The Dissertation Committee for Yiyi Wang Certifies that this is the approved version of the following dissertation:

A NEW SPATIAL MODEL FOR PREDICTING MULTIVARIATE COUNTS: ANTICIPATING PEDESTRIAN CRASHES ACROSS NEIGHBORHOODS AND FIRM BIRTHS ACROSS COUNTIES

Committee:

Kara Kockelman, Supervisor

Paul Damien

Dominique Lord

Michael Walton

Cara Wang

Zhanmin Zhang

A NEW SPATIAL MODEL FOR PREDICTING MULTIVARIATE COUNTS: ANTICIPATING PEDESTRIAN CRASHES ACROSS NEIGHBORHOODS AND FIRM BIRTHS ACROSS COUNTIES

by

Yiyi Wang, B.E.; M.E.

Dissertation

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

Doctor of Philosophy

The University of Texas at Austin

May 2013

For Aaron

Acknowledgements

I would like to express my deep gratitude to my advisor, Dr. Kara Kockelman, who is always generous in providing her advice and guidance to me. Her dedication to teaching and research is a steadfast stimulus for me to endure tough times throughout my doctoral study. She has played a central role in helping me develop the model specification and then editing the text of this dissertation. I also am very thankful to my committee members: to Dr. Paul Damien, for his timely responses to all my questions and leading me to the wonderful world of advanced Bayesian methods; to Dr. Dominique Lord, for his invaluable expertise in the field of crash modeling; to Dr. Michael Walton, for his prompt responses to my email requests and giving his advice and guidance in my academic pursuit; to Dr. Cara (Xiaokun) Wang, for her patiently answering my various questions and her invaluable expertise in spatial data analysis; and to Dr. Zhanmin Zhang, for generously providing information of the pavement management system, which has stimulated some research ideas in my future research plan.

I am deeply indebted to many professors and researchers throughout my education: Dr. Brad Carlin, whose work in disease mapping has laid the foundation for my dissertation work; Dr. Shaw-Ping Miaou, who has patiently answered my minute questions and sharing with me his insight in crash modeling; Dr. Ned Levine, whose work in crime data analysis has stimulated interesting exchanges; Dr. Olivier Parent, who has provided insightful comments and suggestions for my other paper on spatial multinomial probit model; Dr. James LeSage, whose book and toolbox in spatial econometrics open the door for me to the many opportunities in spatial data analysis; Dr. Ming-Chun Lee, whose ArcGIS expertise has greatly contributed to this dissertation work; Dr. Ghislaine De Regge, who has painstakingly read and revised my dissertation draft and with whom I have enjoyed many conversations; Dr. Randy Machemehl, whose graduate courses fortify my understanding of the transportation engineering field and whose warm smile and amiable demeanor always comfort me when I feel low; Dr. Stephen Boyles, who has so generously advised me on academic career and with whom I have so enjoyed working as a teaching assistant. I am also deeply grateful to Ms. Annette Perrone for her administrative and editorial contributions during my doctoral study.

I would like to extend my appreciation to my friends and colleagues in the Civil Engineering department: Dan Fagnant, Donna Chen, Xiaoxia Xiong, Brent Selby, Binny Paul, Sashank Gadda, and Dr. Jason Lemp for their support and terrific teamwork; Dr. Jianming Ma, Dr. Zheng Li, Marisol, Rajesh, Raghu, Yao, Ti, Nan, Ruoyu, and Hui, for their friendship, and many others who I have so enjoyed interacting with in the past years.

Last, I am forever indebted to my family: my parents, Guocheng Wang and Jie Dong, for giving me the sweetest home that I can ever imagine and for believing in me; my son, Aaron (Zi-Chen) for being a wonderful little angel and bearing with his busy mom (you are my best work); my husband, Lei Zhang, for being my rock.

A NEW SPATIAL MODEL FOR PREDICTING MULTIVARIATE COUNTS: ANTICIPATING PEDESTRIAN CRASHES ACROSS NEIGHBORHOODS AND FIRM BIRTHS ACROSS COUNTIES

Yiyi Wang, Ph.D.

The University of Texas at Austin, 2013

Supervisor: Kara M. Kockelman

Transportation research regularly relies on data exhibiting both space and time dimensions. Thanks to the rise of smartphones, Bluetooth, and other devices, geo-referenced data collection enables application of more behaviorally realistic – but complex – models that account for spatial autocorrelation, temporal correlation, and possible time-space interactions (e.g., time-lagged effects from a neighboring unit's response). One promising area is crash count prediction, where crash frequencies (and severities) at zones, intersections, and along roadways will generally exhibit some spatial relationships, due to missing variables, causal mechanisms, and other ties.

This dissertation work proposes and estimates a spatial multivariate count model and provides two case studies to implement such model. One case study is in the context of pedestrian-vehicle crash counts across zones in Austin, Texas, while accounting for network features (e.g., lane-miles and intersection density), land use factors (such as land use entropy and residential accessibility to commercial activities), population and job densities, and school access. Parameter estimates suggest that crash rates fall dramatically as WMT levels rise. Higher shares of residential parcels within one-half mile of commercial parcels are associated with elevated risks for both severe and non-severe pedestrian crashes (after controlling for WMT). Denser freeway and arterial street networks are associated with higher crash rates (for both severity levels), whereas denser local street networks are associated with lower rates. Positive spatial autocorrelation is present across Austin neighborhoods, as expected, due to missing variables that trend in space (such as street design features and demographics). The two crash rates

also exhibit spatially lagged cross-response correlation (spatially clustered and shared across crash types) and aspatial cross-correlation (representing locally omitted variables, like poor lighting conditions and the presence of unusual sight obstructions).

The other case study models new firm births by industry across U.S. counties, while controlling for population density, household incomes, and residents' age. New firms in each studied industry tend to be spatially clustered, perhaps due to agglomeration economies as well as higher chances of attracting more patrons and business opportunities. A younger (and possibly more vital) work force (as quantified by each county's median-age value) is associated with more firm births (in 2009) in each of the three industry categories (basic, retail, and service firms).

The new model specification captures region-wide heterogeneity (thanks to extra variation introduced by the lognormal component in the mean crash-rate specification), correlations across two (or more) count types (in the same zone), and spatial autocorrelation among unobserved components. This new approach and associated application allow analysts to distinguish covariates' effects on multivariate crash and other counts from spatial spillover effects and cross-response correlations. This work adds to the literature by providing guidance on what types of specifications best reflect spatial count data while facilitating estimation (using large data sets) and illuminating the level and nature of spatial autocorrelation, multivariate correlation, and region-wide (latent) heterogeneity that exists in crash data after controlling for a host of observable factors.

Table of Contents

CHAPTER 1: INTRODUCTION 1

1.1 Overview and Motivation 1

1.2 A Brief Overview of Existing Methods for Crash Count Prediction 2

1.2.1 Aspatial Models for Crash Prediction 2

1.2.2 Spatial Models 3

1.2.3 Limitations of Existing Methods 5

1.3 Study Objectives and Organization 7

1.4 Chapter Summary 8

CHAPTER 2: LITERATURE REVIEW 9

2.1 Aspatial Models for Crash Prediction 9

2.2 Spatial Count Models 11

2.2.1 Motivation 11

2.2.2 Development 12

2.3 Pedestrian Crash Predictions 18

2.4 Estimation and Inference Methods 19

2.5 Chapter Summary 21

CHAPTER 3: METHODOLOGY 22

3.1 Univariate Conditional Autoregressive Models 22

3.2 Multivariate Conditional Autoregressive Models 23

3.3 A Flexible MCAR Model 26

3.4 A Poisson Log-Normal MCAR Model 28

3.4.1 Model Specification 28

3.4.2 Sampling Scheme 30

3.4.3 A Trivariate Case 34

3.4.4 Chapter Summary 35

CHAPTER 4: DATA SETS 36

4.1 Pedestrian Safety Data Set 36

4.1.1 Transit Stop Density 38

4.1.2 Land Use 38

4.1.3 Access to Schools 40

4.1.4 Roadway Features 42
4.2 4.3 Chapter Summary 51
CHAPTER 5: ANALYSIS AND RESULTS 53
5.1 Results of Simulated Data Test: Small-Sample Example with Two Response Levels 53
5.2 Results of Simulated Data Test: Large-Sample Example with Three Response Levels 65
5.3 Results of Zone-Level Pedestrian-Crash Model 75
5.3.1 Model for Walk-Miles Traveled (WMT) 75
5.3.2 Two-Response Pedestrian Crash Count Model 78
5.4 Model Results for Firm Birth Counts across Counties 87
5.4.1 Modeling Results of the Firm-Birth Model 89
5.5 Chapter Summary 93
CHAPTER 6: CONCLUSIONS 94
6.1 The Austin Application 95
6.2 The U.S. Firm Birth Application 96
6.3 Opportunities for Model Enhancements 97
6.4 Final Thoughts 99
APPENDIX A. 101
APPENDIX B. 105
REFERENCES 126

CHAPTER 1: INTRODUCTION

1.1 Overview and Motivation

Spatial models are regularly used to analyze behavioral data in transportation, economics, and geography, such as home prices (Case et al. 2003), land use change (Chakir and Parent 2009, Wang and Kockelman 2009, Wang et al. 2012), and roadway crashes (Levine et al. 1995a, 1995b, Miaou et al. 2003, Wang et al. 2009 and 2011). The unique nature of the response variables governs the types of model specification used. For example, land development outcomes or other choice responses are often cast in an unordered setting (leading to the marriage on multinomial probit (or logit) models and standard spatial stochastic processes, as in Chakir and Parent [2009] and Wang et al. [2012]), land intensity in an ordered probit regression setup (which yields the spatial ordered probit model, as described in Wang and Kockelman [2009]), and count data (e.g., traffic crashes [Miaou et al. 2003], disease outbreaks [Jin et al. 2005], and employment).

Compared to the many past studies addressing details of spatial modeling for categorical data, spatial count models have enjoyed relatively little exploration, with empirical studies relying on only a few, rather standard specifications. To this end, this dissertation devises a new spatial model for multivariate count data, while incorporating region-specific heterogeneity, spatial autocorrelation within each response level, cross-correlations across different response levels, and spatially-lagged cross correlations across different response levels. Two case studies are provided here: one for pedestrian crash counts, the other for firm births. The first is the centerpiece of this dissertation, offering highly detailed descriptions and results, whereas the latter showcases a trivariate-response application and makes use of a much larger data set (over 1,316 U.S. counties, rather than 218 Austin Census tracts).

The motivation for a spatial model of pedestrian crash-count data is significant. Walking is advocated as means of addressing multiple social and environmental issues, including air pollution, rising obesity from inactive lifestyles, neighborhood safety, and social cohesion (Ewing 2006 and Leyden 2003). Many nations and communities now target transportation funding to support greater use of non-motorized modes – both walking and biking (Pucher and Renne 2003). Yet pedestrian-vehicle crashes kill nearly 5,000 persons each year, in the U.S. alone, accounting for over 10 percent of the nation's total roadway fatalities (NHTSA 2009).

Motor vehicle data are regularly tabulated and crash count prediction receives significant research attention (Abdel-Aty and Essam-Radwan 2000, Miaou et al. 2003, Lord 2006, Caliendo et al. 2007, Ma et al. 2008, Austroads 2008, Davies et al. 2005). Somewhat surprisingly, relatively little analytical research has tackled the question of pedestrian-vehicle crash rate prediction, especially at the level of zones or neighborhoods, though pedestrians represent the most vulnerable of road users.

Focusing on neighborhood- or zone-level pedestrian crash counts offers several benefits. Spatially aggregated counts complement more focused pedestrian safety investigations, such as those emphasizing intersection counts (e.g., Weir et al. 2009, Naderan and Shahi 2009, Cottrill and Thakuriah 2010). Zone systems do not neglect any (reported) crashes, and almost two thirds of all U.S. pedestrian-related crashes and 76% of all pedestrian *fatalities* occur *away* from intersections (NHTSA 2009, FHWA 2007). Thus, intersection-based analyses miss over half the population of interest. Focused, site-based analyses have also missed the spatial autocorrelation present in such data, due often to missing variables (such as similar shoulder widths, use of planting strips, similar land use settings and local population demographics, and other spatially-correlated variables typically uncontrolled for). Spatial models work well for zone-based data and can identify such patterns (Morency and Cloutier, 2006).

To this end, this dissertation develops and estimates a new multivariate spatial conditional autoregressive (CAR) model that falls into the family of models explored by Cressie (1995), Banerjee et al. (2004), and Jin et al. (2005). This work analyzes zone-based pedestrian crash counts (for severe and non-severe crashes, separately and simultaneously) over a three-year period in Austin, Texas, while allowing for both observed latent heterogeneity (in zones) and spatial autocorrelation (across zones). A second application demonstrates the same techniques for estimating three types of firm starts (by industry type) across a much larger spatial data set (1,316 U.S. counties).

1.2 A Brief Overview of Existing Methods for Crash Count Prediction

1.2.1 Aspatial Models for Crash Prediction

The traffic crash modeling arena provides many *a*spatial specifications, using Poisson count models, negative binomial specifications (based on a Poisson, with latent heterogeneity in the

rate term, via a gamma distribution), and zero-inflated models (for data sets with zero-crash-rate locations).

All these models neglect spatial interactions among nearby sites. As Tobler's (1970) first law notes, "Everything is related to everything else, but near things are more related than distant things." Disregarding spatial relationships may result in sub-optimal estimates and inferences. For example, parameter estimates are biased when one ignores the spatial autoregressive dependencies across response variables (observed or latent), while estimates are unbiased but inefficient when one ignores spatial autoregressive features of unobserved attributes (in the model's error terms).

1.2.2 Spatial Models

Transport data regularly involves time-series (such as the price of gasoline from year to year) and panel data (such as an individual's mode choices from day to day over a week-long survey). One-dimensional temporal autocorrelation can be complicated to model (e.g., gas price fluctuations from day to day or year to year), but is important to recognize when analyzing time-series data. Two-dimensional spatial autocorrelation can be much more complex to control for, but is relatively routine in transportation data sets (since most observations occur somewhere in space, and many sites are proximate) and typically neglected. Examples of such data sets (and citations of associated spatial analysis) include traffic volumes across a network's links (see, e.g., Wang and Kockelman [2009], Selby and Kockelman [2012]), land development decisions across a region's parcels (Chakir and Parent 2009, Munroe et al. 2002, Wang et al. 2012), and crash prediction across zones and roadway segments (e.g., Levine et al.'s [1995a, 1995b] work on zone-level traffic crashes in Hawaii, and Wang et al.'s [2009] analysis of homogenous road segments).

In the case of count data, Cressie (1991) introduced the auto-Poisson model, a term referring to models in which the mean rate, $\boldsymbol{\lambda}$, involves autocorrelated response variables, i.e., $\boldsymbol{\lambda}=\exp(\mathbf{X}\boldsymbol{\beta}+\rho\mathbf{W}\boldsymbol{y})$. More recently, Griffith (2000a) and Chuan (2008) developed a Poisson-based spatial filtering approach to estimate auto-Poisson models. However, these types of Poisson models permit only negative autocorrelation, an unwanted result arising from the peculiar way spatial autocorrelation enters the specification, as shown in the following equation: $\boldsymbol{\lambda}=\exp(\mathbf{X}\boldsymbol{\beta}+\rho\mathbf{W}\boldsymbol{y})$, where $\boldsymbol{\lambda}$ denotes a vector of expected mean rates, $\mathbf{X}$ is an n by k covariate

matrix, $\boldsymbol{\beta}$ is a *k* by 1 vector of unknown coefficients, $\boldsymbol{y}$ represents a vector of observed (count) responses, **W** an *n* by *n* weight matrix, and ρ the spatial autocorrelation coefficient. In addition, the joint likelihood function under an auto-Poisson assumption requires a non-closed-form solution for the normalizing constant (in order for the joint likelihood function under the auto-Poisson specification to be proper, or integrate to1), which impedes successful estimation (Griffith 2000).

In contrast, Besag's (1975) conditional autoregressive (CAR) model allows both positive and negative spatial autocorrelation structures: $\boldsymbol{\gamma} \sim MVN_n[\boldsymbol{\mu}, (\mathbf{I}-\mathbf{C})^{-1}\mathbf{M}]$, where the column vector $\boldsymbol{\gamma}$ is a stacked version of the *n* spatial random effects (γ_i's), $\boldsymbol{\mu}$ is a vector of the mean of the *n* γ_i's, **I** is an identity matrix, **C**=ρ**W** with W being an *n* by *n* weight matrix defined by contiguity or distance and ρ the spatial autocorrelation coefficient, **M** is a diagonal matrix with $\mathbf{M}_{\text{ii}} = \sigma_i^2$ with σ_{i}^2 indicating the variance specific to location *i*. Miaou et al. (2003) used several variations of a Poisson-based CAR specification to demonstrate the existence of spatial autocorrelation among adjacent roadway segments in their analysis of vehicle crash counts along rural two-lane highways in Texas. Wang et al. (2011) examined traffic congestion's influence on crash counts along 70 homogenous segments of a British expressway, while accounting for both heterogeneity and spatial autocorrelation using a series of Poisson-based CAR models.

A spatial autoregressive (SAR) approach can also be used to analyze spatial data. SAR specifications first appeared in Whittle's (1954) seminal examinations of neighboring plants' growth, as he extended time series autoregressive concepts to the two-dimensional spatial setting. Cressie (1993) has since then proved that the SAR model is a special case of the CAR model, at least in a continuous-response context.Wall (2004) compared implications of SAR and CAR covariance structures using location information across the contiguous 48 U.S. states. She found that both models may sometimes generate very counter-intuitive covariance structures, but she did not offer any theoretical reason for such behaviors. Goodchild and Haining (2004) suggested that the CAR model best applies to geographic regions having more "local" spatial effects, like first-order-neighbor influence, whereas other spatial stochastic processes (which include the SAR and spatial error models [SEMs]) are more suitable for situations with higher-order dependencies, and thus more "global" spatial effects or relationships/interactions. In other words, the CAR model may serve as a spatial version of the Markov process (which requires that

the following state is governed only by its previous state), where a location's response is only directly influenced by its immediate neighbors, rather than neighbors of neighbors (i.e., a second- or higher-order [direct] autocorrelation).

In comparison, the SAR model assumes no Markovian property. Goodchild and Haining's (2004) observation is somewhat reinforced by a simple simulation study done for this dissertation,using a 10 by 10 regular grid, wherein the CAR model's covariance matrix died off noticeably faster than that of the SAR model, indicating stronger, lingering correlation among neighbors under a SAR construction, versus a rather localized spatial correlation under the CAR assumption. The CAR's simpler covariance structure reduces computing burdens and requires less computer memory, thereby facilitating applications, especially in the challenging world of discrete response.

Recent years have seen a strong rise in discrete response model research for spatial settings. The choice of the spatial process depends on assumptions of how spatial autocorrelations emerge: whether spatial dependence (or autocorrelation) occurs across the latent response values (resulting in a SAR specification), the error terms (SEM), or the covariates (producing a spatial Durbin model [SDM], as discussed in Lesage and Pace [2009]). The next section examines the history and limitations of such models.

1.2.3 Limitations of Existing Methods

The existing crash-count-forecasting literature tends to rely on spatial models with an "intrinsic" CAR prior, a term invented by Cressie (1991) for CAR models that do not have a spatial autocorrelation coefficient for their covariance matrices. This prior structure implies a series of conditional Gaussian distributions for each location given the remaining locations, which leads to a closed-form multivariate Gaussian distribution for the joint distribution of response values, based on the factorization theorem (Besag 1975). However, due to the absence of the spatial autocorrelation coefficient, its joint distribution is improper or unbounded in the sample space; therefore, this is often referred to as an intrinsic CAR model, to be distinguished from the *proper* CAR model discussed below (Gelfand and Vounatsou 2003). To circumvent the improper joint posterior issue, Besag et al. (1995) suggested imposing a linear constraint on the spatial random effects at each iteration during the estimation algorithm (often implemented using the Gibbs sampler, a type of Markov chain Monte Carlo sampling technique [Carlin and Louis 2009]).

A more serious concern emerges when the precision (or inverse of the variance) parameter of the intrinsic CAR structure is unknown (which is almost always the case), so that the functional form of the joint distribution of those spatial random effects is not identifiable (via regression methods). In other words, the normalizing constant of the conditional posteriors for the spatial random effects (given the precision parameter) is a function of the precision parameter itself (Cressie 1991). Another concern is that this type of (intrinsic) CAR structure provides no information about the overall spatial autocorrelation, due to the omission of such a coefficient, as follows:

$$\boldsymbol{\gamma} \sim MVN_n[\boldsymbol{\mu}, (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}]$$

where the column vector $\boldsymbol{\gamma}$ is a stacked version of the *n* spatial random effects, γ_i's, (as is the vector $\boldsymbol{\mu}$), **I** is an identity matrix, **C** is an *n* by *n* weight matrix defined by contiguity or distance and $\mathrm{C} = \rho \mathrm{W}$, **W** is a row-standardized weight matrix (i.e., $\mathbf{W} = \left[w_{ij}^*\right]$ and $w_{ij}^* = \frac{w_{ij}}{w_{i+}}$), w_{i+} is the i^{th} row sum of **W**, and **M** is a diagonal matrix with $\mathbf{M}_{\text{ii}} = \sigma_i^2$ (more details about the derivation of this specification can be found in Methodology). For example, if γ_i represents house price at location *i*, then $\boldsymbol{\mu}_i$ denotes the expected value of house price at location *i* given a host of explanatory variables, such as number of rooms, lot size, and gardening investment. The quantity γ_i may denote latent response, such as the expected pedestrian crash rate for zone *i* , with $\boldsymbol{\mu}_i$ representing the systematic crash rate (including covariates such as lane-mile density by roadway class, demographics, and land use attributes) in a Poisson-based model.

In contrast, a *proper* CAR model mitigates the aforementioned concerns by incorporating a spatial autocorrelation coefficient (ρ). This setup is used almost exclusively for univariate-response settings (Pettitt et al. 2002, Wall 2004, Wang et al. 2009). Works that attempt to model multivariate counts include Mardia (1988), who modeled multi-spectral images by casting the question into a series of multivariate conditional distributions, but his work was hindered by computational difficulties (at that time). More recent work by Knorr-Held and Rue (2002) used an improper multivariate CAR structure, and by Gelfand and Vounatsou (2003), who revisited Mardia's specification but still encountered substantial computing times. All CAR model analysts have relied on Markov chain Monte Carlo sampling, a technique commonly employed

in Bayesian estimation and works by sampling sequentially from the MCMC chain (see, e.g., Gelman et al. [2004], Carlin and Louis [2009]).

Some transportation researchers have modeled spatial count data from an *ordered response* perspective (Castro et al. 2012), but such specifications neglect the fundamental data-generating process for *count* data (which are cardinal in nature, not just ordinal), and rely on behaviorally arbitrary threshold values for the latent variable's cut points (to classify the integer responses).

Most breakthroughs in spatial count analysis have been made in biostatistics, where researchers study disease occurrence. It is not yet clear which types of spatial count models will work best when analyzing crash counts, especially area- or zone-level counts. This dissertation explores a more general multivariate CAR model that closely follows Jian et al.'s (2005) proposed specification, but with an added random effect to apture zone-specific (latent) heterogeneity.

