2005.12.006

Aguero-Valverde, J., & Jovanis, P. P. (2009). Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record: Journal of the Transportation Research Board, 2136*, 82-91. doi:10.3141/2136-10

Aguero-Valverde, J., & Jovanis, P. P. (2010). Spatial correlation in multilevel crash frequency models. *Transportation Research Record: Journal of the Transportation Research Board, 2165*, 21-32. doi:10.3141/2165-03

Alfò, M., & Maruotti, A. (2010). Two-part regression models for longitudinal zero-inflated count data. *Canadian Journal of Statistics, 38*(2), 197-216. doi:10.1002/cjs.10056

Anastasopoulos, P. C., Shankar, V. N., Haddock, J. E., & Mannering, F. L. (2012). A multivariate tobit analysis of highway accident-injury-severity rates. *Accident Analysis & Prevention, 45*, 110-119. doi:10.1016/j.aap.2011.11.006

Andreev, P., Salomon, I., & Pliskin, N. (2010). Review: State of teleactivities. *Transportation Research Part C, 18*(1), 3-20. doi:10.1016/j.trc.2009.04.017

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Kluwer Academic, Dordrecht, The Netherlands.

Bassett, Jr., D. R., Pucher, J., Buehler, R., Thompson, D. L., & Crouter, S. E. (2008). Walking, cycling, and obesity rates in Europe, North America, and Australia. *Journal of Physical Activity and Health, 5*(6), 795-814.

Beck, L. F., Dellinger, A. M., & O'Neil, M. E. (2007). Motor vehicle crash injury rates by mode of travel, United States: Using exposure-based methods to quantify differences. *American Journal of Epidemiology, 166*(2), 212-218. doi:10.1093/aje/kwm064

Bermúdez, L., & Karlis, D. (2011). Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics, 48*(2), 226-236. doi:10.1016/j.insmatheco.2010.11.001

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. doi:10.1016/j.trb.2011.04.005

Bhat, C. R., Varin, C., & Ferdous, N. (2010). A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered-response model. In Greene, W., Hill, R. C. (eds.), *Advances in Econometrics: Maximum Simulated Likelihood Methods and Applications, 26*, 65-106, Emerald Group Publishing Limited, Bingley, UK.

Bhatia, R., & Wier, M. (2011). Safety in Numbers re-examined: Can we make valid or practical inferences from available evidence? *Accident Analysis & Prevention, 43*(1), 235-240. doi:10.1016/j.aap.2010.08.015

Buck, A. J., Blackstone, E. A., & Hakim, S. (2009). A multivariate Poisson model of consumer choice in a multi-airport region. *iBusiness, 1*(2). doi:10.4236/ib.2009.12012

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. doi:10.1016/j.trb.2011.09.007

Carter, D. and Council, F. (2007). Factors contributing to pedestrian and bicycle crashes on rural highways. Transportation Research Board 86th Annual Meeting Paper #07-2457, *Transportation Research Board*, Washington, DC.

Chib, S., & Winkelmann, R. (2001). Markov chain Monte-Carlo analysis of correlated count data. *Journal of Business & Economic Statistics, 19*(4), 428-435. doi:10.1198/07350010152596673

Chiou, Y.-C., & Fu, C. (2012). Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. *Accident Analysis & Prevention*, forthcoming. doi:10.1016/j.aap.2012.03.030

Cottrill, C. D., & Thakuriah, P. (Vonu). (2010). Evaluating pedestrian crashes in areas with high low-income or minority populations. *Accident Analysis & Prevention, 42*(6), 1718-1728. doi:10.1016/j.aap.2010.04.012

Coogan, M., Karash, K., Adler, T., & Sallis, J. (2007). The role of personal values, urban form and auto availability in the analysis of walking for transportation. *American Journal of Health Promotion, 21*(4 Suppl), 363-370.

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. doi:10.1007/s11116-011-9363-8

DeNavas-Walt, C., Proctor B. D., & Smith J. C. (2010). Income, poverty, and health insurance coverage in the United States: 2009. *Current Population Reports*, U.S. Census Bureau, Washington, DC. Retrieved from: `http://www.census.gov/prod/2010pubs/p60-238.pdf`

El-Basyouny, K., & Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis & Prevention, 41*(4), 820-828. doi:10.1016/j.aap.2009.04.005

Elvik, R. (2009). The non-linearity of risk and the promotion of environmentally sustainable transport. *Accident Analysis & Prevention, 41*(4), 849-855. doi:10.1016/j.aap.2009.04.009

Federal Highway Administration (2006). Strategic highway safety plans: A champion's guide to saving lives. Retrieved from: `http://safety.fhwa.dot.gov/safetealu/guides/guideshsp040506/`