1.3 Study Objectives and Organization

The objectives of this work are both theoretical and empirical in nature. This dissertation provides mathematical formulations for and then successfully estimates a two-response spatial multivariate CAR model of pedestrian crash counts across 218 census tracts in Austin, Texas. The application is then extended to a three-response vector of firm births across 1,316 U.S. counties, and guidelines are provided for higher-dimension applications. Spatial analysis of pedestrian crash data is a relative novelty. Covariates include zone-level residential and jobs densities, bus-stop densities (transit access), network features, sidewalk densities, and other demographic and land use characteristics. Bayesian estimation schemes are presented for use of R code, as well as more user-friendly software, such as WinBUGS. The trivariate firm-birth case is provided to showcase the applicability of such models in higher dimensions, across more sites.

The dissertation is divided into five chapters, following this introductory chapter. They are the Literature Review, Methodology, Data Sets, Analysis and Results, and Conclusions. Chapter 2 (Literature Review) synthesizes specifications and techniques employed in crash prediction modeling, along with results that highlight important contributing factors for pedestrian crashes. Chapter 3 (Methodology) focuses on the proposed spatial multivariate CAR models (with two and three response levels, respectively) and the Bayesian sampling schemes used. Chapter 4 (Data Sets) describes data processing for the various explanatory variables and response

variables, with summary statistics provided. Chapter 5 (Analysis and Results) reports and interprets estimation outputs for a simulated (test) data set and Austin's 3-year pedestrian crash counts, with a comparative look at empirical results from a aspatial models (with and without cross-type correlation) and a spatial model without cross-type correlation (i.e., assuming independence of counts by crash type). Chapter 5 ends with the firm-birth (trivariate response) application. Chapter 6 (Conclusions) explains the planning and policy implications for pedestrian safety improvement, and summarizes the work's key contributions from both theoretical and empirical perspectives, while also suggesting several paths forward for new modeling efforts.

1.4 Chapter Summary

This chapter introduced the concept of spatial count models as well as the importance of pedestrian crash modeling, and briefly described relevant existing methods and their limitations. The objectives of this study are to 1) propose and successfully estimate a multivariate CAR count model, to account for cross-count correlations, spatial dependence, and zone-specific heterogeneity, and 2) provide insights for pedestrian-safety planning and policy. A thorough review of competing modeling methods and a discussion of how this work contributes to existing literature and practice are summarized in Chapter 2.

CHAPTER 2: LITERATURE REVIEW

This chapter provides a synthesis of research studies in the field of crash modeling, with an emphasis on the methods commonly employed, including both aspatial and spatial modeling techniques. It also identifies how the work fits within existing literature and allows for important improvements in analysis of spatial count data.

2.1 Aspatial Models for Crash Prediction

Crash analysts have relied on many model specifications and estimation methods. Due to the discrete nature of crash counts (aggregated over time and space, such as a year's worth of crashes along a homogenous roadway segment), continuous-response models are generally not favored (except for highly aggregated data sets, like an entire state's annual crash counts). The Poisson regression model serves as a key starting point for more complex specifications. A Poisson process can describe counts of phenomena with very low occurrence probability (e.g., disease and the occurrence of rare natural disasters). Its application for transportation engineering includes modeling car arrivals under low traffic volume and roadway crashes. The mathematical formulation is expressed as:

$$p(y_i|x_i;\beta) = \frac{{\lambda_i}^{y_i} e^{-\lambda_i}}{y_i!}$$

where $\lambda_i = \exp(x_i\beta)$ is defined as the rate for observation unit i, y_i indicates the observed count over fixed time period and over a fixed length of roadways, and the symbol "!" denotes factorial.

A caveat to employ such model relates to the equi-dispersion assumption where the mean equals the variance, expressed mathematically: $E(y_i) = VAR(y_i) = \lambda_i$. In empirical crash studies, the analysts are more likely to encounter data that exhibit over-dispersion (where the variance is larger than the mean) due to individual heteroscedasticity and unobserved liaisons across observation units. Therefore, Poisson models are often relaxed to the negative binomial or the Poisson-lognormal cases, which allow extra variations in the error terms across individuals.

The negative binomial model differs from Poisson models by adding an error term ε_i whose exponential follows a gamma distribution: $\exp(\varepsilon_i) \sim \text{gamma}(1, \alpha)$, where the parameter α is often referred to as the over-dispersion parameter. The expected sample variance is linked to the

expected sample mean by the equality: $VAR(y_i) = E(y_i) + \alpha[E(y_i)]^2$, which suggests that the variance is no less than the mean and negative binomial models collapse into Poisson models if and only if the over-dispersion parameter $\alpha = 0$. However, the negative binomial model does not apply to situations where under-dispersion is prevalent since by construction $\alpha > 0$, and estimation of the dispersion parameter is problematic when low sample mean and small sample sizes occur (Lord and Mahlawat 2009, Lord and Miranda-Moreno 2008).

The Poisson-lognormal model differs from the negative binomial model by assuming that the error term follows a normal distribution, rather than a gamma distribution (for the exponential of the error term), allowing for more flexibility in describing heterogeneity (and dispersion). Its limitations include more estimation complexity (due to a non-closed form of the Poisson-lognormal distribution) and biasness in the presence of insufficient sample sizes and low sample means (Miaou et al. 2003).

Lord and Mannering (2010) synthesized many model specifications for analyzing crash counts, comparing those mentioned above to zero-inflated models (built under an assumption of a dichotomous process [using a binary probit or logit model]: some locations are crash-free, while others carry a positive crash risk), Conway-Maxwell-Poisson models (which are capable of capturing both under-dispersion and over-dispersion, but are subject to biased estimator issues in the presence of low sample means), a gamma model (which, similar to zero-inflated models, assumes a dual data generating process), generalized estimating equation model, generalized additive models, random-effects, and negative multinomial models (i.e., a negative binomial model with multiple levels of responses that are cross-correlated through the latent error terms). No clear cutoff line can be drawn in terms of which model is superior; the choice of model forms depends on the characteristics of the data and the availability and run-times of computing resources. Nevertheless, they suggest that random parameter model is more easily implemented using MCMC methods, with certain limitations in terms of issues associated with run-times.

There have also been many multivariate crash count studies, to recognize severity levels in a system of simultaneous equations (Song et al. 2006, Ma et al. 2008, and El-Basyouny and Sayed 2009). A simpler way to anticipate counts by severity is to use separate models of injury severity (such as an ordered probit for each crash), conditioned on the total crash count estimate (see, e.g., Carson and Mannering 2001 and Lee and Mannering 2002).

Ongoing advances in crash count modeling and prediction stem from several issues common to such data. For example, zero-inflated Poisson and negative binomial models were developed as a remedy for the preponderance of zeroes in crash data – a phenomenon particularly common for fatal crash counts. Lord and Mannering (2010) argue that a high share of zero counts (which lead to rather low sample mean values) can create biased estimators, as seen in Lord's (2006) small-sample estimate of the negative binomial model's dispersion parameter. The incorrect estimation of dispersion parameters also negatively affects parameter-based inferences. As expected, underreporting of crashes (most common for property-damage-only-type crashes) and missing data also affect estimator consistency and efficiency (Ma 2009).

2.2 Spatial Count Models

2.2.1 Motivation

Spatial models and methods enjoy increasing relevance and opportunity, thanks to advances in geo-referenced data collection and visualization. For example, police crash reports generally have location information, in the form of x-y coordinates using global positioning systems (GPS) and/or the more traditional distance-from-origin (DFO) descriptions common in the past; and many agencies have shifted to sophisticated software (like ESRI's very popular ArcGIS package) to visualize their geo-referenced data (as now required of U.S. state DOTs, for use in the FHWA's Highway Pavement Management System).

A fundamental motivation for a trend toward spatially explicit models lies in the relationships of geographically close observations, due largely to omitted variables (or, in some cases, causal influences). If all influencing factors (such as demographics, topography, rainfall, and so forth) are captured in a stochastic model, one can argue that observations are not related to each other via missing variables, so all error terms are spatially independent. However, it is unrealistic for most analysts to exhaustively characterize and control for every influencing factor. For example, a set of nearby highway segments in an area prone to short but severe storms, which greatly impair drivers' visibility while reducing the roadway's surface friction, experience a higher crash risk. Annual precipitation data only relate to the area's average rainfall conditions. Storm severity and duration variables enter the model via the segment's error terms, inducing spatial autocorrelations for nearby sites.

2.2.2 Development

Earlier work tends to rely on descriptive spatial statistics and aspatial modeling techniques with spatial indicators. Levine et al. (1995a) examined the spatial patterns of Honolulu-motor-vehicle crashes for 1990 by crash types and crash times (i.e., hours of the day, weekdays, and weekends). They concluded that more crashes occurred in the vicinity of employment centers than residential areas and crashes are in general more serious (involving death or severe injury) in suburban and rural areas. Khan et al. (2008) studied weather-related crash counts aggregated at the county level in Wisconsin. They used spatial statistics (e.g., Getis-Ord's G statistic) to identify spatial clusters of crashes and established the link between snow and clusters of weather-related crashes. To gauge the spatial effects within a modeling framework, Shankar et al. (1998) compared the random-effects negative binomial (RENB) and the cross-sectional negative binomial (NB) model results for all median sections longer than 800 meters without median barriers on divided state highways in Washington State. They found that RENB's benefits were notable when spatial and temporal indicator variables were *not* explicitly controlled for in the model's geometric and traffic variables (such as average daily traffic [ADT], maximum shoulder width, access control, and speed limit). The RENB specification lost its advantage when spatial and temporal effects were explicitly specified in the model (using simple indicators for year, route, and the interactions between year and route [i.e., the interaction between time and space]). However, they attempted to allow for spatial correlations in a rather ad-hoc way by employing an indicator spatial variable.

Song et al. (2006) analyzed Texas's county-based crash data using a series of multivariate intrinsic CAR models, with different assumptions on the priors of the spatial random effects. Their work offered statistical insights for model formulation and provided sufficient conditions to assure the propriety of posterior distributions. However, their segments were spatially coarse observational units, and they controlled only for three indicator variables: wet location (to reflect more rainy locations), the presence of horizontal curvature, and obstruction (to indicate roadside conditions). As mentioned in Chapter 1, their intrinsic-CAR specifications do not offer an overall measure of spatial dependence, the spatial autocorrelation coefficient (so it is difficult to examine the significance of spatial dependence using their specifications), and such specifications lead to improper posterior distributions.

Valvade and Jovanis (2008) tested a space-time CAR model (proposed by Bernardinelli et al. [1995]) for county-based fatal crashes in Pennsylvania. They assumed a mean linear time-trend and time-varying coefficients in the logarithm of crash rates: $\log(\theta_{ij}) = \alpha + \sum_k \beta_k x_{ijk} + v_i + u_i + (\varphi + \delta_i)t_j$, where *i* denotes the i^{th} county, *j* the j^{th} time interval, *k* the k^{th} explanatory variable, v_i indicates uncorrelated heterogeneity, u_i captures spatial autocorrelation described by a CAR kernel, φ is the linear time trend, and δ_i captures the interaction between time and county with t_j indicating the time interval *j*. They accounted for county-level demographic (e.g., population, age, and wealth) and weather condition variables (e.g., precipitation and total number of rainy days in a year) as covariates. They estimated that counties with higher shares of persons below the poverty line, young people (ages range from 0 to 24), and elderly people (ages over 64), and a higher road density (lane miles per square mile – which essentially proxies for the exposure/vehicle-miles-travelled term that they did not have have significantly higher crash rates (Precipitation, however, did not appear to be significant.) Spatially correlated structures pose various problems for estimation, as discussed in more detail below (in the Spatial Count Models subsection).

Much work has sought to explicitly recognize spatial dependence in count models. Kaiser and Cressie's (1997) spatial count model assumes that a site's expected or average count μ takes the form: $\mu_i = \exp(\rho \sum_{j \neq i}^{N} w_{ij} y_j + x_i \beta)$, where ρ represents the spatial autocorrelation coefficient, w_{ij} indicates the proximity between locations *i* and *j*, *y* denotes response variables, x_i is a vector of covariates, and β the corresponding coefficients. This form leads to the CAR Poisson model, but with an intractable Leontief inverse and negative spatial autocorrelation coefficient, ρ. Other works include Schabenberger and Pierce's (2002) attempt to use direct representation of error processes, Rasmussen's (2004) CAR model with neighborhood contiguity, eigenvector-based spatial filtering methods for an auto-Poisson process (by Griffith (2002) and Haining et al. (2009)), and Bayesian hierarchical methods (see, e.g., LeSage et al.'s [2007] study on knowledge spillovers using a Poisson spatial interaction model and Flores et al.'s [2009] investigation into relationship between spatial autocorrelation and zero-inflation using ecological data). Among these models, the CAR specification (mostly of the intrinsic variety) has by far enjoyed the most application and investigation for spatial count data analysis (see Wang et al. [2009] and Guo et

al. [2010], and Mariella and Tarantino [2010] for a spatial-temporal model), thanks to relative computational ease and open-source statistical routines.

In analyzing lung cancer risks across Ohio State for four demographic groups (male vs. female, and white vs. non-white), Waller et al. (1997) assumed that latent heterogeneity (represented by a random-effect term ${\theta_i}^t$) and clustering patterns vary across time (i.e., the corresponding spatial error terms are specified for each time period, denoted by ${\phi_i}^t$). They used an expected predictive deviance (EPD) method to compare different reduced forms, and found that proper priors for the heterogeneity error term and space-time error terms can help alleviate identification issues over their space-time model's two error terms. They also acknowledged that the two error terms may be viewed as surrogates for unobserved regional covariates. That is, as more important covariates are considered, the time-space structure may become redundant. In a similar vein, some covariates may have strong collinearity with the spatially correlated error term, making the spatial noise terms difficult to identify and rendering the models difficult to fit.

Waller et al. (1997) began with a univariate version Metropolis-Hasting algorithm, wherein "associated with each parameter was a univariate normal candidate density centered at the current value of this parameter and having some variance". Using an elementary transformation (${\eta_i}^t = {\theta_i}^t + {\phi_i}^t$) and conditional independence assumption (conditioned on η, the observed count responses are iid[1] Poisson distributed), the full conditionals for η and θ were written as: $P(\eta_i|\eta_{-i}, \theta, y) \propto L(\eta_i; y_i)P(\eta_i - \theta_i|\{\eta_{-i} - \theta_{-i}\})$ and $P(\theta_i|\theta_{-i}, \eta, y) \propto P(\theta_i)P(\eta_i - \theta_i|\{\eta_{-i} - \theta_{-i}\})$. Hence, the conditionals of θ no long depend on the data, which they initially regarded as a "serendipitous side benefit" of a normally distributed full conditional – but later recognized as Bayesian unidentifiability (as identified by Eberly and Carlin (2000), and alluded to in Chapter 1). The presence of unidentified parameters through the likelihood has repercussions for the MCMC's convergence rate, as well as convergence monitoring and diagnosis.

Eberly and Carlin (2000) investigated convergence and Bayesian learning using Scotland cancer data set under a CAR framework. The model's individual-level latent heterogeneity (represented by the error terms θ_i) and spatial effects (described by the error terms ϕ_i) capture the amount of extra-Poisson variability allocated to latent heterogeneity and spatial clustering. The

[1] iid stands for "independent and identically distributed".

unidentifiability issue arises when writing out the conditional posterior for θ_i (and ϕ_i), in that the kernel does not depend on the data, as encountered by Waller et al. (1997). In some sense, unidentifiability can be avoided so long as informative priors are judiciously assigned to these two random variables. However, Bayesian estimation methods often rely on vague priors in order to amplify the influences of observed data, rather than let prior assumptions overwhelm parameter estimation and inference. But in this particular case, the variance for θ_i (and ϕ_i) cannot simply be chosen to be arbitrarily large, since then θ_i (and ϕ_i) would be unidentified.

Of course, the sum of these two error terms can indeed be identified. But the purpose of spatial models is often to distinguish such effects. Under this motivation, Eberly and Carlin (2000) examined Bayesian learning behavior for the combined term, $\psi = \frac{SD(\phi)}{SD(\phi)+SD(\theta)}$ (first proposed by Best et al. [1999]), where *SD* denotes the marginal posteriors' empirical standard deviation and $SD(\phi)$ can be approximated using Bernardinelli et al.'s (1995) findings. They maintained that Bayesian learning/identification can still take place for ψ, even under the shadow of unidentified θ_i and ϕ_i. The trick is to use an appropriate scale for the precision parameters (i.e., the inverse of the prior variance) for the heterogeneity and spatial clustering error terms, since the learning pattern can change dramatically under different scale values. It is also of value to investigate the effects on Bayesian learning when using hyperpriors for these precision parameters, rather than using fixed values (as done in their study). They concluded that several factors impact convergence rates, including the selection of starting values, choice of prior distributions, and even the response variable and covariates themselves.

Kim and Lim (2007) specified a multiplicative log-linear mixed model:
$p_{ijk} = Z_i^* \theta_j^* \left(\mu_j^* W_i^*\right)^{(t_K - t)} e_{ijk}^*$, where e_{ijk}* is the exponential of the residual e_{ijk}, Z_i* denotes the effects of the i^{th} county, θ_j* denotes the effects of the j^{th} age group, μ_j* the overall rates of change over time for the j^{th} age group, and W_i* the rates of change in the i^{th} county over time. The model was applied to Missouri state's lung cancer mortality data. They maintained that the assumptions on error structures (e.g., whether e_{ijk}* is assumed to follow a lognormal or gamma distribution) exerts more influence on estimation than the assumptions on spatial patterns (e.g., SAR vs. CAR). The SAR error structure takes the form: $U_i = \rho \sum_j c_{ij} U_j + \tau_i$, where U_i is any spatially correlated random variable, ρ is the spatial autocorrelation parameter, τ_i the uncorrelated white

noise, and $[c_{ij}]$ the adjacency weight matrix. The CAR error structure is expressed as $f(U_i|U_j, j \neq i) = (\frac{1}{2\pi\delta_u})^{1/2}\exp\{-\frac{1}{2\delta_u}(U_i - \rho\sum_{j\neq i} c_{ij} U_j)^2\}$, where δ_u is the variance for the conditional normal distributions and the other parameters defined as they are for the SAR model. Gelman and Rubin's (1992) diagnostics were used to examine the Gibbs sampler's convergence. Kim and Lim (2007) acknowledged difficulty when writing the conditional posteriors for $Z_i^*|Z_j^*$, $j \neq i$ and noted that "contrary to the CAR model, it is difficult to write the conditional distributions for the SAR model in a higher dimension. Most statisticians prefer to use a CAR model." (p. 319) In addition, no exogenous covariates were considered in their multiplicative log-linear mixed model.

Using Bayesian hierarchical modeling scheme, Hoef and Jasen (2007) compared a spatial-time zero-inflated Poisson (ZIP) and hurdle model (for a detailed discussion on the difference between ZIP and hurdle models, see Ridout et al. [1998] and Potts and Elith [2006]) in their analysis of harbor-seal haul-out patterns on glacial ice. Similar to the specifications of Waller et al. (1997), each time period has a separate and independent realization for the random error terms ε_i (for the count model phase) and δ_i (for the binary logit phase), which follow Besag's (1974) CAR specification:

$$\delta_i \sim N(\mathbf{0}, \sigma_\delta^2(\mathbf{I} - \rho_\delta \mathbf{C})^{-1}\mathbf{M})$$

$$\varepsilon_i \sim N(\mathbf{0}, \sigma_\varepsilon^2(\mathbf{I} - \rho_\varepsilon \mathbf{C})^{-1}\mathbf{M})$$

where *C* is an *n* by *n* spatial weight matrix with the ij^{th} element = 1 if the two grids are within 1 km, and then row-standardized; *M* is a diagonal matrix wherein the diagonal elements contain the reciprocal of the number of neighbors. They used diffuse or non-informative priors for all regression parameters, and spatial autocorrelation parameters (ρ_δ and ρ_ε) were assumed constant across time. However, since the model was estimated on a log scale, extremely large parameter values caused computational instability; so they set each regression parameter to have a normally distributed prior with a variance of ten. Their model was estimated using MCMC sampling in WinBUGS software, and the stationarity of parameter draws was evaluated using R's CODA package. Liang et al. (2010) employed a heterogeneous spatio-temporal Poisson process to

analyze major crime data in Cincinnati, using Bayesian methods. They utilized Cressie (1991)'s approach for examining residuals to detect spatial and temporal anomalies.

In a simulation study, Banerjee et al. (2004) showed that the CAR model's ρ term can mislead interpretation of spatial association, and allow for only very limited spatial pattern (with Moran's *I* or Geary's C taking small values, even when ρ gets close to 1). Similarly, in her simulation study for the 48 contiguous U.S. states, Wall (2004) showed how intrinsic-CAR model correlations, among pairs of observations, can change in unintuitive ways.

SAR specifications invite application to analysis of count data. Lambert et al. (2010) proposed a SAR-Poisson model, estimated via a two-step limited-information maximum likelihood (LIML) method. However, they found it hard to generalize the properties of the SAR-Poisson estimator, and detecting AR-lag processes was far from straightforward, given the test statistics used. Using simulated data, they found their estimator performed relatively well in estimating the true autocorrelation, based on size tests. These results may not be too surprising, given that the two-stage estimator applied offers gains in consistency, at the cost of efficiency.

Lambert et al.'s SAR-Poisson model assumed spatial dependence across neighbors' latent rates (λ_i). However, this specification may not explain the data generating process behind traffic crashes well. It is not reasonable to assume the crash rates or counts at one location or on one roadway segment *directly* influence those of neighboring segments (like friends may influence one's consumption patterns), though they are likely correlated, even after controlling for a host of factors. In reality, crash risks correlate in more subtle ways, through associations in their error terms: some unobserved factors (such as climate and topography) cause spatially and temporally correlated error structures, which can be conveyed via a spatial error model (SEM) specification.

McMillen (1992) discussed both SAR and SEM specifications for a binary probit model. He suggests that spatial autocorrelation generally presents heteroskedasticity, reduces OLS estimators' efficiency, and leads to inconsistent OLS estimates. He proposed two categories of estimators for probit models with spatial heterogeneity. One is based on the EM algorithm and is suitable for models with a lagged dependent variable or autoregressive errors. Two disadvantages of these estimators exist: one is computing efficiency, since the inverse of an *n* by

n matrix must be computed in each main iteration. The other disadvantage is that consistent covariance matrix estimates are not readily available.

The second estimator category or estimation method applies to models in which a functional form can be assumed for the heteroskedasticity. An example is a model derived using the spatial expansion method, which is useful in cases where errors have non-constant variance. It is "fairly easy to estimate, requiring only iterated weighted least squares, and can be applied to large data sets" (p. 137). The model generates consistent estimates as long as the form of the heteroskedasticity is specified correctly. The model also produces efficient estimates if the errors are not autocorrelated. Thus, he concluded that the spatial expansion model seems preferable to the SEM and SAR models for "most applications". OLS estimates for the SEM model are consistent, but OLS results in inconsistent estimators for the SAR model. In either model, maximum-likelihood estimates are more *efficient* than OLS estimates. Consistent and efficient estimates are obtained by maximizing the log-likelihood functions for the SEM and SAR models (p. 4 of McMillen [1992]). To aid in evaluating the log-likelihood, McMillen suggested using Ord's (1975) approximation for computing the determinants (which is also known as the normalizing factor: |I-$\rho\times$ W|) as functions of the eigenvalues of *W*: $|I - \rho W| = \prod_i (1 - \rho\omega_i)$. While the SEM and SAR models were designed to help reflect spatial autocorrelation, their implied covariance matrixes have heteroskedastic, not just spatially correlated, error terms. A simulated test showed that the average variance (measured by the average of the diagonal elements of the covariance matrix) increases as ρ increases, and the coefficient of variation (CV) suggests that variance increases too. Also, there is a spatial pattern to the heteroskedasticity, with variances decreasing toward the border of the geographic area under study.

2.3 Pedestrian Crash Predictions

Few tools are available for safety and planning agencies to analyze and forecast pedestrian crashes. Examples include the Pedestrian and Bicycle Crash Analysis Tool packet, which helps analysts identify crash-causing maneuvers while suggesting candidate countermeasures (PBCAT, 2007), and Crossroads software, which serves as a GIS-based database and analysis software for studying pedestrian- and cyclist-involved crashes in the San Francisco Bay Area (Crossroads, 2007).

Weir et al. (2009) studied vehicle-pedestrian injury collisions across 176 San Francisco census tracts, while controlling for local traffic volumes, shares of arterial streets with and without transit service, some land use attributes, population, employment, and residents' income levels. . Their log-linear OLS results suggest that pedestrian injury/fatality counts rise with traffic volumes, shares of arterial streets lacking transit, share of land zoned for neighborhood commercial and mixed residential/neighborhood commercial uses, numbers of residents and (resident) workers, and share of persons living in poverty. land area and proportion of senior residents were not significant crash predictors They did not normalize crash counts by an exposure measure (such as land area or walk-miles traveled), as done here (as discussed in Chapter 4), so many of the effects modeled are size effects (proxying for exposure), which is fundamental to count prediction.

Miranda-Moreno et al. (2011) simultaneously modeled pedestrian activity (in log-linear form) and crash counts (using a standard negative binomial specification) at signalized intersections in the City of Montreal, Canada. They concluded that many built environment, transport system, and traveler attributes (such as land use types, network connectivity, transit supply, and demographic characteristics) in the vicinity of an intersection are strong predictors of pedestrian activity (the exposure variable), but have rather small effects on collision frequency (after controlling for exposure). This result was found here too, as described in Chapters 5 and 6.

2.4 Estimation and Inference Methods

Spatial models with limited dependent variables (like crash counts) tend to be of large dimension, and it is challenging to successfully estimate them. Empirical studies often resort to nonlinear generalized method of moments (GMM) techniques (Klier and McMillen 2008), conditional autoregressive general linear models (Schabenberger and Pierce, 2002), and Bayesian MCMC methods (LeSage et al. 2007). Recent studies also utilize the long-standing composite maximum likelihood (CML) methods which first appeared in Cox's (1975) seminal work and have been revived in Cox and Reid (2004), Varin and Vidoni (2005, 2006, 2009), Varin (2008), and Varin and Czado (2010). The CML approach constructs pseudo-likelihoods by compounding low-dimensional margins (Cox and Reid 2004), in order to achieve computational savings from a minor loss in efficiency. It has been applied to a broad realm of scientific topics, including gene-mapping (Larribe and Lessard 2008), population evolution (Andrieu 2008), and

land use and transportation (Bhat 2011). Despite the reported efficiency gains in empirics, problems may still occur when the analyst is dealing with data set with massive-scale dependence (e.g., in a spatial context, the weight matrix often derives from a large region, with thousands of zones, road segments, and persons), estimation can slow down (Cox and Reid 2004).

Among these estimation methods, the Bayesian MCMC approach appears to enjoy the most applications, thanks to various techniques developed over many years. For example, Damien et al. (1999) described how to sample non-standard posteriors using auxiliary (or latent) variables with case studies for generalized linear mixed models, nonlinear mixed models, and nonlinear random-effects models. Eberly and Carlin (2000) discussed issues surrounding how to properly identify the heterogeneity and clustering error terms in a spatial count model. Kass and Wasserman (1996) offered insights into prior selection, and Best et al. (1999) explored spatially correlated disease and exposure data using Bayesian methods.

Metropolis-Hastings (M-H) algorithms are commonly used in estimating complex spatial models. But these can be difficult to implement and typically require substantial "tuning": data analysts need to judiciously adjust the tuning parameter (i.e., the variance of the proposal distribution) in order to achieve a better mixing of the target distributions and proposal distributions. Many other algorithms have been developed to alleviate non-convergence and improve the robustness of the Gibbs sampler for nonlinear hierarchical models; see, for example, Jungbacker and Koopman's (2007) additional rejection algorithms, the differential evolution MCMC approach of ter Braak (2006), the delayed rejection adaptive Metropolis (DRAM) sampler proposed by Haario et al. (2006), the multiple very fast simulated annealing (MVFSA) algorithm of Villagran et al. (2008), the differential evolution adaptive Metropolis (DREAM) algorithm of Vrugt et al. (2009), the t-walk general-purpose MCMC sampler of Christen and Fox (2010), and the generalized direct sampling (GDS) proposed by Walker et al. (2011). Higdon et al. (2008) noted that relatively simple single-component Metropolis updates can achieve good convergence results and are as efficient as the more complex sampling schemes. The adoption of any sampler depends on the context and is explored in more detail here, in Chapter 3.

An important way to validate any model is to compare its predictions with observed "hold out" data. Hauer (2004a) introduced Cumulative Residuals (CURE) methods for measuring fit of

negative binomial model prediction. CURE methods work by visually examining the cumulative residuals as a function of the independent variable of interest, with a good CURE plot being one oscillating around zero. However, the absolute values of the deviation of predictions from observed values can mask the varying influence of under- and over-prediction. For example, in some cases, over estimating an outcome may have more negative impacts than underestimating an outcome (e.g., over-estimating crash occurrence on a roadway segment may provide false alarm to roadway maintenance departments, but cause no further harm, while under-estimating crashes can divert attention to other segments, leading to unnecessary loss of lives). To this end, an asymmetric loss function can be used to evaluate such model behaviors (Varian 1975, Zellner 1986). Root-mean squared error (RMSE) terms can also be used to compare among models: for applications, see Lambert et al. 2010 (who compared an aspatial Poisson model to a spatial Poisson SAR model).

2.5 Chapter Summary

This chapter synthesizes the various model specifications and estimation techniques employed in the count model literature. Standard aspatial count models include the Poisson, negative binomial, Poisson-lognormal, and zero-inflated family of models. Spatial autoregressive (SAR) and conditional autoregressive (CAR) structures are regularly used to describe spatial dependence. Research shows that estimation differences across the SAR and CAR kernels are not as notable as the differences that result from assumptions made for the heterogeneity term. In addition, spatial count model involving a SAR structure often require formidable computing times.

CHAPTER 3: METHODOLOGY

This dissertation develops a more flexible multivariate conditional autoregressive (MCAR) model, following in the lines of Jin et al. (2005). It extends Jin et al.'s continuous-response model to a count response setting by incorporating a non-Gaussian (Poisson-based) first stage, plus error terms for additional, latent heterogeneity. This chapter first describes the univariate CAR Gaussian model, then describes a restrictive multivariate CAR Gaussian model (as proposed by Gelfand and Vounatsou [2003]), and then introduces the flexible MCAR model.

3.1 Univariate Conditional Autoregressive Models

CAR specifications appear to begin with Besag (1975), and are mostly estimated using Bayesian methods. Conditional distributions of CAR-model response variables are, in most cases, defined by a series of conditional distributions, as shown in Equation 3.1.1 (Cressie 1993).

$$\gamma_i|\gamma_{-i} \sim N\left[\mu_i + \sum_{j=1}^{n} c_{ij}\left(\gamma_j - \mu_j\right),\ \sigma_i^2\right] \tag{3.1.1}$$

where γ_i indicates the spatially autocorrelated variable (e.g., spatial random effects centered at zero, or a response variable -- like traffic flows or household incomes), γ_{-i} denotes such variables at neighboring locations (other than location *i*), μ_i is the expected/mean value of γ_i (i.e., $E(\gamma_i) = \mu_i$), σ_i^2 is the conditional variance, and c_{ij} are weights (either known or unknown) describing the proximity or closeness between locations *i* and *j*.

These conditional distributions lead to a multivariate normal (MVN) joint distribution of the spatially correlated variables (shown in Equation [3.1.2]), based on the factorization theorem (Besag 1975).

$$\boldsymbol{\gamma} \sim MVN_n[\boldsymbol{\mu}, (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}] \tag{3.1.2}$$

where the column vector $\boldsymbol{\gamma}$ is a stacked version of the *n* γ_i's (as is the vector $\boldsymbol{\mu}$), **I** is an identity matrix, **C** is an *n* by *n* weight matrix (defined by site contiguity or inter-observation distances), with $\mathbf{C} = \left[c_{ij}\right]$, and **M** is a diagonal matrix, with $\mathbf{M}_{\text{ii}} = \sigma_i^2$. This joint distribution is used along with the likelihood function of the data set to implement the Gibbs sampler to estimate the posterior distributions of all parameters. Note that the Equations (3.1.1) and (3.1.2) are often referred to as a Markov random field (MRF) because of the way they are derived: achieving a closed-form joint distribution by first specifying a set of conditional distributions (Banerjee et al. 2004).

The validity of the MVN distribution shown in Equation (3.1.2) requires that its covariance matrix, $(\mathbf{I}-\mathbf{C})^{-1}\mathbf{M}$, be symmetric and positive-definite (like any covariance matrix must), thereby necessitating certain constraints on the forms of the matrices $\mathbf{C}$ and $\mathbf{M}$. For example, one may let $\mathbf{C}=\rho\mathbf{W}$ and $\sigma_i^2=\frac{\sigma^2}{w_{i+}}$, where ρ is referred to as the spatial autocorrelation coefficient, $\mathbf{W}$ is a row-standardized weight matrix (i.e., $\mathbf{W}=\left[w_{ij}^*\right]$ and $w_{ij}^*=\frac{w_{ij}}{w_{i+}}$), and w_{i+} is the i^{th} row sum of $\mathbf{W}$.

The CAR specification permits contiguity and distance-based weight matrices, but precludes the K^{th}-nearest-neighbor weighting scheme because such weights violate the symmetry condition. First-order contiguity weights are defined such that $w_{ij}=1$ if *i* and *j* share a common border (else $w_{ij}=0$), and $\mathbf{W}$'s diagonal elements are all zeros by construction (Cressie 1991). As alluded to in Chapter 2, this type of CAR model is called a *proper* CAR model, and is commonly estimated using Bayesian techniques in the open-source WinBUGS software package (Spiegelhalter 2003), where "BUGS" stands for Bayesian inference Using Gibbs Sampling.

As discussed in Chapter 2, the "intrinsic" CAR model does not have a spatial autocorrelation coefficient for its covariance matrix, so it has just one parameter, σ^2, to describe the spatial attributes of data (e.g., the strength of spatial dependence and the variation of spatial dependence). This can lead to counterintuitive interpretations: e.g., when σ^2 (or the unscaled variance term in the conditional distribution) is small, the spatially-correlated effect is strongly dependent on the neighboring values. However, the overall contribution to the mean is small (Spiegelhalter 2003). The intrinsic CAR model is not used here, and should not be used by others.

3.2 Multivariate Conditional Autoregressive Models

The first multivariate CAR model was discussed in Mardia (1988). Similar to the univariate CAR setting, it was formulated as a series of full conditional distributions under the MRF assumption:

$$p\left(\boldsymbol{\gamma}_i\middle|\boldsymbol{\gamma}_{j\neq i},\boldsymbol{\Gamma}_i^{-1}\right)=N\left(\mathbf{Q}_i\sum_{i\sim j}\mathbf{B}_{ij}\boldsymbol{\gamma}_j,\boldsymbol{\Gamma}_i^{-1}\right) \tag{3.2.1}$$

where $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{ip})'$ denotes a p×1 vector of spatial random effects at location i (and p is the number of response types -- like k=2 for children's heights and weights, respectively, or k=3 for passenger car, SUV, and truck indicators), $\mathbf{Q}_i$ is a k×k matrix describing the overall spatial strength of the k types, $\mathbf{B}_{ij}$ is a p×p matrix of exogenous weights across different response types across locations, and $\boldsymbol{\Gamma}_i^{-1}$ is the covariance matrix capturing remaining correlations between the p types of. Analogous to the univariate case, the joint distribution can be derived using Brook's Lemma (Banerjee et al. 2004).

$$p(\boldsymbol{\gamma}) = N\left(\mathbf{0}, \left[\boldsymbol{\Gamma}(\mathbf{I} - \mathbf{B}_Q)\right]^{-1}\right) \tag{3.2.2}$$

where the np×1 vector $\boldsymbol{\gamma}$ is a stacked version of the n $\boldsymbol{\gamma}_i$'s, $\mathbf{B}_Q$ is an np×np matrix with $\left(\mathbf{B}_Q\right)_{ij} = \mathbf{Q}_i\mathbf{B}_{ij}$, and by construction $\left(\mathbf{B}_Q\right)_{ii} = 0$. $\boldsymbol{\Gamma}$ is an np×np block diagonal matrix: $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\Gamma}_i)$. Note that for Equation (3.2.2) to exist, the covariance matrix, $\left[\boldsymbol{\Gamma}(\mathbf{I} - \mathbf{B}_Q)\right]^{-1}$, must again be symmetric and positive definite (Gelfand and Vounatsou 2003).

A variety of MCAR models arise from Equations (3.2.1) and (3.2.2) depending on different parameterizations of $\boldsymbol{\Gamma}$ and $\mathbf{B}_Q$, which govern the propriety of the likelihood function. For example, one may assume that $\mathbf{Q}_i = \rho \cdot \mathbf{I}_{p\times p}$ across locations, where the scalar ρ measures the overall level of spatial autocorrelation, and $\boldsymbol{\Gamma} = \mathbf{D}\otimes\boldsymbol{\Lambda}$, where $\mathbf{D}$ is usually a diagonal matrix, $\mathbf{D} = diag(m_i)$, with m_i denoting the i^{th} row sum of the n×n weight matrix (defined using contiguity or distance, though the former is more common in empirical studies, probably due to the computational benefits of sparse matrices); and $\boldsymbol{\Lambda}$ is a p×p matrix capturing the non-spatial correlations among the p response types at any location and must be positive definite and symmetric. Under these parameterizations, the MCAR model can be expressed as:

$$p(\boldsymbol{\gamma}) = N\left(\mathbf{0}, \left[\left(\mathbf{D}(\mathbf{I} - \rho\mathbf{B})\right)\otimes\boldsymbol{\Lambda}\right]^{-1}\right) \tag{3.2.3}$$

where $\mathbf{B}$ is an n× n row-standardized weight matrix, $\mathbf{B} = \mathbf{D}^{-1}\mathbf{W}$, and the weight matrix $\mathbf{W}$ can be defined by contiguity or (inverse) distance.

The *intrinsic* MCAR specification will emerge is one assumes $\rho = 1$. Although the symmetry condition holds, so long as $\mathbf{W}$ and $\boldsymbol{\Lambda}$ are symmetric, the covariance matrix is singular when

$\rho = 1$ because $\mathbf{D}(\mathbf{I} - \mathbf{B}) \cdot \mathbf{1} = \mathbf{0}$. This model is dubbed intrinsic because the positive definite criterion can be omitted. The *proper* MCAR model results when $\rho \in (-1, 1)$; typically, $\rho \in [0, 1)$ since negative spatial autocorrelation is rare. This model was used in Gelfand and Vounatsou's (2003) analysis of children's height and weight data, and in Carlin and Banerjee's (2003) work.

Equations (3.2.1) through (3.2.3) are derived when arranging the individual spatial random effects in a way such that $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1p}, \ldots\ , \gamma_{n1}, \gamma_{n2}, \ldots, \gamma_{np})'$. Alternatively, these *np* random effects can be grouped by response types (Jin et al. [2005]), leading to the following form (for *k*=2):

$$\begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} (\mathbf{D} - \rho\mathbf{W})\Lambda_{11} & (\mathbf{D} - \rho\mathbf{W})\Lambda_{12} \\ (\mathbf{D} - \rho\mathbf{W})\Lambda_{12} & (\mathbf{D} - \rho\mathbf{W})\Lambda_{22} \end{pmatrix}^{-1} \right) \tag{3.2.4}$$

where $\boldsymbol{\gamma}_1 = (\gamma_{11},\ \gamma_{21}, \ldots, \gamma_{n1})'$ and $\boldsymbol{\gamma}_2 = (\gamma_{12},\ \gamma_{22}, \ldots, \gamma_{n2})'$ encompass all the spatial random effects for response types 1 and 2, respectively, across the *n* locations; $\boldsymbol{\Lambda} = [\Lambda_{ij}]$, *i*, *j*=1, 2, describes the non-spatial correlations between the two types (e.g., cancer types, traffic crash types) at any given locations; and **W** serves as the unnormalized weight matrix; with remaining parameters defined as above.

As Jin et al. (2005) noted, it is not logical to use the same spatial autocorrelation coefficient ρ throughout covariance matrix, since different observation types are likely to exhibit somewhat different spatial clustering patterns. An intuitive improvement is to specify *three* distinct spatial coefficients, one for each response type and one for their interaction terms, resulting in a new form of the covariance matrix appearing in Equation (3.2.4):

$$\boldsymbol{\Sigma} = \begin{pmatrix} (\mathbf{D} - \rho_1\mathbf{W})\Lambda_{11} & (\mathbf{D} - \rho_3\mathbf{W})\Lambda_{12} \\ (\mathbf{D} - \rho_3\mathbf{W})\Lambda_{12} & (\mathbf{D} - \rho_2\mathbf{W})\Lambda_{22} \end{pmatrix}^{-1} \tag{3.2.5}$$

Alas, it is difficult to evaluate the positive definiteness for such a flexible covariance matrix and the resulting model is often hard to implement via Markov chain Monte Carlo estimation (Jin et al. 2005). Thus, a tradeoff is made here, to allow only two distinct spatial autocorrelation coefficients, as proposed by Carlin and Banerjee (2003) and Gelfand and Vounatsou (2003).

They utilized matrix decomposition methods to parameterize the "precision matrix" (the inverse of the covariance matrix, $\boldsymbol{\Sigma}$) in a way such that $\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \mathbf{R}_1'\mathbf{R}_1\Lambda_{11} & \mathbf{R}_1'\mathbf{R}_2\Lambda_{12} \\ \mathbf{R}_2'\mathbf{R}_1\Lambda_{12} & \mathbf{R}_2'\mathbf{R}_2\Lambda_{22} \end{pmatrix}$, where $\mathbf{R}_k$ is an upper-triangular matrix computed using either Cholesky or spectral decomposition, and $\mathbf{R}_k'\mathbf{R}_k = \mathbf{D} - \rho_k\mathbf{W}$, *k*=1, 2. In other words, the spatial autocorrelation coefficients for the off-diagonal elements are determined as a function of the diagonal elements' spatial autocorrelation coefficients. However, different MCAR models can result from the same covariance matrix because the decomposition of $(\mathbf{D} - \rho_k\mathbf{W})$ is not unique (Jin et al. [2005]), which may cause the model to be unidentified.

3.3 A Flexible MCAR Model

When successfully specifying a MCAR structure, an important consideration is the validity of the joint covariance matrix's inverse. This precision matrix needs no inversion and so is faster to compute than the covariance matrix itself, and the computation can rely on several techniques -- like the decomposition methods employed by Carlin and Banerjee (2003) and Gelfand and Vounatsou (2003). However, working directly with the precision matrix, instead of the covariance matrix, often obscures the interpretation of the correlation structure of the phenomenon under study (Jin et al. 2005). A judiciously designed *covariance* matrix allows one to incorporate more behavioral realism, while ensuring the resulting model's estimability. Jin et al. (2005) proposed a "generalized" MVCAR model by working directly with the covariance matrix. Their two-response-level model (*k*=2) is expressed as:

$$\begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right) \tag{3.3.1}$$

where $\boldsymbol{\gamma}_k$ contains the spatial random effects across *n* locations for a given response type *k* (with *k* =1 and 2), and $\boldsymbol{\Sigma}_{kl}$ represents $n \times n$ covariance matrices (*k*, *l* =1, 2). Standard multivariate-normal theory (? Theory or equations?) leads to the following formulation:

$$\boldsymbol{\gamma}_1|\boldsymbol{\gamma}_2 \sim \mathrm{N}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\gamma}_2, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}')$$

$$\boldsymbol{\gamma}_2 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{22}) \tag{3.3.2}$$

For ease of presentation, let $\mathbf{A} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$, $\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}'$. Therefore, Equation (3.3.2) can also be written as (Jin et al. 2005):

$$\boldsymbol{\gamma}_1|\boldsymbol{\gamma}_2 \sim \mathrm{N}(\mathbf{A}\cdot\boldsymbol{\gamma}_2,\ \boldsymbol{\Sigma}_{11\cdot 2})$$

$$\boldsymbol{\gamma}_2 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{22})$$

Given Equations (3.3.2) and (3.3.3), the joint distribution of $\boldsymbol{\gamma}$ can be expressed as:

$$\begin{pmatrix}\boldsymbol{\gamma}_1\\ \boldsymbol{\gamma}_2\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{11\cdot 2} + \mathbf{A}\cdot\boldsymbol{\Sigma}_{22}\mathbf{A}' & \mathbf{A}\cdot\boldsymbol{\Sigma}_{22}\\ (\mathbf{A}\boldsymbol{\Sigma}_{22})' & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right) \tag{3.3.3}$$

Equation (3.3.3) allows one to write the joint distribution as $p(\boldsymbol{\gamma}) = p(\boldsymbol{\gamma}_1|\boldsymbol{\gamma}_2)\cdot p(\boldsymbol{\gamma}_2)$, which exists as long as the covariance matrices are symmetric and positive-definite. The conditions that ensure this property are that $\boldsymbol{\Sigma}_{11\cdot 2}$ and $\boldsymbol{\Sigma}_{22}$ are positive definite (Harville [1997], as cited in Jin et al. [2005]). The crux of the problem is then to specify the matrices $\mathbf{A}$, $\boldsymbol{\Sigma}_{11\cdot 2}$, and $\boldsymbol{\Sigma}_{22}$, which will uniquely determine the functional form of the covariance matrix of the joint distribution for all response variables, as shown in Equation (3.3.3).

Jin et al. (2005) assumed that $\mathrm{var}(\boldsymbol{\gamma}_1|\boldsymbol{\gamma}_2) = \boldsymbol{\Sigma}_{11\cdot 2} = [(\mathbf{D} - \rho_1\mathbf{W})\tau_1]^{-1}$ and $\mathrm{var}(\boldsymbol{\gamma}_2) = \boldsymbol{\Sigma}_{22} = [(\mathbf{D} - \rho_2\mathbf{W})\tau_2]^{-1}$, with τ_1 and τ_2 serving as scale parameters. Intuitively, the covariance structure of $\boldsymbol{\phi}_2$ is independent of type 1's spatial autocorrelation coefficient, and $\boldsymbol{\phi}_2$'s mean values are centered around zero (as shown in Equation [3.3.2]). Likewise, the conditional covariance of $\boldsymbol{\phi}_1$ does not depend on type 2's spatial autocorrelation coefficient. However, the conditional mean of $\boldsymbol{\phi}_1$ is a weighted average of $\boldsymbol{\phi}_2$ and $\mathbf{A}$ serves as a transformation matrix. $\mathbf{A}$ is the final undetermined quantity, needed to uniquely identify the covariance matrix of the full conditional distribution in Equation (3.3.3). Jin et al. (2005) assumed that $\mathbf{A} = \eta_0\mathbf{I} + \eta_1\mathbf{W}$ and $E(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2) = (\eta_0\mathbf{I} + \eta_1\mathbf{W})\boldsymbol{\phi}_2$, with scalars η_0 and η_1 dubbed the bridging parameters. The term "bridging" is used because it associates ϕ_{i1} with ϕ_{i2} and ϕ_{j2}. In other words, this type of MCAR model treats the conditional mean of $\boldsymbol{\phi}_{i1}$ at a given location *i* as a weighted average of neighboring $\boldsymbol{\phi}_{j2}$ values along with a scaled $\boldsymbol{\phi}_{i2}$ value at its own location, *i*. They also prove that the proper MCAR model, shown in Equation (3.2.4) is a special case of the MCAR model developed in Equations (3.3.1) through (3.3.3); it emerges when assuming $\alpha_1 = \alpha_2 = \alpha$ and

$\eta_1 = 0$, as in Gelfand and Vounatsou's (2003) model. In such cases, the covariance matrix's positive-definiteness property is ensured when $|\alpha_1| < 1$ and $|\alpha_2| < 1$.