Franzese, R.J. & Hays, J.C., 2008. Empirical models of spatial interdependence. In Box-Steffensmeier, J.M., Brady, H.E., Collier, D., (eds.), *The Oxford Handbook of Political Methodology*, 570-604, Oxford University Press, Oxford.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics, 31*(4), 1208-1211. doi:10.1214/aoms/1177705693

Gotschi, T., & Mills, K. (2008). Active transportation for America: The case for increased federal investment in bicycling and walking. Rails to Trails Conservancy, Washington, DC. Retrieved from: `http://www.railstotrails.org/resources/documents/whatwedo/atfa/ATFA_20081020.pdf`

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. doi:10.1016/j.compenvurbsys.2010.12.004

Herriges, J. A., Phaneuf, D. J., & Tobias, J. L. (2008). Estimating demand systems when outcomes are correlated counts. *Journal of Econometrics, 147*(2), 282-298. doi:10.1016/j.jeconom.2008.09.026

Humes K. R., Jones, N. A., & Ramirez, R. R. (2011). Overview of race and Hispanic origin: 2010. *2010 Census Briefs*, U.S. Census Bureau, Washington, DC. Retrieved from: `http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf`

Jacobsen, P. L. (2003). Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *Injury Prevention, 9*(3), 205-209. doi:10.1136/ip.9.3.205

Kim, J.-K., Kim, S., Ulfarsson, G. F., & Porrello, L. A. (2007). Bicyclist injury severities in bicyclemotor vehicle accidents. *Accident Analysis & Prevention, 39*(2), 238-251. doi:10.1016/j.aap.2006.07.002

Ladrón de Guevara, F., Washington, S., & Oh, J. (2004). Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. *Transportation Research Record: Journal of the Transportation Research Board, 1897*, 191-199. doi:10.3141/1897-25

LaScala, E. A., Gerber, D., & Gruenewald, P. J. (2000). Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. *Accident Analysis & Prevention, 32*(5), 651-658. doi:10.1016/S0001-4575(99)00100-1

Lee, B. H.-Y., Schofer, J. L., & Koppelman, F. S. (2005). Bicycle safety helmet legislation and bicycle-related non-fatal injuries in California. *Accident Analysis & Prevention, 37*(1), 93-102. doi:10.1016/j.aap.2004.07.001

Lee, A. H., Wang, K., Scott, J. A., Yau, K. K. W., & McLachlan, G. J. (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research, 15*(1), 47-61. doi:10.1191/0962280206sm429oa

LeSage, J. P., & Pace, R. K. (2009). *Introduction to Spatial Econometrics.* CRC Press, Taylor & Francis Group, Boca Raton, FL.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A, 44*(5), 291-305. doi:10.1016/j.tra.2010.02.001

Loukaitou-Sideris, A., Liggett, R., & Sung, H.-G. (2007). Death on the crosswalk: A study of pedestrian-automobile collisions in Los Angeles. *Journal of Planning Education and Research, 26*(3), 338-351. doi:10.1177/0739456X06297008

Lyons, R. A., Towner, E., Christie, N., Kendrick, D., Jones, S. J., Hayes, M., Kimberlee, R., Sarvotham, T., Macey, S., Brussoni, M., Sleney, J., Coupland, C., & Phillips, C. (2008). The advocacy in action study: A cluster randomized controlled trial to reduce pedestrian injuries in deprived communities. *Injury Prevention, 14*(2), e1-e5. doi:10.1136/ip.2007.017632

McAndrews, C. (2011). Traffic risks by travel mode in the metropolitan regions of Stockholm and San Francisco: a comparison of safety indicators. *Injury Prevention, 17*(3), 204-207. doi:10.1136/ip.2010.029306

McMillan, T. E. (2007). The relative influence of urban form on a child's travel mode to school. *Transportation Research Part A, 41*(1), 69-79. doi:10.1016/j.tra.2006.05.011

McMillen, D. P. (2010). Issues in spatial data analysis. *Journal of Regional Science, 50*(1), 119-141. doi:10.1111/j.1467-9787.2009.00656.x

Metropolitan Transportation Commission (2009). Regional bicycle plan for the San Francisco Bay Area 2009 Update. Retrieved from: `http://www.mtc.ca.gov/planning/bicyclespedestrians/MTC_Regional_ Bicycle_Plan_Update_FINAL.pdf`

Miaou, S.-P., & Song, J. J. (2005). Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis & Prevention, 37*(4), 699-720. doi:10.1016/j.aap.2005.03.012

Mitra, S. (2009). Spatial autocorrelation and Bayesian spatial statistical method for analyzing intersections prone to injury crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2136*, 92-100. doi:10.3141/2136-11

Müller, G., & Czado, C. (2005). An autoregressive ordered probit model with application to high-frequency financial data. *Journal of Computational and Graphical Statistics, 14*(2), 320-338. doi:10.1198/106186005X48687

National Highway Traffic Safety Administration (2012). Quick facts 2010. National Highway Traffic Safety Administration, U.S. Department of Transportation. Retrieved from: `http://www-nrd.nhtsa.dot.gov/Pubs/811616.pdf`

New York City Department of Transportation (2010). Green light for midtown evaluation report. Retrieved from: `http://www.nyc.gov/html/dot/downloads/pdf/broadway_report_final2010_web.pdf`

Pace L., Salvan A. & Sartori N (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica 21*(1): 129-148.