Jin et al. (2005) applied the multivariate CAR model to standardized mortality ratios (SMRs) as the continuous response variables. Its applicability in a non-Gaussian (first-stage) setting, with crash counts, for example, has not been tested until now, as the focus of this dissertation.

3.4 A Poisson Log-Normal MCAR Model

Jin et al.'s (2005) MCAR structure is adopted here, while incorporating a Poisson first-stage link function with added region-specific heterogeneity. Rather than having to transform the aggregated counts to continuous response (like Jin et al.'s SMR values), this work's log-normal MCAR model directly analyzes spatial count data (common in the study of transportation and other systems), while accounting for region-specific heterogeneity. Here, a new, Poisson log-normal MCAR model will be applied to analyze area-level pedestrian crash count data in Travis County, as well as county-level firm births across the country. The following paragraphs discuss this new model's formulation and sampling scheme, in the context of zone-level pedestrian crash counts with two response variables (*k*=1 for fatal and severe injury crashes, and *k*=2 for light or no injury pedestrian crashes).

3.4.1 Model Specification

The first stage is expressed as a Poisson process:

$$y_{ik} \sim \text{Poisson}(\lambda_{ik}) \tag{3.4.1}$$

where y_{ik} is the observed pedestrian crash count by severity level (*k*=1, 2) for the i^{th} polygon/zone in Travis County, and the mean crash rates of the second-stage, λ_{ik}, represent the (continuous) expected crash counts:

$$\ln(\lambda_{ik}) = \ln(E_{ik}) + x_i'\beta_k + \phi_{ik} + u_i \tag{3.4.2}$$

where E_{ik} is an exposure term (like walking-miles traveled in each zone), which may be a function of local employment and population (e.g., $E_{ik} = EMP^a \cdot POP^b$); $\boldsymbol{x}_{ik}$ is a vector of zone- and crash-type-specific covariates; $\boldsymbol{\beta}_k$ is a vector of parameter coefficients, specific to each outcome type *k*; and ϕ_{ik} represents the spatial random effect defined by the MCAR structure described earlier. The heterogeneity error term, u_i, captures zone-specific heterogeneity

or latent variation that is not explained by spatial effects and is often assumed to follow an iid normal distribution, $u_i \sim N(0, \sigma_u^2)$, leading to the Poisson-lognormal spatial model. Alternatively, its exponential term may take on a gamma distribution, $\exp(u_i) \sim Gamma(\theta, \theta)$, leading to a negative binomial model (Miaou et al. 2003).

The parameter ϕ_{ik} is defined such that: $\boldsymbol{\phi}_2 \sim N(\mathbf{0}, [(\mathbf{D} - \alpha_2\mathbf{W})\tau_2]^{-1})$ and $\boldsymbol{\phi}_1|\boldsymbol{\phi}_2 \sim N(\mathbf{A}\boldsymbol{\phi}_2, [(\mathbf{D} - \alpha_1\mathbf{W})\tau_1]^{-1})$, where τ_1, τ_2, α_1, α_2, **W**, **D**, and **A** are defined in the previous section.

Analogous to the spatial random effects ϕ_{ik}, which are zero-centered, as shown in Equation (3.4.2), the average logarithmic crash rates, $\ln(\lambda_{ik})$, can be expressed using an MCAR structure. The only difference between ϕ_{ik} and $\ln(\lambda_{ik})$ is that that the latter's mean value is no longer centered at zero, but rather $[\ln(E_{ik}) + x_i'\beta_k + u_i]$. For ease of presentation, column vectors $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are used here to substitute for these latent continuous values: $\boldsymbol{Z}_1 = \ln(\boldsymbol{\lambda}_1) = [\ln(\lambda_{11}),\ \ln(\lambda_{21}), \ldots, \ln(\lambda_{n1})]'$ and $\boldsymbol{Z}_2 = \ln(\boldsymbol{\lambda}_2) = [\ln(\lambda_{12}),\ \ln(\lambda_{22}), \ldots, \ln(\lambda_{n2})]'$. In this case, the conditional distributions for $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are multivariate normal and expressed as:?

$$\boldsymbol{Z}_1|\boldsymbol{Z}_2 \sim N(\ln(\boldsymbol{E}) + \boldsymbol{X}\cdot\boldsymbol{\beta}_1 + \boldsymbol{u} + (\eta_0\mathbf{I} + \eta_1\mathbf{W})(\boldsymbol{Z}_2 - \ln(\boldsymbol{E}) - \boldsymbol{X}\cdot\boldsymbol{\beta}_2 - \boldsymbol{u}),\ [\tau_1(\mathbf{D} - \alpha_1\mathbf{W})]^{-1})$$

$$\boldsymbol{Z}_2 \sim N(\ln(\boldsymbol{E}) + \boldsymbol{X}\cdot\boldsymbol{\beta}_2 + \boldsymbol{u}, [\tau_2(\mathbf{D} - \alpha_2\mathbf{W})]^{-1}) \quad (3.4.3)$$

where $\boldsymbol{E}$ is an *n* by 1 vector of exposure values (with any unknown parameters to be estimated), $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are column vectors specific to each of the two crash types, $\boldsymbol{X}$ is the covariance matrix (with the i^{th} row being the observed explanatory variables, including a constant term, for location/neighborhood) *i*, and $\boldsymbol{u}$ is a vector of the *n* site-specific error terms: $\boldsymbol{u} = (u_1, \ldots, u_n)'$.

The two "bridging" parameters, η_0 and η_1, associate ϕ_{i1} with ϕ_{i2} and with ϕ_{j2} ($j \neq i$), respectively. The parameter η_0 captures the relationship between the spatial random effects of each region's severe (including fatal and incapacitating injury crashes) and non-severe (i.e., non-incapacitating injury, light, possible, or no injury crashes), while η_1 links neighboring zones' influences across the two crash types. (Note: for simplicity, η_1 can be set to zero, letting the spatial autocorrelation coefficients in the covariance structures capture the interactions among neighboring regions.)

The spatial autocorrelation coefficients α_1 and α_2 describe the spatial dependence for the two crash types respectively and should lie within the range $[\frac{1}{\min}, \frac{1}{\max}]$ for the covariance matrix, $[\tau_2(\mathbf{I} - \alpha_2\mathbf{D}^{-1}\mathbf{W})]^{-1}\mathbf{D}^{-1}$, to be positive definite and thus invertible (Jin et al. 2005), where max and min denote the maximum and minimum eigenvalues of the weight matrix, $\mathbf{D}^{-1}\mathbf{W}$. Note that the matrix, $\mathbf{D}^{-1}\mathbf{W}$, is row-standardized (i.e., normalized) by construction. Negative spatial dependence is rare, so the lower bound on α_1 and α_2 is often set to 0; the maximum eigenvalue of a row-standardized weight matrix is guaranteed to be 1. **D** is a diagonal matrix with the i^{th} diagonal element representing the i^{th} row sum of **W**. The precision parameters τ_1 and τ_2 scale the covariance structures in order to do capture any noise that is not being captured by the covariance matrix.

3.4.2 Sampling Scheme

Having specified the conditional distributions of the average crash rates at each location, $\boldsymbol{Z} = (\boldsymbol{Z}_1,\ \boldsymbol{Z}_2)$, the focus is now on the joint posterior distribution of all model unknowns: $p(\beta, \boldsymbol{u}, Z, \tau, \alpha, \eta|Y) \propto L(Y|\beta, \boldsymbol{u}, Z, \tau, \alpha, \eta) \cdot \pi(\beta) \cdot \pi(\boldsymbol{u}) \cdot \pi(Z) \cdot \pi(\tau) \cdot \pi(\alpha) \cdot \pi(\eta)$. Each of these components, for use in the MCMC process of draws, is developed and discussed below.

The posterior distribution, $p(\beta, \boldsymbol{u}, Z, \tau, \alpha, \eta|Y)$:

$$p(\beta, \boldsymbol{u}, Z, \tau, \alpha, \eta|Y) \propto L(Y|\ Z) \cdot p(Z|\beta, \boldsymbol{u}, \tau, \alpha, \eta) \cdot \pi(\beta) \cdot \pi(\boldsymbol{u}) \cdot \pi(\tau) \cdot \pi(\alpha) \cdot \pi(\eta)$$

$$\propto L(Y|\ Z) \cdot p(Z_1|Z_2, \beta, \boldsymbol{u}, \tau, \alpha, \eta) \cdot p(Z_2|\beta, \boldsymbol{u}, \tau, \alpha, \eta) \cdot \pi(\beta) \cdot \pi(\boldsymbol{u}) \cdot \pi(\tau) \cdot \pi(\alpha) \cdot \pi(\eta)$$

$$\propto \left(\prod_{k=1}^{2}\prod_{i=1}^{n} Z_{ik}^{y_{ik}} \cdot e^{-Z_{ik}}\right) \cdot \tau_1^{\frac{n}{2}} \cdot |D - \alpha_1 W|^{\frac{1}{2}} \cdot \exp\left\{-\frac{\tau_1}{2}[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]' \cdot (D - \alpha_1 W)[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]\right\} \cdot \tau_2^{\frac{n}{2}} \cdot |D - \alpha_2 W|^{\frac{1}{2}} \cdot \exp\left\{-\frac{\tau_2}{2}(\boldsymbol{Z}_2 - \boldsymbol{m}_2)'(D - \alpha_2 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)\right\} \cdot e^{\sum_i u_i} \cdot [\text{Gamma}(\theta, \theta)]^n \cdot [\text{Gamma}(1, 0.1)]^2 \cdot [\text{Unif}(0, 1)]^2[\text{N}(0, 100)]^2 \quad (3.4.4)$$

where $\boldsymbol{m}_1 = X'\boldsymbol{\beta}_1 + \ln(\boldsymbol{E}) + \boldsymbol{u}$ and $\boldsymbol{m}_2 = X'\boldsymbol{\beta}_2 + \ln(\boldsymbol{E}) + \boldsymbol{u}$.

Here, response-type-specific covariates $\beta = (\beta_1, \beta_2)$ are assumed to follow a flat normal prior, centered around zero with a large variance term: $\beta_1 \sim N(\mathbf{0}, 10^5 I)$ and $\beta_2 \sim N(\mathbf{0}, 10^5 I)$. The precision parameters $\tau = (\tau_1, \tau_2)$ are assumed to follow a rather diffuse Gamma distribution:

Gamma$(1, 0.1)$ with mean 10 and variance 100. Spatial autocorrelation coefficients, $\alpha = (\alpha_1, \alpha_2)$, are assigned a uniform prior over the interval (0, 1), denoted by Unif$(0, 1)$. The two "bridging" parameters η_0 and η_1 follow a diffuse normal prior, $N(0, 10^2)$.

Conditional distributions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

Assuming a diffuse prior $\boldsymbol{\beta}_1 \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\beta}_2 \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = 10^5\mathbf{I}$:

$$p(\boldsymbol{\beta}_1|\cdot) \propto \exp\Big\{-\frac{\tau_1}{2}[\boldsymbol{Z}_1 - X'\boldsymbol{\beta}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u} - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})]'(D - \rho_1 W)[\boldsymbol{Z}_1 - X'\boldsymbol{\beta}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u} - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})]\Big\} \cdot \exp\Big\{-\frac{1}{2}\beta_1'\boldsymbol{\Sigma}^{-1}\beta_1\Big\}$$

$$\propto \exp\Big\{\Big[\boldsymbol{\beta}_1 - \tfrac{1}{2}{\Lambda_1}^{-1}\Omega_1\Big]' \cdot \Lambda_1 \cdot \Big[\boldsymbol{\beta}_1 - \tfrac{1}{2}{\Lambda_1}^{-1}\Omega_1\Big]\Big\}$$ (using the completing-the-squares technique

$$\propto N\Big(\tfrac{1}{2}{\Lambda_1}^{-1}\Omega_1, {\Lambda_1}^{-1}\Big)$$

where $\Lambda_1 = -\frac{\tau_1}{2}X(D - \rho_1 W)X' - \frac{1}{2}\boldsymbol{\Sigma}^{-1}$ and $\Omega_1 = -\frac{\tau_1}{2}[2X(D - \rho_1 W)(\boldsymbol{Z}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u}) - X(D - \rho_1 W)(\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})]$.

$$p(\boldsymbol{\beta}_2|\cdot) \propto \exp\Big\{-\frac{\tau_1}{2}\Big[-(\boldsymbol{Z}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u})'(D - \rho_1 W)(\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u}) + (X'\boldsymbol{\beta}_1)'(D - \rho_1 W)(\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u}) - (\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})'(\eta_0 I + \eta_1 W)'(D - \rho_1 W)\big(\boldsymbol{Z}_1 - X'\boldsymbol{\beta}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u} - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})\big)\Big]\Big\} \cdot \exp\Big\{-\frac{\tau_2}{2}[(\boldsymbol{Z}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})'(D - \rho_2 W)(-X'\boldsymbol{\beta}_2) - (X'\boldsymbol{\beta}_2)'(D - \rho_2 W)(\boldsymbol{Z}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u}) + (X'\boldsymbol{\beta}_2)'(D - \rho_2 W)(X'\boldsymbol{\beta}_2)]\Big\}$$

$$\propto \exp\Big\{\Big[\boldsymbol{\beta}_2 - \tfrac{1}{2}{\Lambda_2}^{-1}\Omega_2\Big]' \cdot \Lambda_2 \cdot \Big[\boldsymbol{\beta}_2 - \tfrac{1}{2}{\Lambda_2}^{-1}\Omega_2\Big]\Big\} \propto MVN\Big(\tfrac{1}{2}{\Lambda_2}^{-1}\Omega_2, {\Lambda_2}^{-1}\Big)$$

where $\Lambda_2 = \frac{\tau_2}{2} X(D - \rho_2 W)X' - \frac{\tau_1}{2} X(\eta_0 I + \eta_1 W)'(D - \rho_1 W)(\eta_0 I + \eta_1 W)X' - \frac{1}{2}\boldsymbol{\Sigma}^{-1}$ and $\Omega_2 = -[\tau_1((\boldsymbol{Z}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})'(\eta_0 I + \eta_1 W)'(D - \rho_1 W)(\eta_0 I + \eta_1 W)X' - (\boldsymbol{Z}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u})'(D - \rho_1 W)(\eta_0 I + \eta_1 W)X' + \beta_1' X(D - \rho_1 W)(\eta_0 I + \eta_1 W)X') + \tau_2(\boldsymbol{Z}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})'(D - \rho_2 W)X']$.

Conditional distributions of $u_1, u_2, \ldots, u_n$:

$$p(\boldsymbol{u}|\cdot) \propto \exp\left\{-\frac{\tau_1}{2}[\boldsymbol{Z}_1 - X'\boldsymbol{\beta}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u} - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})]'(D - \rho_1 W)[\boldsymbol{Z}_1 - X'\boldsymbol{\beta}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u} - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})] - \frac{\tau_2}{2}(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})'(D - \rho_2 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})\right\} \cdot e^{\sum_i u_i}.$$

It is difficult to draw $\boldsymbol{u} = u_1, u_2, \ldots, u_n$ simultaneously. The conditional posteriors of u_i values do not follow any known distribution and so cannot be sampled using Gibbs' method. The Metropolis-Hastings algorithm (Metropolis et al. 1953, Carlin and Louis 2009) and a more recent development, the generalized direct sampling (GDS) method (Walker et al. 2011), can be used for such draws.

$$p(u_i|\cdot) \propto \exp\left\{-\frac{\tau_1}{2}[\boldsymbol{Z}_1 - X'\boldsymbol{\beta}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u} - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})]'(D - \rho_1 W)[\boldsymbol{Z}_1 - X'\boldsymbol{\beta}_1 - \ln(\boldsymbol{E}) - \boldsymbol{u} - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})] - \frac{\tau_2}{2}(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})'(D - \rho_2 W)(\boldsymbol{Z}_2 - X'\boldsymbol{\beta}_2 - \ln(\boldsymbol{E}) - \boldsymbol{u})\right\} \cdot e^{u_i} \quad (i=1, 2, \ldots, n)$$

Conditional distribution of $\boldsymbol{Z}_1$:

$$p(\boldsymbol{Z}_1|\cdot) \propto \left(\prod\nolimits_{i=1}^{n} Z_{i1}^{y_{i1}} \cdot e^{-Z_{i1}}\right) \cdot \exp\left\{-\frac{\tau_1}{2}[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]' \cdot (D - \rho_1 W)[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]\right\}$$

where $\boldsymbol{m}_1 = X'\boldsymbol{\beta}_1 + \ln(\boldsymbol{E}) + \boldsymbol{u}$ and $\boldsymbol{m}_2 = X'\boldsymbol{\beta}_2 + \ln(\boldsymbol{E}) + \boldsymbol{u}$. Due to the model's non-Gaussian first stage (thanks to integer responses [e.g., crash counts]), the conditional posterior of $\boldsymbol{Z}_1$ does not follow a known form.

Conditional distribution of $\boldsymbol{Z}_2$:

$$p(\boldsymbol{Z}_2|\cdot) \propto \left(\prod_{i=1}^{n} Z_{i2}^{y_{i2}} \cdot e^{-Z_{i2}}\right)$$
$$\cdot \exp\left\{-\frac{\tau_1}{2}[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]'\right.$$
$$\cdot (D - \rho_1 W)[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]$$
$$\left. -\frac{\tau_2}{2}(\boldsymbol{Z}_2 - \boldsymbol{m}_2)'(D - \rho_2 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)\right\}$$

Like the conditional posterior for $\boldsymbol{Z}_1$, the conditional posterior $p(\boldsymbol{Z}_2|\cdot)$ does not follow a standard distribution.

Conditional distribution of τ_1:

$$p(\tau_1|\cdot) \propto \tau_1^{n/2} \exp\left(-\frac{\tau_1}{2} \cdot T_1\right) \propto \text{Gamma}\left(\frac{n}{2} + 1, \frac{T_1}{2}\right)$$

where $T_1 = [\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]' \cdot (D - \rho_1 W)[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]$.

Conditional distribution of τ_2:

$$p(\tau_2|\cdot) \propto \tau_2^{n/2} \exp\left(-\frac{\tau_2}{2} \cdot T_2\right) \propto \text{Gamma}\left(\frac{n}{2} + 1, \frac{T_2}{2}\right)$$

where $T_2 = (\boldsymbol{Z}_2 - \boldsymbol{m}_2)'(D - \rho_2 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)$.

Conditional distribution of ρ_1:

$$p(\rho_1|\cdot) \propto |D - \rho_1 W|^{\frac{1}{2}}$$
$$\cdot \exp\left\{-\frac{\tau_1}{2}[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]'\right.$$
$$\left. \cdot (D - \rho_1 W)[\boldsymbol{Z}_1 - \boldsymbol{m}_1 - (\eta_0 I + \eta_1 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)]\right\}$$

Conditional distribution of ρ_2:

$$p(\rho_2|\cdot) \propto |D - \rho_2 W|^{\frac{1}{2}} \cdot \exp\left\{-\frac{\tau_2}{2}(\boldsymbol{Z}_2 - \boldsymbol{m}_2)'(D - \rho_2 W)(\boldsymbol{Z}_2 - \boldsymbol{m}_2)\right\}$$

Conditional distribution of η_0:

$$p(\eta_0|\cdot) \propto \exp\left\{-\frac{\tau_1}{2}[(\boldsymbol{Z}_2-\boldsymbol{m}_2)'(\eta_0 I+\eta_1 W)'(D-\rho_1 W)(\eta_0 I+\eta_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2) - 2(\boldsymbol{Z}_1-\boldsymbol{m}_1)'(D-\rho_1 W)(\eta_0 I+\eta_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2)]\right\}$$

For aspatial models with multivariate correlation (i.e., $\eta_1 = 0$) , $p(\eta_0|\cdot)$ is then written as:

$$p(\eta_0|\cdot) \propto \exp\left\{-\frac{\tau_1}{2}[{\eta_0}^2(\boldsymbol{Z}_2-\boldsymbol{m}_2)'(D-\rho_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2) - 2\eta_0(\boldsymbol{Z}_1-\boldsymbol{m}_1)'(D-\rho_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2)]\right\}$$

$$\propto \exp\left\{-\frac{\tau_1}{2}(\boldsymbol{Z}_2-\boldsymbol{m}_2)'(D-\rho_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2)\left(\eta_0-\frac{(\boldsymbol{Z}_1-\boldsymbol{m}_1)'(D-\rho_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2)}{(\boldsymbol{Z}_2-\boldsymbol{m}_2)'(D-\rho_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2)}\right)^2\right\}$$

$$\propto N\left(\frac{(\boldsymbol{Z}_1-\boldsymbol{m}_1)'(D-\rho_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2)}{(\boldsymbol{Z}_2-\boldsymbol{m}_2)'(D-\rho_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2)}, [\tau_1(\boldsymbol{Z}_2-\boldsymbol{m}_2)'(D-\rho_1 W)(\boldsymbol{Z}_2-\boldsymbol{m}_2)]^{-1}\right)$$

Note: An aspatial model (with cross-type correlations) will have $\eta_1 = 0$, $\rho_1 = 0$, and $\rho_2 = 0$.

3.4.3 A Trivariate Case

Analogous to the two-response MCAR model, a trivariate MCAR model assumes that the spatial random effects are represented as $\boldsymbol{\phi} = \left(\boldsymbol{\phi}'_1,\ \boldsymbol{\phi}'_2,\ \boldsymbol{\phi}'_3\right)'$. In a crash-count context, $\boldsymbol{\phi}_i$ could be the *n* by 1 vectors of spatial random effects for the latent rates of crash types 1 (e.g., fatal and incapacitating-injury crashes), 2 (e.g., non-incapacitating injury crashes), and 3 (possible and no injury cases)**.** A question then emerges as to the sequence of these conditional distributions. One way to settle such a question is to try all 6 possible arrangements and choose the model with best goodness-of-fit.

For ease of exposition, assume the sequence of conditional distributions as such: $p(\boldsymbol{\phi}) = p(\boldsymbol{\phi}_1|\ \boldsymbol{\phi}_2,\ \boldsymbol{\phi}_3) \cdot p(\boldsymbol{\phi}_2|\ \boldsymbol{\phi}_3) \cdot p(\boldsymbol{\phi}_3)$. Based on multivariate normal theory, the joint distribution of $\boldsymbol{\phi}$ takes the form: $\begin{pmatrix}\boldsymbol{\phi}_1\\ \boldsymbol{\phi}_2\\ \boldsymbol{\phi}_3\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\\ \boldsymbol{\mu}_3\end{pmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12} & \Sigma_{13}\\ \Sigma_{12}' & \Sigma_{22} & \Sigma_{23}\\ \Sigma_{31}' & \Sigma_{32}' & \Sigma_{33}\end{bmatrix}\right)$, where the *n* by 1 vector $\boldsymbol{\mu}_p$ indicates the mean for response type *p* (*p* = 1, 2,3), Σ_{pl} is an *n* by *n* matrix describing the covariance structure between response types *p* and *l*. The marginal distribution of $\boldsymbol{\phi}_3$ can be written as: $p(\boldsymbol{\phi}_3) \sim \boldsymbol{N}(\boldsymbol{\mu}_3, \Sigma_{33})$ and assume $\boldsymbol{\mu}_3 = \mathbf{0}$ and $\boldsymbol{\Sigma}_{33} = \left[\tau_3(\mathbf{D}-\rho_3\mathbf{W})\right]^{-1}$. The marginal

distribution of $(\boldsymbol{\phi}_2, \boldsymbol{\phi}_3)$ can be obtained by removing irrelevant elements (with respect to $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$) from the full distribution and follows a multivariate normal distribution:
$\begin{pmatrix}\boldsymbol{\phi}_2\\ \boldsymbol{\phi}_3\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{\mu}_2\\ \boldsymbol{\mu}_3\end{pmatrix}, \begin{bmatrix}\Sigma_{22} & \Sigma_{23}\\ \Sigma'_{23} & \Sigma_{33}\end{bmatrix}\right)$.

$\boldsymbol{\phi}_2|\boldsymbol{\phi}_3 \sim N\left(\mathrm{A}_{23}\boldsymbol{\phi}_3,\ [(D-\rho_2 W)\tau_2]^{-1}\right)$, where A_{23} describes the aspatial correlation between response types 2 and 3, as well as the spatially-lagged correlation between the two response types, formally: $\mathrm{A}_{23} = \eta_{0,23} I + \eta_{1,23} W$.

$\boldsymbol{\phi}_1|\boldsymbol{\phi}_2, \boldsymbol{\phi}_3 \sim N\left(\mathrm{A}_{13}\boldsymbol{\phi}_3 + \mathrm{A}_{12}\boldsymbol{\phi}_2,\ [(D-\rho_1 W)\tau_1]^{-1}\right)$, where A_{13} and A_{12} capture the aspatial and spatially-lagged correlation across response types 1 and 3, and response types 1 and 2, formally: $\mathrm{A}_{13} = \eta_{0,13} I + \eta_{1,13} W$ and $\mathrm{A}_{12} = \eta_{0,12} I + \eta_{1,12} W$.

3.4.4 Chapter Summary

This chapter discusses the specification of the new multivariate spatial count model and explains the behavioral realism that the model conveys. The chapter extends the multivariate spatial structure (that first appeared in Jin et al. 2005 in a continuous response setting) to a count model setting with site-specific error terms. Bayesian estimation technique is the most common approach used to address complex spatial models like this thanks to its ability to uncover non-closed-form likelihood function. Bayesian sampling scheme is also provided, along with the conditional posteriors associated with the proposed model.

CHAPTER 4: DATA SETS

This chapter discusses the two data sets used in this dissertation to demonstrate application of the new MCAR model for count responses, as described in Chapter 3. The first data set consists of a 3-year aggregate of pedestrian crash counts with two-level response over 218 neighborhoods in Austin, Texas. The other contains new-firm counts across 1,316 contiguous U.S. counties in the 2008 period and is used to demonstrate the model's application in cases involving a larger sample and three (rather than two) response levels.

4.1 Pedestrian Safety Data Set

The city of Austin, the capital of Texas, is located within a medium-sized urban region. Part of Travis County, with a county population of close to 1 million people, Austin has a fair amount of pedestrian activity, thanks to generally sunny conditions, a large college student population, and a walk-friendly culture. The county's 3-year pedestrian crash counts (reflecting all severity levels, between 2007 and 2009, as reported by police) were aggregated using ArcGIS's *spatial join* function over Thiessen polygons built around each census tract's centroid.

The Thiessen (or Voronoi) polygon is determined by the distances to a set of objects (e.g., points or polygons) in a two- or higher-dimension space. Its mathematical derivation can be found in Aurenhammer (1991), although its history traces back to the 19th Century thanks to the seminal work of Johann Peter Gustav Lejeune Dirichlet, a German mathematician. The distance from all points within a polygon to the original, corresponding Census tract's centroid is less than any point outside that polygon. Thiessen polygons are commonly used in epidemiology to find correlations between infections and contributing factors, and in ecology to study species distributions. To the best of the author's knowledge, this study is the first to utilize Thiessen polygons in the area-level crash analysis, which can be computed using the *proximity* function in ESRI's ArcGIS program.

This work relies on the 218 Thiessen polygons, rather than the original census tracts, to ensure that high-crash locations, regularly along tract edges (important roadways) and often at tract corners (important intersections), can be uniquely assigned to a polygon zone rather than be unassigned or arbitrarily assigned to adjacent tracts. By default, ArcGIS creates Thiessen polygons based on a given set of polygons (or their centroids) within a rectangular area that covers the given geographic area, resulting in several unreasonably large polygons at the

periphery of Austin's Travis County. For this reason, the County's boundary file was used to cut away the excess portions of the polygons, to yield a new boundary that closely follows the County's true shape as shown in Figure 4.1.

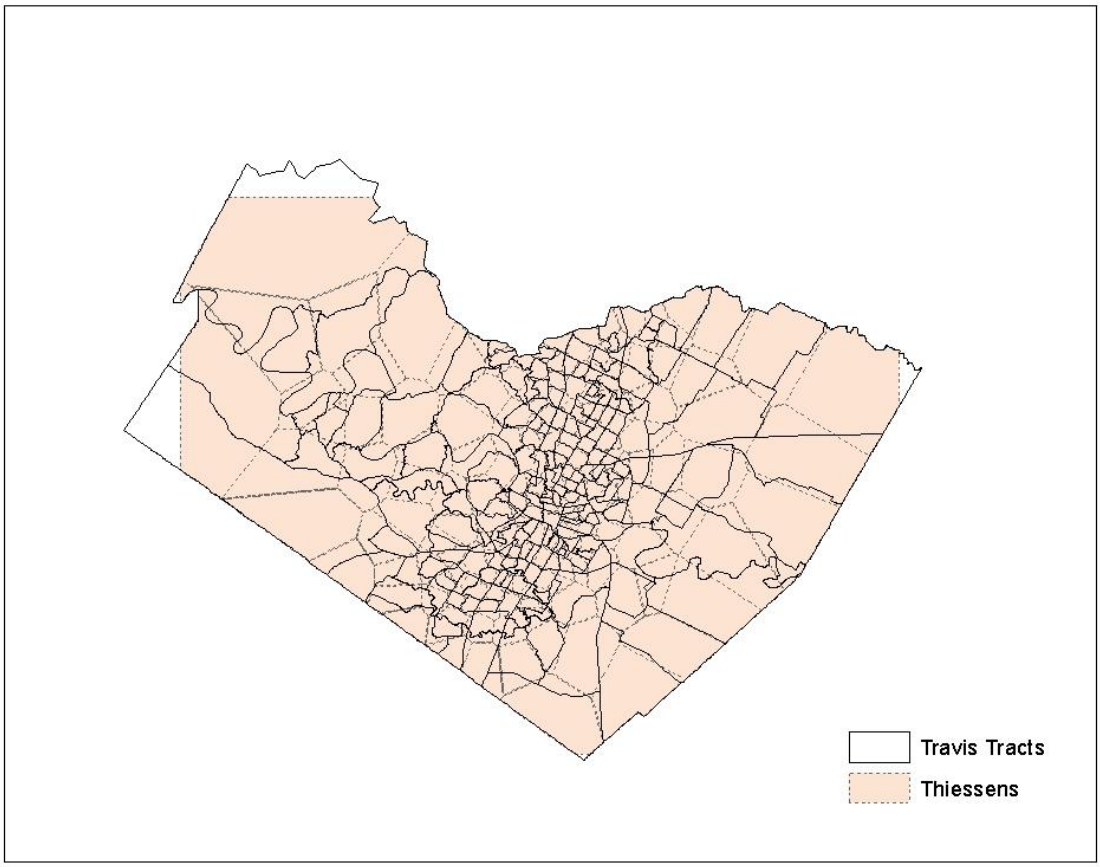


Figure 4.1: Thiessen polygons overlaid on Travis County's 218 Census Tracts

As described in Chapter 3, both contiguity- and distance-based weight matrices (**W**) are explored here. Contiguity (of a certain order) can be determined using the *spdep* library in R, an open-source statistical package. Euclidean distance weights were computed using R code developed here, based on the coordinate information for the polygon's centroids obtained via the *Feature-to-Point* routine in ArcGIS's Data Management Toolbox. The unnormalized weight matrix (**D**) was then computed such that $D_{ij} = \frac{1}{d_{ij}}$, if $d_{ij} < d_{\max}$, and $D_{ij} = 0$ if $d_{ij} \geq d_{\max}$, where $d_{\max}$ is the maximum distance defining a polygon's neighborhood. As noted earlier, the nearest-neighbor

scheme cannot be used because it yields an asymmetric weight matrix, thus violating the CAR model's necessary symmetry condition.

4.1.1 Transit Stop Density

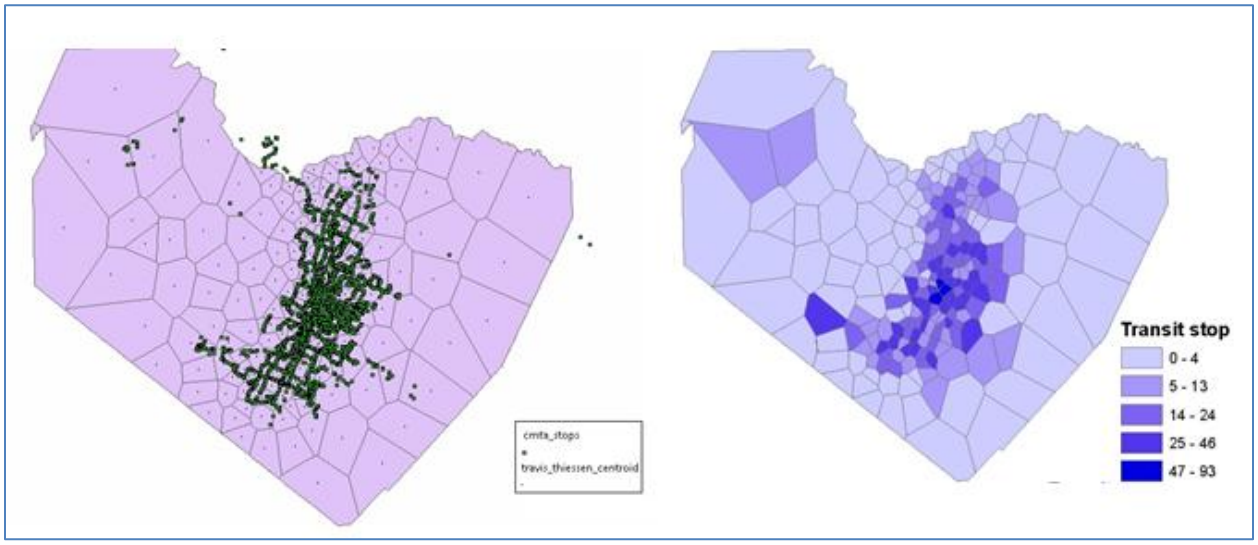


Figure 4.2: Capital Metro Transit Stop Locations (left panel) and Counts by Polygon (right panel)

Information on transit stop locations for January 2012 was obtained through Austin Capital Metro's online database, and is shown in Figure 4.2. While stop locations may differ over the 2007-2009 period (when the crash data were obtained), it is unlikely the changes are significant and will affect this analysis. There are 2,680 transit stops within Travis County's boundary (out of a total of 2,732 Capital Metro stop locations), with an average of 12 stops per polygon (and a standard deviation of 13). The maximum number (93) of transit stops is found in a polygon just south of Austin's downtown.

4.1.2 Land Use

The influence of land use patterns and the built environment on pedestrian safety has been well documented (Dumbaugh 2005, Dumbaugh and Rae 2009). Such factors affect walking frequency (which is tightly linked to pedestrian exposure), traffic volumes (vehicle exposure), and environmental complexity (which influences the general likelihood of collisions) (see, e.g., Clifton et al. [2008] and Miranda-Moreno et al. [2011]). Clifton et al. (2004) have also showed

how areas with high transit access are associated with much higher pedestrian crash rates, and with crashes involving more children.

Several land use variables are controlled for here, in the models of crash counts. These include land use balance or entropy, the percentage of residential parcels that are close to transit stops, and the percentage of residential parcels that are close to commercial activities. Such variables were developed using the City of Austin's 2006 land use map. 2006 was chosen because it immediately precedes the 2007-2009 data period, and covers all of Travis County, whereas the next available map, from 2008, only records land use patterns around the center of the county. Hence, this study uses the 2006 map with the assumption that relatively few locations experience significant land use changes over a two-year period (2007-2009).

The land use balance terms is developed using an entropy measure, like that used in Cervero and Kockelman (1997): $LU\ Balance = -\sum_{k=1}^{4} p_k \ln(p_k) / \ln 4$

where p_k is the proportion of each land use *k* (residential, commercial, office, and industry uses) in the polygon, and an even or uniform balance (25% of land in each of the four categories) yields the largest entropy value of 1. Figure 4.3 illustrates the spatial distribution of these resulting entropy values across the Travis County polygons. Interestingly, more balanced land use patterns are evident in the western region.

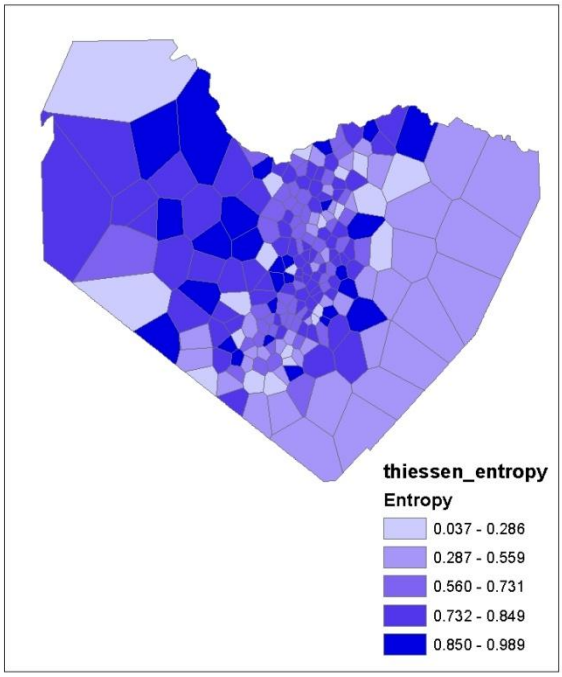


Figure 4.3: Entropy Values for Thiessen Polygons across Travis County

Another relevant variable is the proximity of residential parcels to transit stops and service, which can relate to pedestrian exposure and vehicle-pedestrian interactions, along with some bus-induced sight obstructions that can facilitate pedestrian-vehicle crashes (Clifton et al. 2008). This study computed the share of each polygon's residential parcels (number of single-family dwelling units [SFDUs] plus multi-family parcels [mostly apartment buildings]) that lie within one-half mile of a transit stop for each polygon using ArcGIS's *buffer* and *spatial join* functions. Here, multi-family parcels include group quarter, duplexes, apartments, and condos (as defined by the City of Austin's land use archive).

4.1.3 Access to Schools

A positive association between the presence of children (under 14 years of age) and pedestrian crashes also has been established in the literature (see, e.g., Clifton et al. [2004] and NHTSA [2011]). Plausible causes include children's shorter stature (making them harder for motorists to see), their (often) less-developed sense of motion, and an inexpert ability to judge traffic conditions and signal lights. This study computed the share of residential parcels (within each

polygon) within one-half mile of K-12 schools, which may proxy this particularly vulnerable population. The only available data year for the Texas Education Agency's school locator (http://wgisprd.tea.state.tx.us/sdl/) is/was 2010, so we assumed that school locations remained constant over the 4-year period (from 2007 to 2010).

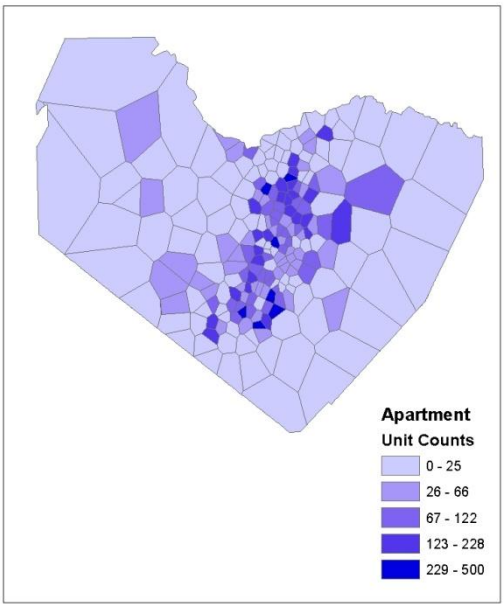


Figure 4.4: Apartment Parcel Counts within One-Half Mile of Schools across Thiessen Polygons.

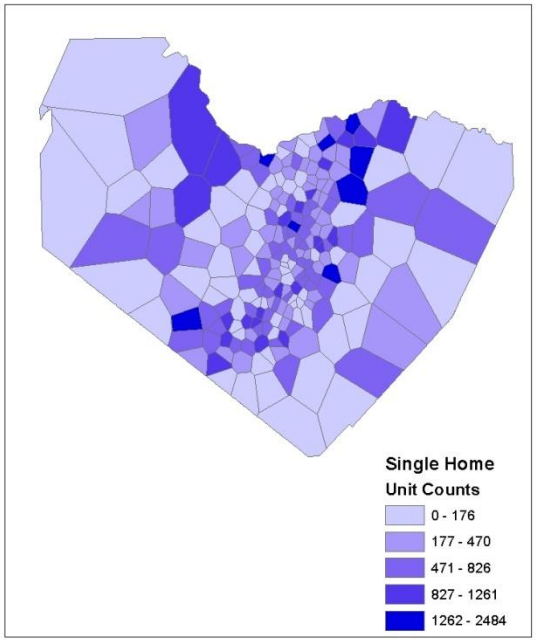


Figure 4.5: Single-Family Dwelling Unit Counts within One-Half Mile of Schools across Thiessen Polygons.

4.1.4 Roadway Features

Lane-miles (by functional class of roadway) affect vehicle exposure and traffic conditions. The Capital Area Metropolitan Planning Organization's (CAMPO's) coded 2005 network was used to extract roadway information for Travis County. This network file provides the number of lanes, travel speeds from travel demand modeling estimates, and vehicle counts (imputed from travel demand modeling, rather than real/observed counts) for roadways across all CAMPO-coded links (for purposes of demand modeling), and so does not include local streets. Census TIGER/Line® data, as shown in Figure 4.6, complement CAMPO's coded network by providing all local street information, but lack traffic counts and land counts.

Figure 4.6 A Snapshot of Travis County's TIGER/Line® Network (2008), across Thiessen Polygons.

CAMPO's three types of coded roadways are (1) freeways (including interstates, expressways, associated direct connectors and ramps), (2) arterials (including major and minor arterials, and frontage roads), and (3) collectors (but only 36 links were coded as collectors).

Local streets were identified using the 2008 TIGER/Line map. Such streets typically have a single lane of in each direction, so we assume that all local streets in the expanded network have two lanes. Given the small number (36 links) of collectors, local streets and collectors were grouped together as "local streets". Thus, in the end lane-miles were computed/estimated for the following roadway classes: freeway, arterials, and local streets. Traffic count data and link lengths are used to gauge vehicle miles traveled (VMT) by roadway class. The VMT variable did not show up as significant in the walk-miles-traveled model.

A sidewalk map also was obtained, from the City of Austin's transportation archive, and the lengths of these were aggregated across polygons (with two sidewalks, on either side of a single street, counting twice). Table 4.2 provides summary statistics for all the variables discussed above.

Total vehicle miles traveled (VMT) were computed using flow values on all links lying within each polygon, times the length of each of those links (within each polygon). These data come from CAMPO's 2010 *coded* network links, which offer estimated flow rates (from travel demand model forecasts) for 203 lane-miles of freeways, 1,061 lane-miles of arterials, and 3,231 lane-miles of local streets. It is important to note that CAMPO networks omit most local streets and many collector roadways, so the VMTs estimated here serve only as a rough (and biased-low) estimate of the true VMT in each polygon. Moreover, these volume measures were not statistically (or practically) significant covariates in the crash-count models, and so were removed from the final model specifications.

Table 4.2: Summary Statistics of Covariates and Response Variables across Thiessen Polygons (n=218)

	Mean	*Std Dev*	*Min*	*Max*
Transit Access				
% SFDU[a] near Transit (1/2 mi.)	0.628	0.433	0	1
% APT[b] near Transit (1/2 mi.)	0.655	0.432	0	1
Transit Density (# of bus stops per sq. mile)	13.66	17.57	0	98.6
Land Use				
Land Use Entropy	0.647	0.229	0.037	0.989
% Residential near Commercial (within 1/2 mi.)	0.759	0.304	0	1
Network Density				
LnMiDenFWY	4.228	6.435	0.000	44.430
LnMiDenART	8.836	6.783	0.104	51.207
LnMiDenLOC	2.435	3.770	0.000	18.932
Sidewalk Density	6.718	6.076	0.000	28.851
Population & Employment *Density (# per sq. mi) (2007)* [c]				
Population Density	2,470	2,611	5	15,633

Basic Emp. Density	356	653	0	5,137
Retail Emp. Density	235	279	0	1,842
Service Emp. Density	598	762	1	5,308
Access to Schools				
% SFDU near K-12 schools (within 1/2 mi.)	0.514	0.352	0	1
% APT near K-12 schools (within 1/2 mi.)	0.487	0.386	0	1
Vehicle Miles Traveled (Traffic Volume × Length)				
VMTFWY	159,208	293,354	0	1,525,955
VMTART	322,692	332,826	937	3,616,047
VMTLOC	13,785	29,345	0	245,631
Exposure Measure				
Walk-Miles Traveled (WMT[d]) (in miles)	4978.6	6740.5	20	50,132
Response Variable				
Severe Crash Count per polygon (Fatal & Incapacitating Crashes, 2007- 2009)	0.89	1.53	0	15
Non-Severe Crash Counts per polygon (Incapacitating, Possible Injury, & No Injury Crash Counts, 2007-2009)	3.23	7.4	0	100

Notes: [a.]SFDU stands for single-family dwelling units, including single family and large-lot single family dwelling units; [b.]APT denotes apartments (e.g., group quarter, duplex, apartment/condo defined by the City of Austin's land use archive); [c.].population and employment densities are computed as the estimated counts (by overlaying traffic-analysis-zone-level count information obtained from CAMPO) divided by polygon size; [d] WMT is the crash exposure measure, estimated using 2006 Austin travel survey and least squares regression (with details provided in this paper's Results section).