Park, E. S., & Lord, D. (2007). Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record: Journal of the Transportation Research Board, 2019*, 1-6. doi:10.3141/2019-01

Prezza, M., Alparone, F. R., Cristallo, C., & Luigi, S. (2005). Parental perception of social risk and of positive potentiality of outdoor autonomy for children: The development of two instruments. *Journal of Environmental Psychology, 25*(4), 437-453. doi:10.1016/j.jenvp.2005.12.002

Pucher, J., & Dijkstra, L. (2003). Promoting safe walking and cycling to improve public health: lessons from the Netherlands and Germany. *American Journal of Public Health, 93*(9), 1509-1516. doi:10.2105/AJPH.93.9.1509

Pucher, J., Komanoff, C., & Schimek, P. (1999). Bicycling renaissance in North America? *Transportation Research Part A, 33*(6), 625-654. doi:10.1016/S0965-8564(99)00010-5

Pucher, J., Buehler, R., Bassett, D. R., & Dannenberg, A. L. (2010). Walking and cycling to health: A comparative analysis of city, state, and international data. *American Journal of Public Health, 100*(10), 1986-1992. doi:10.2105/AJPH.2009.189324

Quddus, M. A. (2008). Time series count data models: An empirical application to traffic accidents. *Accident Analysis & Prevention, 40*(5), 1732-1741. doi:10.1016/j.aap.2008.06.011

Schrank, D., Lomax, T., & Eisele, B. (2011). The 2011 urban mobility report. Texas Transportation Institute. Retrieved from: `http://mobility.tamu.edu`

Sener, I. N., Eluru, N., & Bhat, C. R. (2009). Who are bicyclists? Why and how much are they bicycling? *Transportation Research Record: Journal of the Transportation Research Board, 2134*, 63-72. doi:10.3141/2134-08

Siddiqui, C., Abdel-Aty, M., & Choi, K. (2012). Macroscopic spatial analysis of pedestrian and bicycle crashes. *Accident Analysis & Prevention, 45*, 382-391. doi:10.1016/j.aap.2011.08.003

Sidharthan, R., & Bhat, C. R. (2012). Incorporating spatial dynamics and temporal dependency in land use change models. *Geographical Analysis, 44*(4), 321-349. doi:10.1111/j.1538-4632.2012.00854.x

Sidharthan, R., Bhat, C. R., Pendyala, R. M., & Goulias, K. G. (2011). Model for children's school travel mode choice. *Transportation Research Record: Journal of the Transportation Research Board, 2213*, 78-86. doi:10.3141/2213-11

Southern California Association of Governments (2012). Bike Ped Plan Wiki. Accessed on 06/25/2012 from: `http://bikepedwiki.scag.ca.gov/bikepedtransportation/node/7`

U.S. Department of Transportation, Federal Highway Administration (2005). Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU). *Joint Explanatory Statement of the Committee of the Conference, House Report 109-203*, 866-867. Retrieved from: `http://safety.fhwa.dot.gov/saferoutes/overview/legislation.cfm`

U.S. Department of Transportation, Federal Highway Administration (2009). 2009 National Household Travel Survey. Retrieved from: `http://nhts.ornl.gov`

Ver Hoef, J. M., & Jansen, J. K. (2007). Spacetime zero-inflated count models of harbor seals. *Environmetrics, 18*(7), 697-712. doi:10.1002/env.873

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. doi:10.1016/j.aap.2011.05.016

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. doi:10.1016/j.aap.2008.10.001

Winters, M., Davidson, G., Kao, D., & Teschke, K. (2010). Motivators and deterrents of bicycling: comparing influences on decisions to ride. *Transportation, 38*(1), 153-168. doi:10.1007/s11116-010-9284-y

Xing, Y., Handy, S. L., & Mokhtarian, P. L. (2010). Factors associated with proportions and miles of bicycling for transportation and recreation in six small US cities. *Transportation Research Part D, 15*(2), 73-81. doi:10.1016/j.trd.2009.09.004

Zhao, Y., & Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics, 33*(3), 335-356. doi:10.1002/cjs.5540330303