4.2 County-level Firm Growth Data

As mentioned earlier, a U.S. county-level firm birth data set (across three industry types, in 2008/2009) was also used in this thesis, to showcase the applicability of the proposed model in a relatively large sample setting (N_{obs}=1,316) with a third dimension to for the multivariate response values. Firm births are characterized by the number of new establishments, which are available annually from the Statistics of U.S. Businesses, as provided by the U.S. Census Bureau (http://www.census.gov//econ/susb/data/download_susb2009.html). The North American Industry Classification System (NAICS) code groups these business establishments into basic, retail, and service categories, based on the first two digits of NAICS' 2007 definition, as shown in Table 5.12. Basic jobs refer to those involved in industries driven largely by external demands, from outside the region (Quintero 2007). These include agriculture, manufacturing, and

wholesale trade, since many products are transported outside the source regions. Non-basic industries mostly sell their goods and services locally, and include, for example, local grocery stores and restaurants. These were grouped into retail and service firms here, as listed in Table 4.3.

Table 4.3: NAICS 2007 Definition of Industry Classification and General Codes.

NAICS Code	NAICS Industry Description	General Class
11	Agriculture, forestry, fishing & hunting	1
21	Mining, quarrying, & oil & gas extraction	1
22	Utilities	3
23	Construction	3
31-33	Manufacturing	1
42	Wholesale trade	1
44-45	Retail trade	2
48-49	Transportation & warehousing	1
51	Information	3
52	Finance & insurance	3
53	Real estate & rental & leasing	3
54	Professional, scientific, & technical services	3
55	Management of companies & enterprises	3
56	Administrative support, waste management & remediation services	3
61	Educational services	3
62	Health care & social assistance	3
71	Arts, entertainment, & recreation	3
72	Accommodation & food services	3
81	Other services (except public administration)	3

Note: General classes are basic (class 1), retail (class 2), and service (class 3) industries. More NAICS information can be found at http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2007.

The number of new establishments (in basic, retail, and service industries) in each U.S. county in from 2008 to 2009 serves as the response vector here. Figures 4.7 through 4.11 illustrate the spatial distribution of the three new-establishment counts, and Figures 4.12 through 4.15 illustrate the spatial profile of total employment counts by industries.

Covariates evaluated include population density (persons per acre of land in the county), share of African American population, share of vacant housing units, median age of resident, share of families living in poverty, median annual income (for persons older than 16 years of age), and an indicator variable to distinguish metropolitan counties from non-metro areas (a definition based on a rural-urban continuum coding [USDA 2003]). Unemployment rates and educational attainment may also influence economic grown and thus new-firm starts, but these variables are only estimated in the American Community Survey for about 800 of the nation's 3,200 counties and so were not controlled for here.

County sizes, existing employment, and firm count (number of establishments) serve as the exposure measures for the rate of new-firm starts (with more details provided in Chapter 5). The data used to generate all these variables come from the U.S. Census 2010 summary files, which focus on social, economic, and housing characteristics collected on the Census short form (summary files 1 and 2) and the long-form questionnaire (from a sample of 19 million housing units, about 1 in 6 households, as provided in summary files 3 and 4).

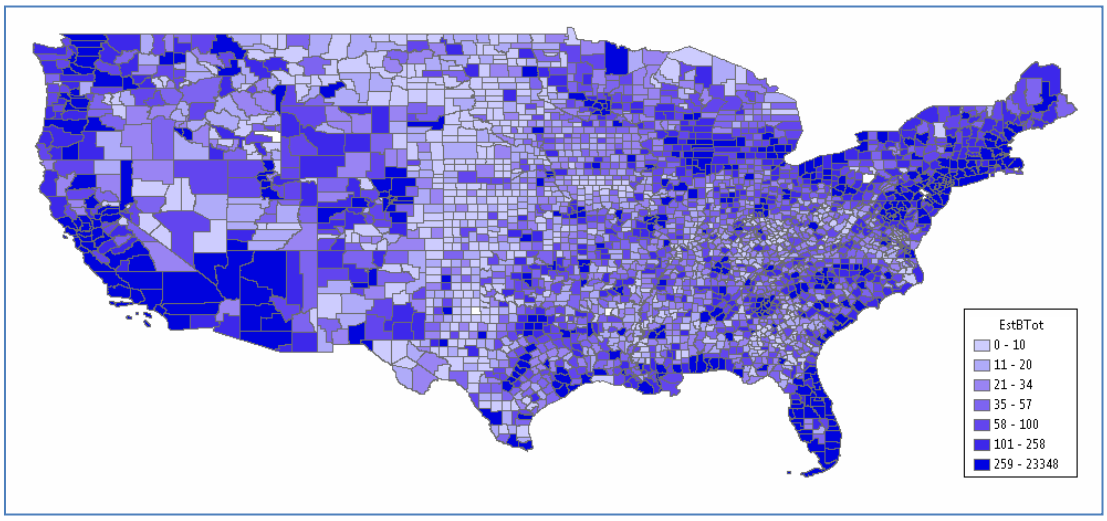


Figure 4.7 a): New Establishments across U.S. Counties in the Lower 48 States during 2008-2009.

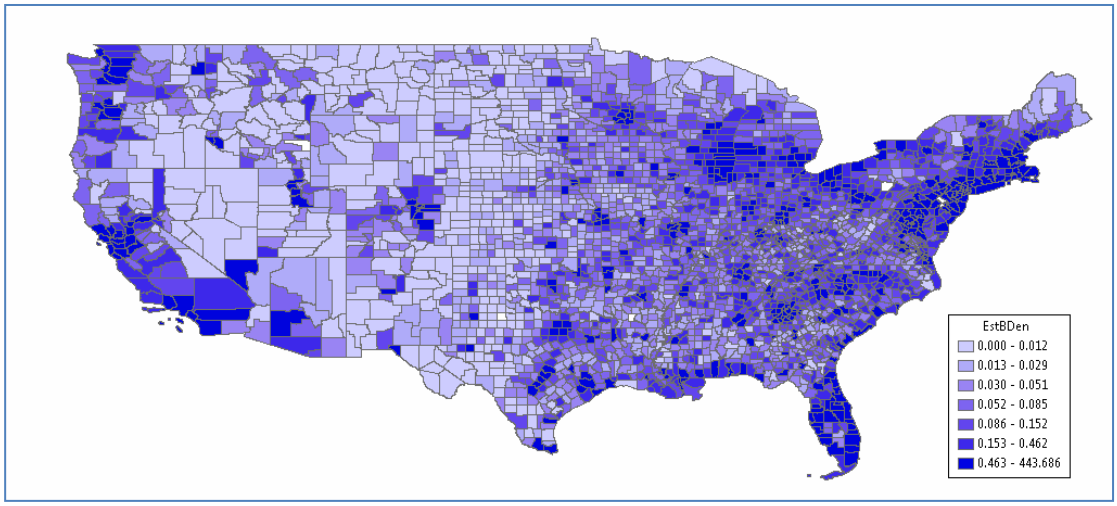


Figure 4.7 b): Density of New Establishments across U.S. Counties in the Lower 48 States during 2008-2009 (counts per sq. mi. of land)

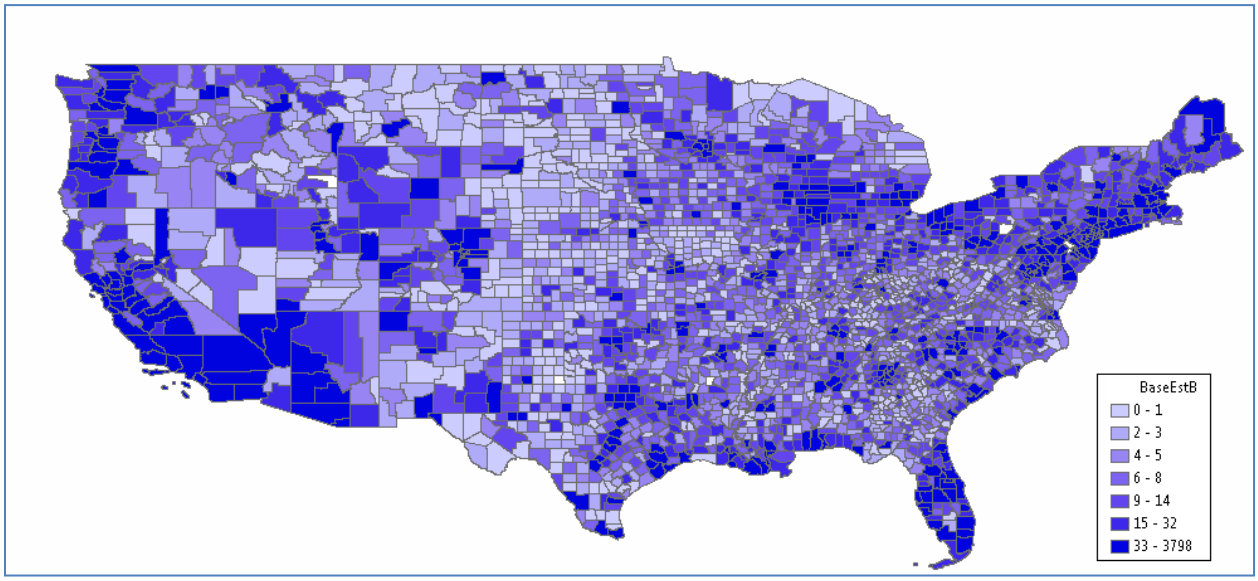


Figure 4.8: New Basic Establishments across U.S. Counties in the Lower 48 States during 2008-2009.

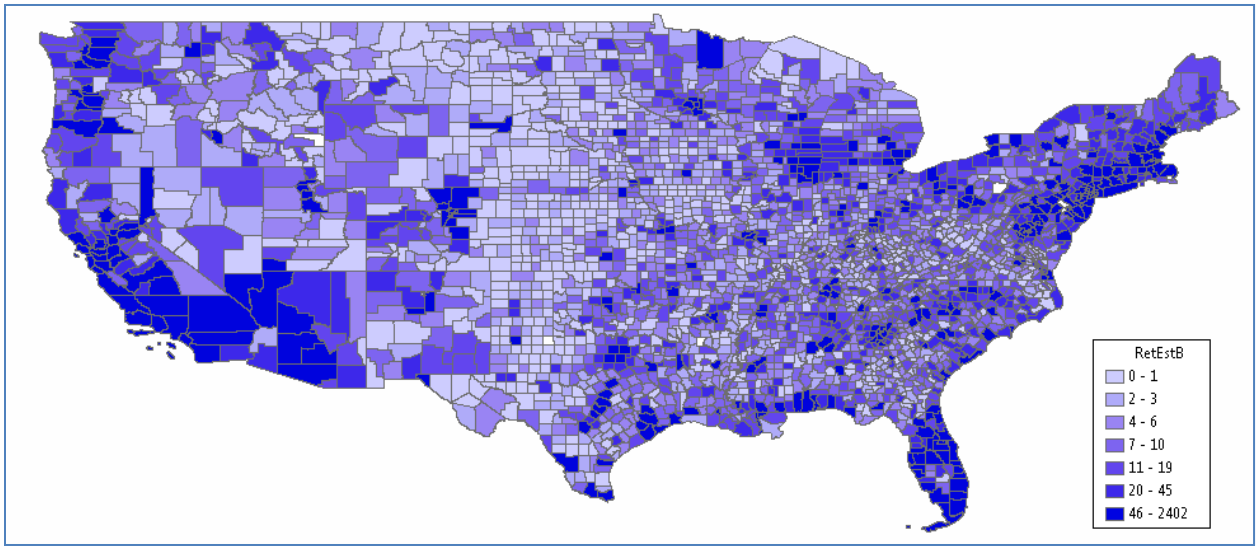


Figure 4.9: New Retail Establishments across U.S. Counties in the Lower 48 States during 2008-2009.

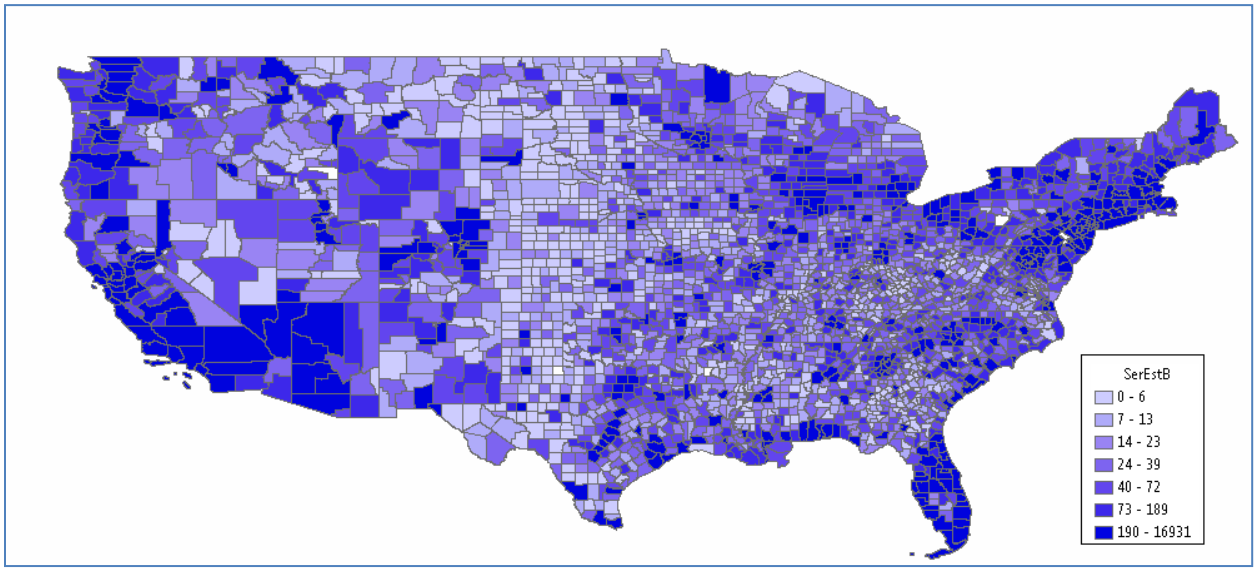


Figure 4.10: New Service Establishments across U.S. Counties in the Lower 48 States during 2008-2009.

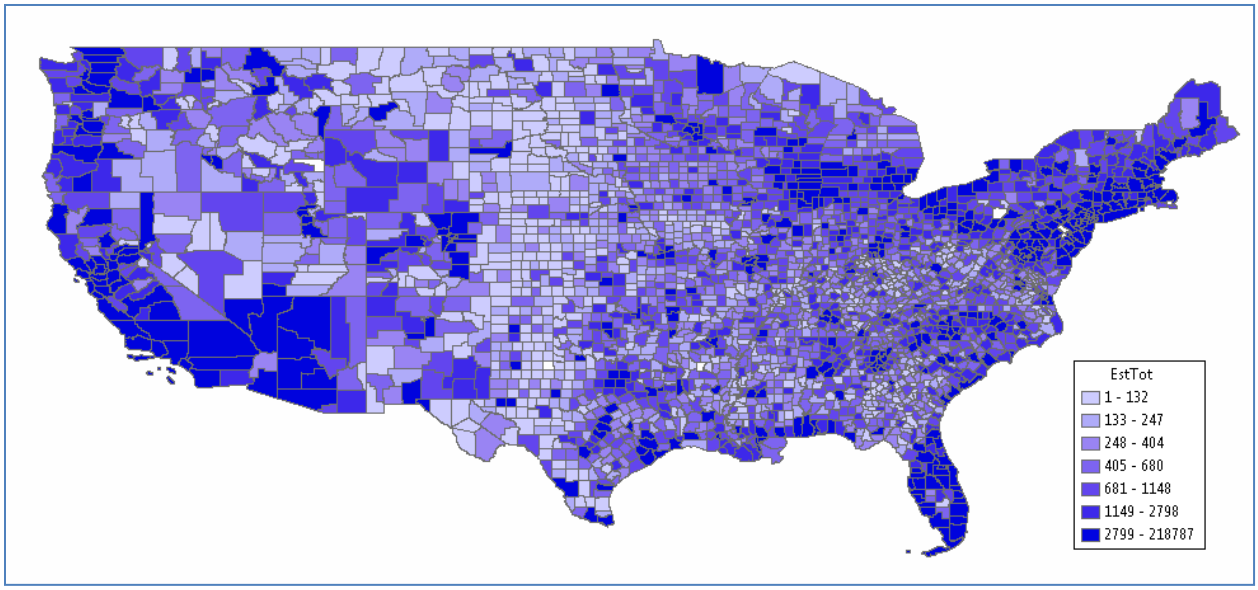


Figure 4.11: Total Number of Establishments across U.S. Counties in the Lower 48 States during 2008-2009.

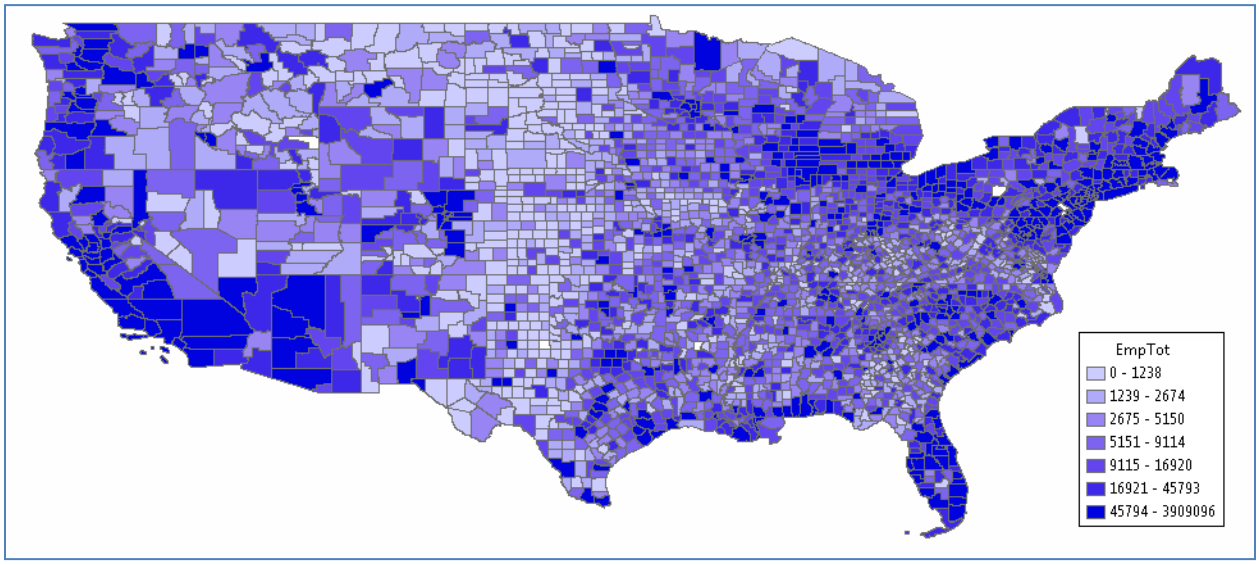


Figure 4.12: Total Number of Jobs in U.S. Counties across the Lower 48 States during 2008-2009.

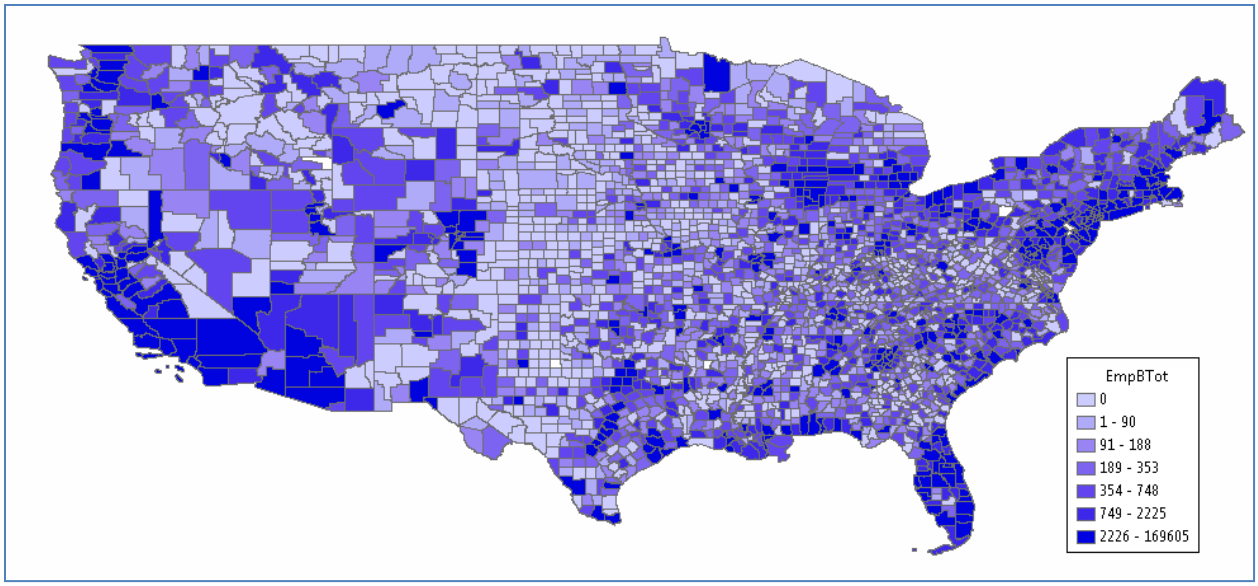


Figure 4.13: Total Number of New Jobs in U.S. Counties across the Lower 48 States during 2008-2009.

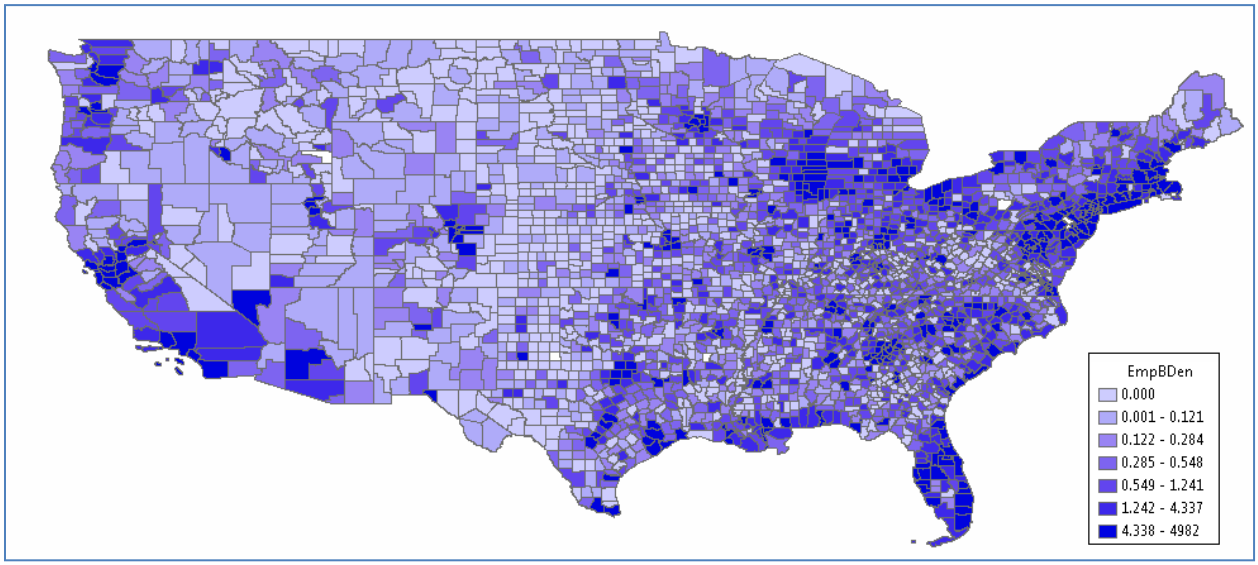


Figure 4.14: Density of New Jobs across U.S. Counties in the Lower 48 States during 2008-2009 (counts per sq. mi. of land).

4.2 4.3 Chapter Summary

This chapter discussed the details of developing polygon-level pedestrian crash and county-level firm birth data sets, including covariates used, to be fitted using the Poisson-based multivariate conditional auto-regressive (CAR) models in Chapter 5. The pedestrian crash data set covers

218 Travis County Thiessen polygons (constructed around the centroids of census tracts), with controls for transit stop density, land use characteristics, network features, school access, and sidewalk provision. The firm birth data set was assembled for a 1,316-county sample of the nation's lower 48 states (which hold a total of 3,109 counties), with controls for population and land use characteristics. Chapter 5 describes the model estimation and results analysis stemming from these two data sets.

CHAPTER 5: ANALYSIS AND RESULTS

This chapter discusses the results of the multivariate conditional autoregressive (MCAR) model developed in Chapter 3 of this dissertation, as applied first to bivariate and then trivariate simulated data sets (to ensure the code is converging on proper parameter values) and then to the data sets described in Chapter 4 (i.e., the Austin pedestrian crash counts and U.S. firm-birth data). In the context of spatial data analysis, for discrete response, the first of these two real data sets can be considered a typical or small sample (n = 218 polygon neighborhoods) and the latter a large sample (n = 1,316 counties). For this reason, these two sample sizes were also used in the initial, simulated data set tests, as described below.

5.1 Results of Simulated Data Test: Small-Sample Example with Two Response Levels

To test Chapter 3's estimation algorithm (coded in R and WinBUGS, as shown in Appendix A of this dissertation), a simulated data set with four covariates (including a constant term) was generated for Travis County's 218 Thiessen polygons. Recall that the first stage is expressed as a Poisson process: $y_{ik} \sim \text{Poisson}(\lambda_{ik})$, where y_{ik} is the observed counts by response types and k = 1, 2 for the i^{th} polygon of Travis County. The expected crash count of response level k, λ_{ik}, of the second-stage is formulated as shown:

$$\ln(\lambda_{ik}) = \ln({E_i}^{\alpha}) + \beta_0 + x_{i1} \cdot \beta_{1k} + x_{i2} \cdot \beta_{2k} + x_{i3} \cdot \beta_{3k} + \phi_{ik} + v_{ik}$$

For model testing purposes, the exposure measure, E_i, was generated from a uniform distribution, $\text{unif}(0,\ 50)$. The true $\boldsymbol{\beta}$ values for the two types of response were set to $\boldsymbol{\beta}_{\cdot 1} = (\beta_{01}, \beta_{11}, \beta_{21}, \beta_{31})' = (0.5, 1, -1.2, 1.5)'$ and $\boldsymbol{\beta}_{\cdot 2} = (\beta_{02}, \beta_{12}, \beta_{22}, \beta_{32})' = (1, 1.5, -1, 2)'$. The covariates ($x_i$'s) were random draws from the standard normal distribution.

The spatial error term, ϕ_{ik}, was simulated in two steps: first, a vector of zero-centered spatial error terms, $\boldsymbol{\phi}_{\cdot 2}$, was generated from a multivariate normal distribution, $N(\mathbf{0},\ [(\mathbf{D} - \rho_2\mathbf{W})\tau_2]^{-1})$, where the square matrix **W** is the unnormalized weight matrix (defined via first-order contiguity) and **D** is a diagonal matrix containing the *n* row-sums of **W**. Both **W** and **D** are known and derived using the 218 Thiessen polygons of Travis County. The values of parameters α_2 and τ_2 were set to 0.6 and 2, respectively.

The second step drew $\boldsymbol{\phi}_{\cdot 1}$ from a multivariate normal distribution conditional on $\boldsymbol{\phi}_{\cdot 2}$, such that $\boldsymbol{\phi}_{\cdot 1}|\boldsymbol{\phi}_{\cdot 2} \sim N(\mathbf{A}\boldsymbol{\phi}_{\cdot 2},\ [(\mathbf{D}-\rho_1\mathbf{W})\tau_1]^{-1})$, where $\mathbf{A} = (\eta_0\mathbf{I} + \eta_1\mathbf{W})$ with $\eta_0 = 0.8$ and $\eta_1 = 0.5$. The true values for parameters ρ_1 and τ_1 were set at 0.75 and 1.5, respectively. The other random error term, v_{ik}, describes zone-specific latent heterogeneity and is shared across different types of responses. It was assumed to follow a log-normal prior: $v_{ik} \sim N\left(0, \frac{1}{\tau_{vk}}\right)$ with τ_{vk} assigned a gamma prior, leading to a lognormal MCAR model. Heteroskedasticity is allowed here, thanks to the response-specific variance term, $\frac{1}{\tau_{vk}}$. Alternatively, the exponential of v_{ik} could be assumed to follow a gamma prior: $\exp(v_{ik}) \sim \text{Gamma}(\theta_k, \theta_k)$, with θ_k assigned a gamma prior, leading to a negative-binomial-type model.

Tables 5.1 and 5.2 summarize the true parameter values and estimation results, respectively. Figure 5.1 shows the trace plots of parameter draws from two chains generated using two sets of different starting values, and Figure 5.2 illustrates the corresponding density plots, with black dots on the horizontal axis denoting each sampled point.

Table 5.1: True Parameter Values of the Two-Response Example.

Parameter	True Value	Parameter	True Value
β_{01}	0.5	β_{02}	1
β_{11}	1	β_{12}	1.5
β_{21}	-1.2	β_{22}	-1
β_{31}	1.5	β_{32}	2
ρ_1	0.75	ρ_2	0.6
τ_1	1.5	τ_2	2
η_0	0.8	η_1	0.5
τ_{v1}	0.5	τ_{v2}	0.2

Table 5.1: Estimation Results of the Small Sample Example.

Parameter	Estimated Mean	STD	2.50%	Median	97.50%	MC error	MC Error/STD
ρ_1	0.765	0.103	0.536	0.778	0.929	1.57E-03	1.52%
ρ_2	0.640	0.103	0.429	0.645	0.827	2.35E-03	2.28%
β_{01}	0.102	0.153	-0.210	0.105	0.400	6.31E-03	4.12%
β_{11}	1.127	0.092	0.950	1.125	1.313	4.23E-03	4.58%
β_{21}	-1.217	0.088	-1.394	-1.214	-1.047	3.95E-03	4.49%
β_{31}	1.776	0.098	1.585	1.775	1.972	4.34E-03	4.44%
β_{02}	0.948	0.055	0.839	0.948	1.056	1.76E-03	3.19%
β_{12}	1.371	0.016	1.339	1.371	1.403	3.31E-04	2.03%
β_{22}	-0.826	0.022	-0.869	-0.826	-0.784	4.71E-04	2.18%
β_{32}	2.032	0.033	1.968	2.031	2.098	9.98E-04	3.01%
η_0	0.814	0.082	0.625	0.823	0.944	1.39E-03	1.69%
η_1	0.519	0.125	0.258	0.522	0.761	6.04E-03	4.83%
τ_{v1}	0.464	0.067	0.345	0.459	0.611	1.70E-03	2.53%
τ_{v2}	0.217	0.036	0.160	0.212	0.303	1.70E-03	4.73%
τ_1	1.625	0.467	0.872	1.579	2.674	1.66E-02	3.56%
τ_2	1.963	0.298	1.256	1.972	2.518	1.35E-02	4.54%
n.chain=2; n.iter=8,500; n.burn-in=2,000; sample=16,000							

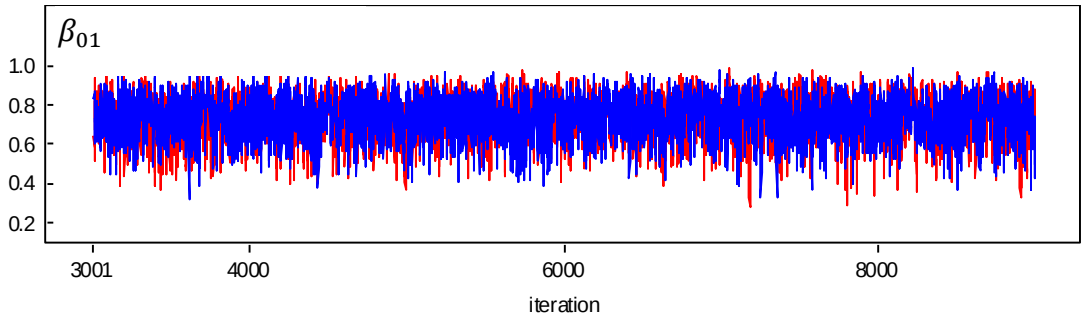
Notes: n.chain = number of chains; n.iter = number of iterations; n.burn-in = number of burn-in; sample = total draws (excluding burn-in period) generated = (n.iter−n.burn-in)×n.chain. STD and MC stand for standard deviation and Monte Carlo, respectively.

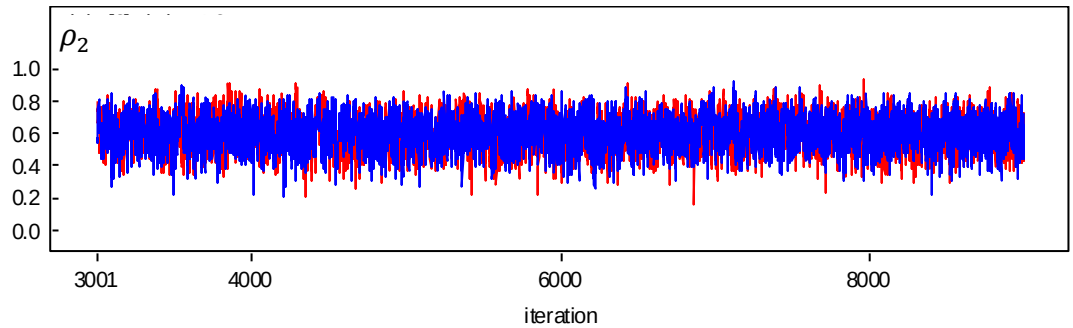
Monte Carlo errors were also reported. These reflect the performance of the posterior estimates and can be used to determine the number of iterations needed to achieve reasonable estimates. Generally, Monte Carlo errors should be no greater than 5% of the standard deviation of the sample after burn-in. The small-sample simulated-data results suggest that the proposed MCAR model and estimation algorithm can recover the true parameter values and (appear to) achieve convergence based on Geweke's (1992) diagnostic test (as described below) using the first 650 draws as compared to the last 3,250 draws (after eliminating the first 2,000 draws). Table 5.4 displays these results.

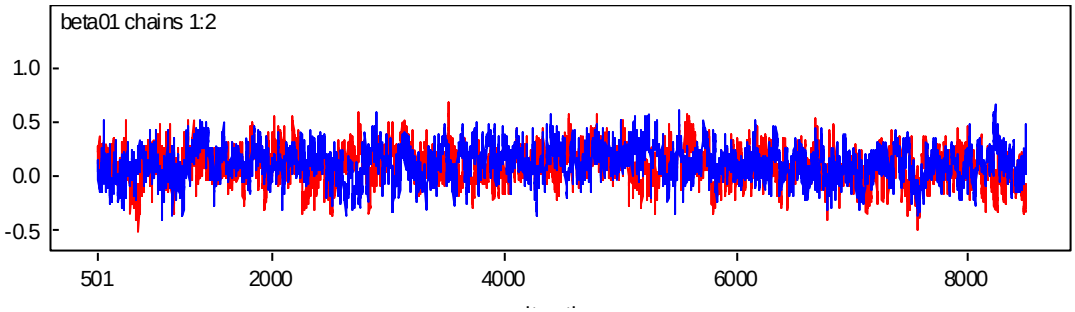
Table 5.4 Results of Geweke's Diagnostic Test for Small-Sample Example.

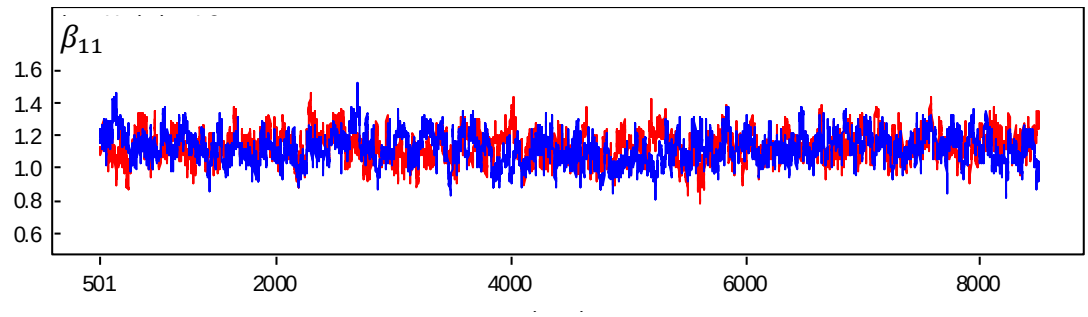
Parameter	Geweke's Test Statistic	Parameter	Geweke's Test Statistic
β_{01}	-0.212	β_{02}	-0.498
β_{11}	-1.315	β_{12}	-0.762
β_{21}	-0.663	β_{22}	-0.122
β_{31}	0.905	β_{32}	0.492
ρ_1	0.661	ρ_2	1.002
τ_1	0.102	τ_2	-1.290
η_0	0.832	η_1	-2.021
τ_{v1}	1.629	τ_{v2}	1.591

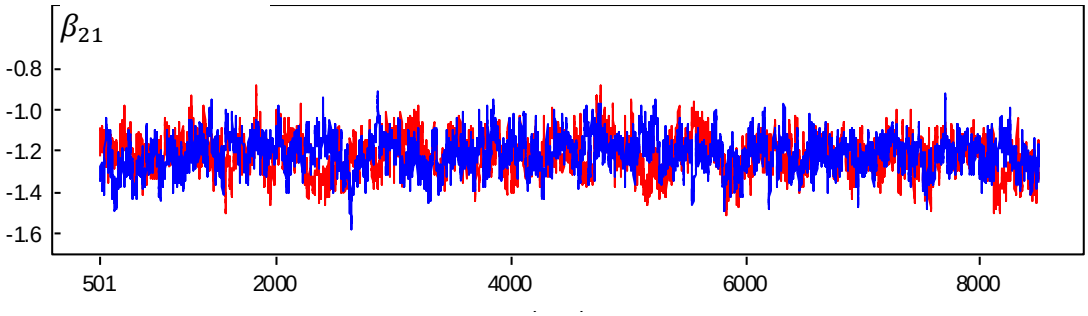
Geweke's (1992) test relies on comparing the two sample means of a sample drawn at the beginning and the end of MCMC outputs. If the two sample means show no difference, then convergence is thought to be achieved, using a Z statistic: $Z = \frac{\bar{\theta}^A - \bar{\theta}^B}{\sqrt{S_\theta^A/T_A + S_\theta^B/T_B}}$, where $\bar{\theta}^A$ denotes the sample mean of the subsample drawn at the beginning of the MCMC output, $\bar{\theta}^B$ denotes the sample mean of the subsample drawn at the end of the MCMC output, S_θ^A and S_θ^B indicate the sample variance corresponding to the two samples (with sample size T_A and T_B, respectively). This test has been coded in R's CODA package, where samples A and B are set to be the first 10 percent and last 50 percent of the MCMC outputs, respectively. Best et al. (1996) showed how this Z statistic asymptotically follows a standard-normal distribution, such that |Z| values greater than 2.0 hint at non-convergence, due to the noticeable discrepancy between the two samples.

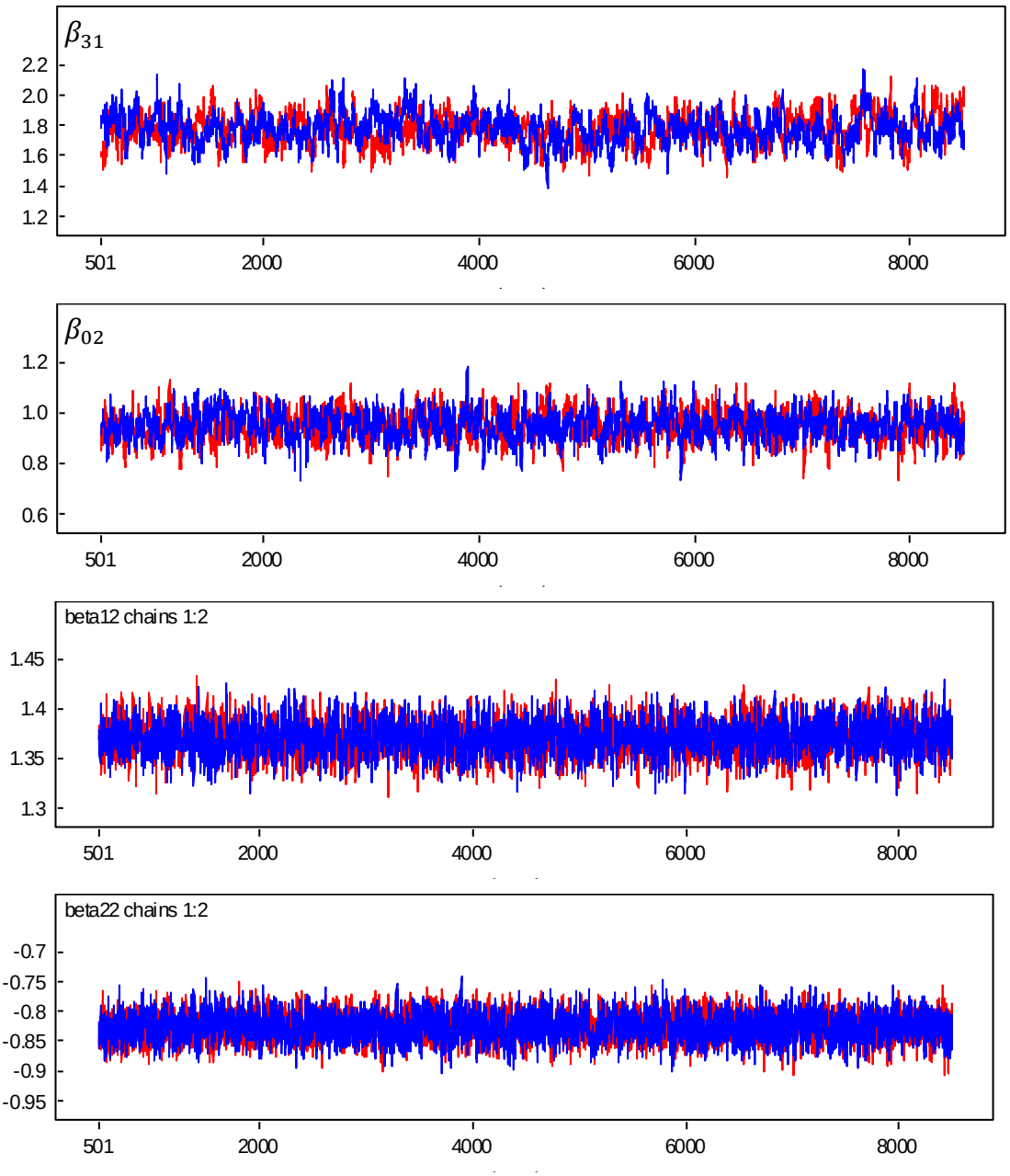

β_{01}
1.0
0.8
0.6
0.4
0.2
3001
4000
6000
8000
iteration

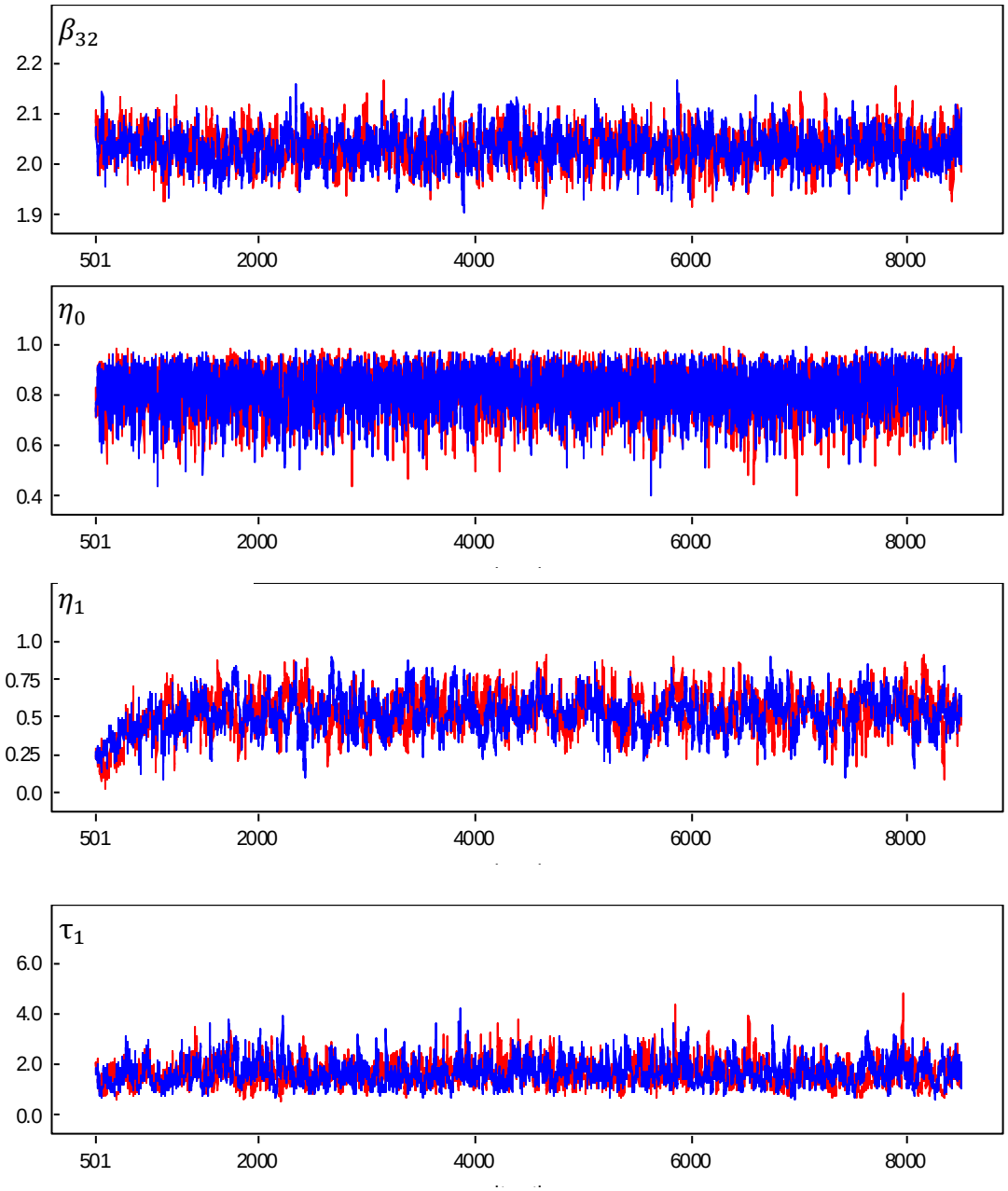

ρ_2
1.0
0.8
0.6
0.4
0.2
0.0
3001
4000
6000
8000
iteration


beta01 chains 1:2
1.0
0.5
0.0
-0.5
501
2000
4000
6000
8000


β_{11}
1.6
1.4
1.2
1.0
0.8
0.6
501
2000
4000
6000
8000


β_{21}
-0.8
-1.0
-1.2
-1.4
-1.6
501
2000
4000
6000
8000


β_{31}
2.2
2.0
1.8
1.6
1.4
1.2
501
2000
4000
6000
8000
β_{02}
1.2
1.0
0.8
0.6
501
2000
4000
6000
8000
beta12 chains 1:2
1.45
1.4
1.35
1.3
501
2000
4000
6000
8000
beta22 chains 1:2
-0.7
-0.75
-0.8
-0.85
-0.9
-0.95
501
2000
4000
6000
8000


β_{32}
2.2
2.1
2.0
1.9
501
2000
4000
6000
8000
η_0
1.0
0.8
0.6
0.4
501
2000
4000
6000
8000
η_1
1.0
0.75
0.5
0.25
0.0
501
2000
4000
6000
8000
τ_1
6.0
4.0
2.0
0.0
501
2000
4000
6000
8000

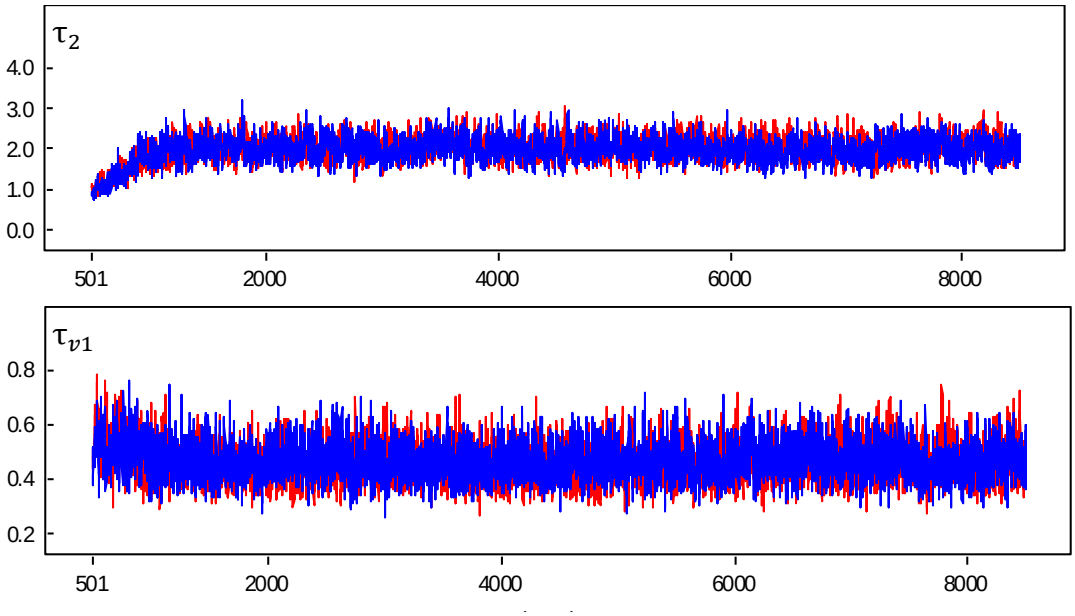
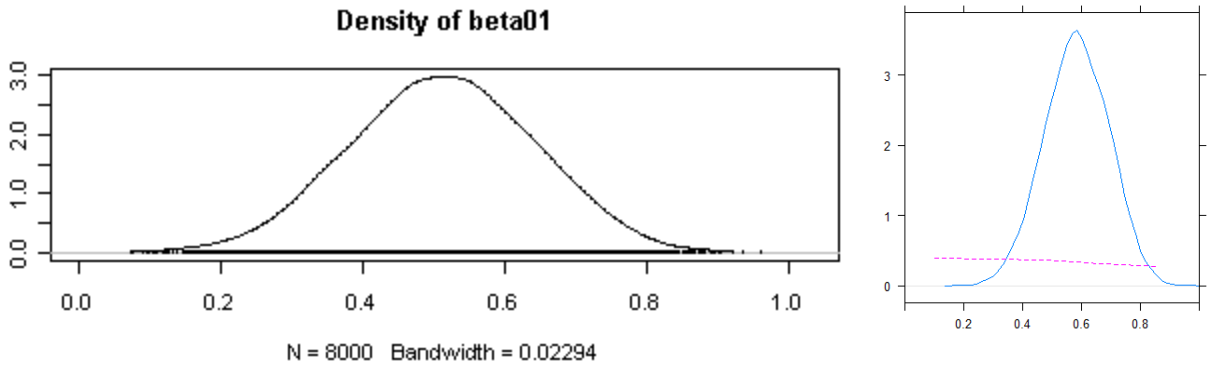
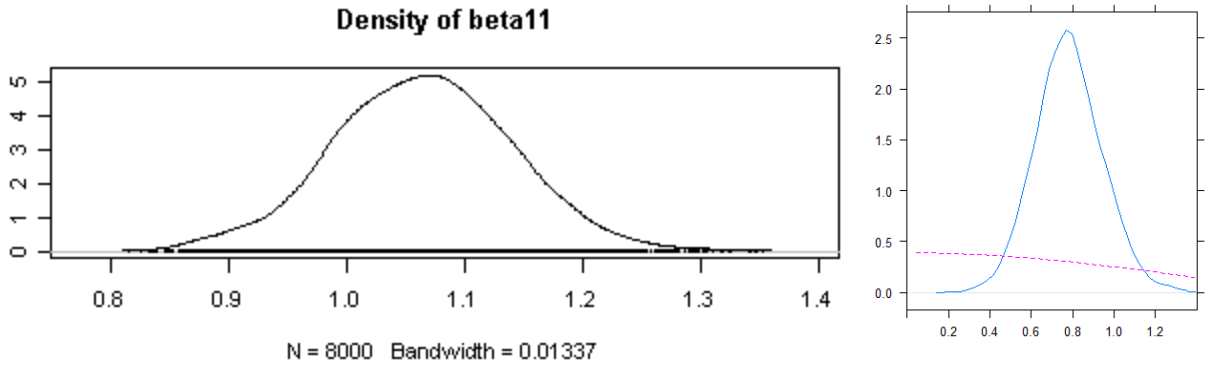
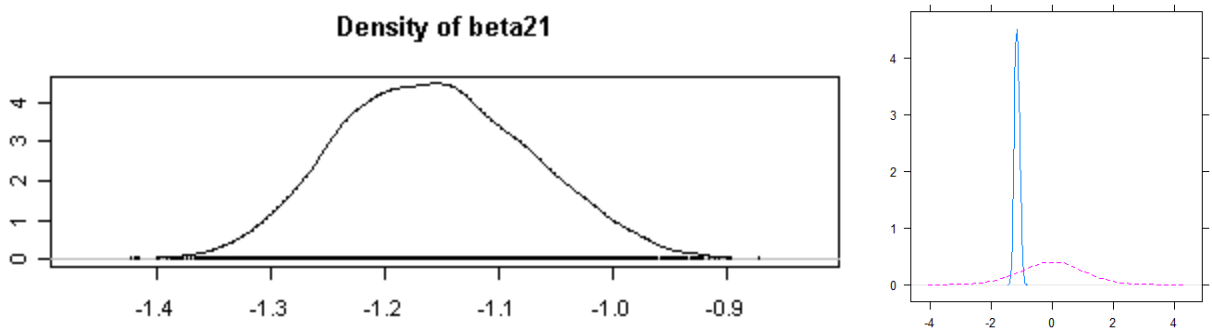
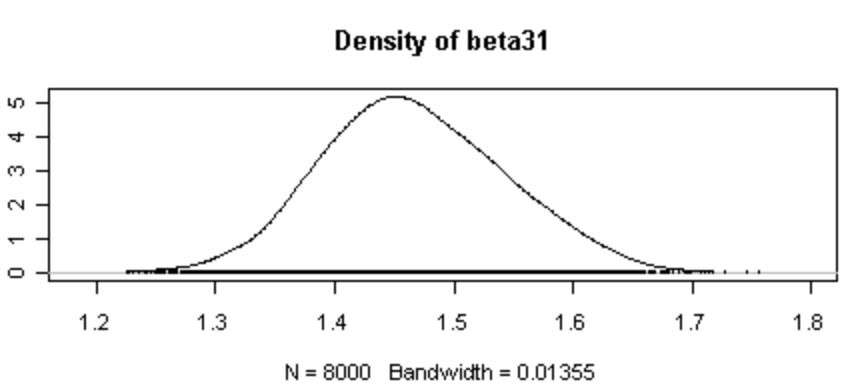


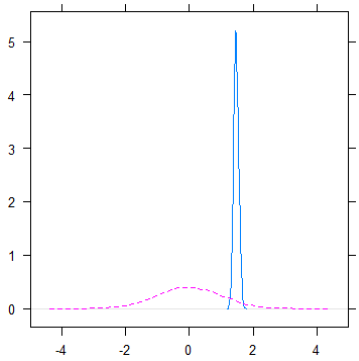
Figure 5.1: Trace Plots for Parameter Estimates of the Small Sample Example (after burn-in sample).

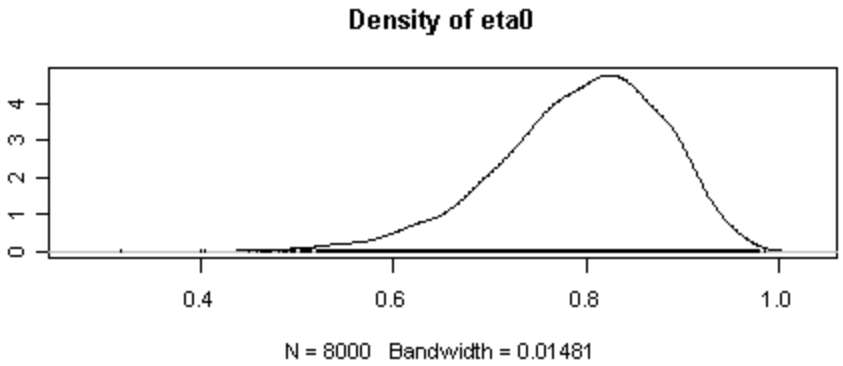

Density of beta01
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
1.0
N = 8000 Bandwidth = 0.02294
0
1
2
3
0.2
0.4
0.6
0.8

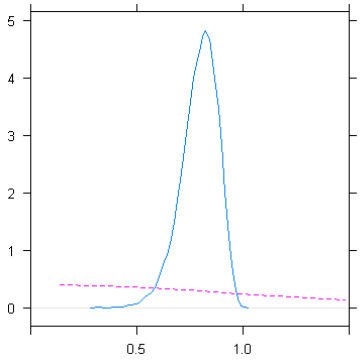

Density of beta11
0
1
2
3
4
5
0.8
0.9
1.0
1.1
1.2
1.3
1.4
N = 8000 Bandwidth = 0.01337
0.0
0.5
1.0
1.5
2.0
2.5
0.2
0.4
0.6
0.8
1.0
1.2

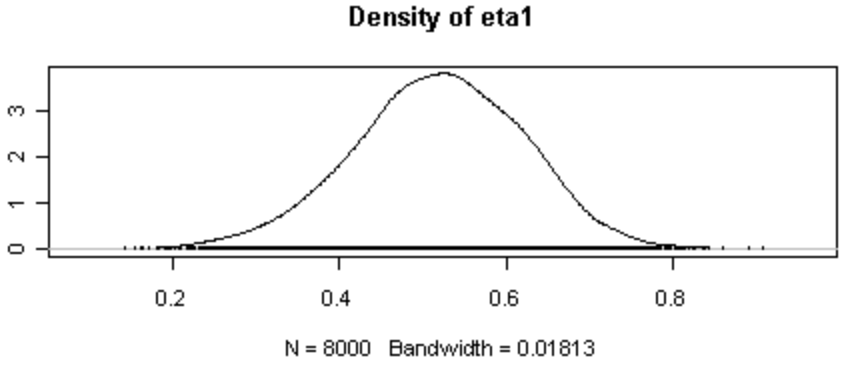

Density of beta21
0
1
2
3
4
-1.4
-1.3
-1.2
-1.1
-1.0
-0.9
0
1
2
3
4
-4
-2
0
2
4

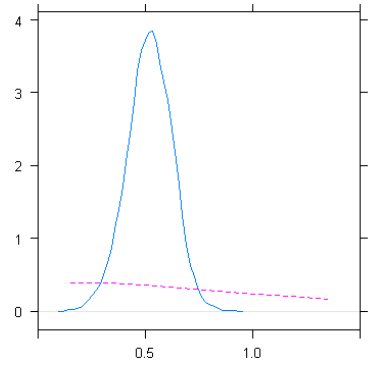

Density of beta31
0 1 2 3 4 5
1.2 1.3 1.4 1.5 1.6 1.7 1.8
N = 8000 Bandwidth = 0.01355

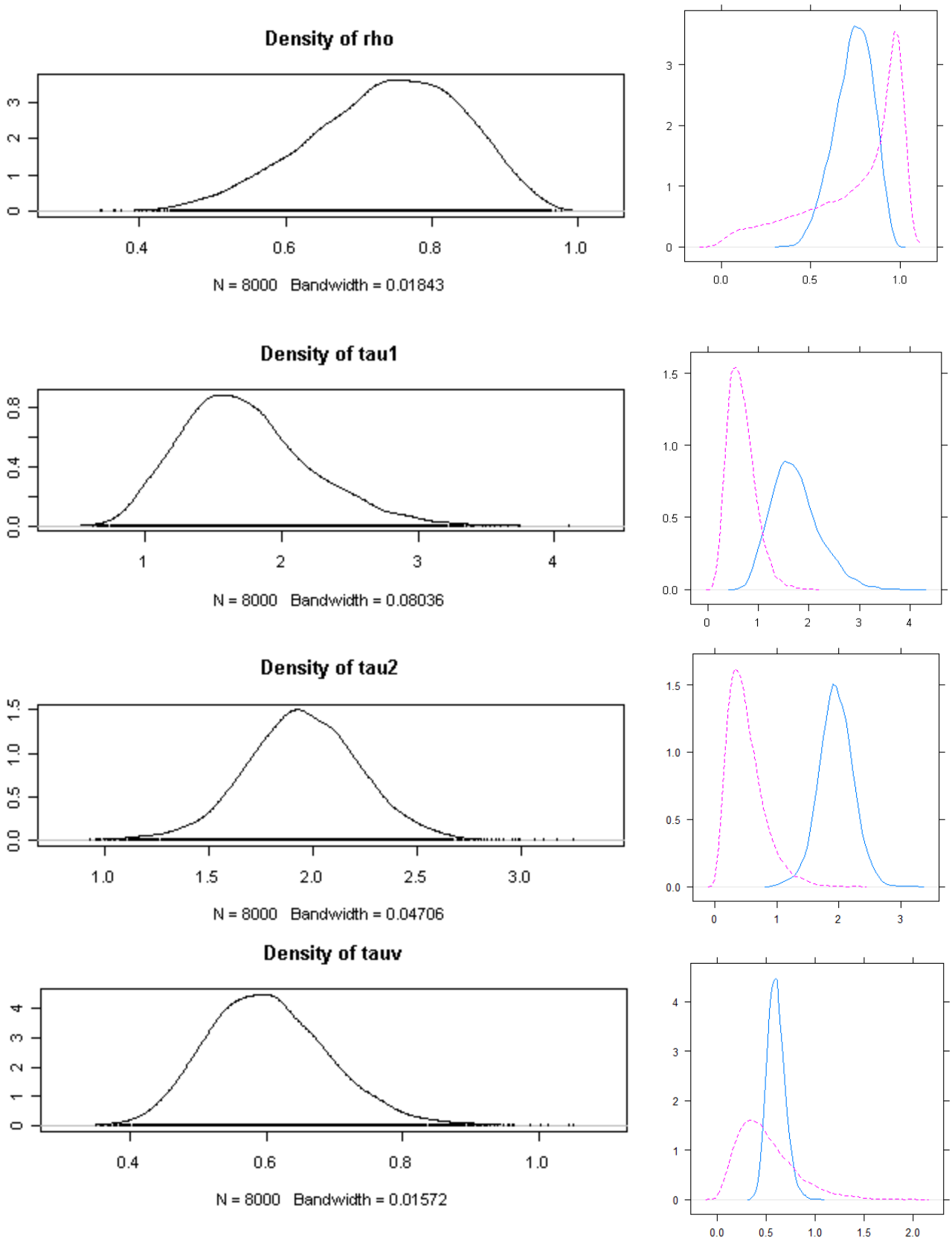

5
4
3
2
1
0
-4 -2 0 2 4

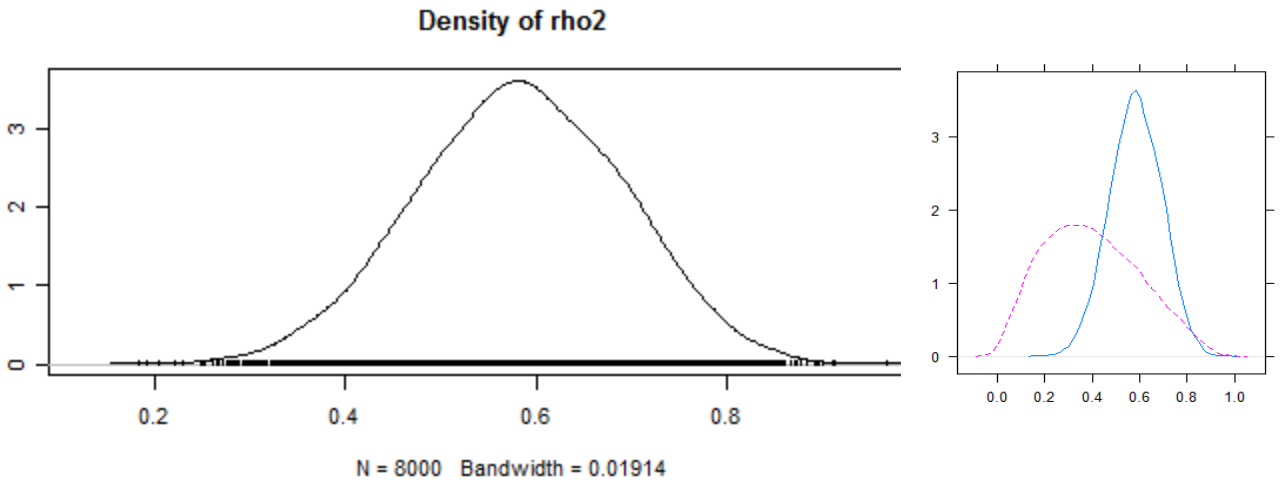

Density of eta0
0 1 2 3 4
0.4 0.6 0.8 1.0
N = 8000 Bandwidth = 0.01481

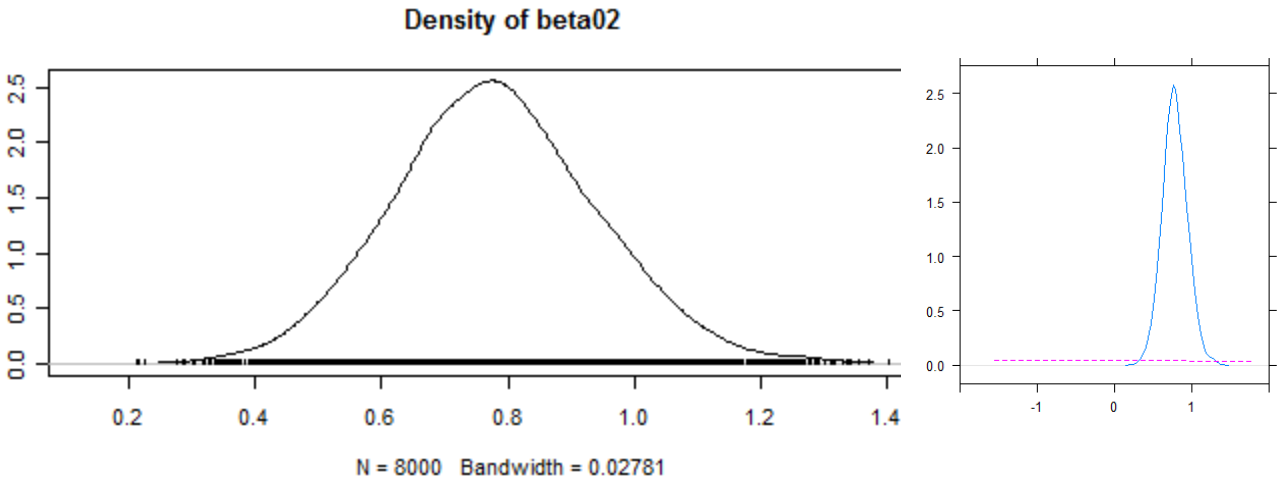

5
4
3
2
1
0
0.5 1.0

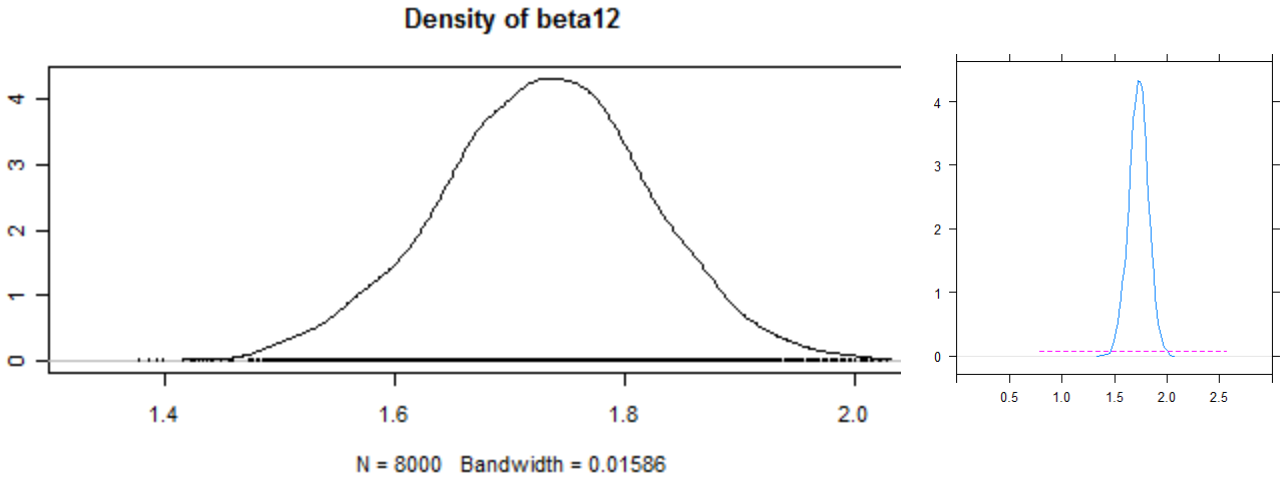

Density of eta1
0 1 2 3
0.2 0.4 0.6 0.8
N = 8000 Bandwidth = 0.01813


4
3
2
1
0
0.5 1.0


Density of rho
N = 8000 Bandwidth = 0.01843
Density of tau1
N = 8000 Bandwidth = 0.08036
Density of tau2
N = 8000 Bandwidth = 0.04706
Density of tauv
N = 8000 Bandwidth = 0.01572


Density of rho2
N = 8000 Bandwidth = 0.01914


Density of beta02
N = 8000 Bandwidth = 0.02781


Density of beta12
N = 8000 Bandwidth = 0.01586

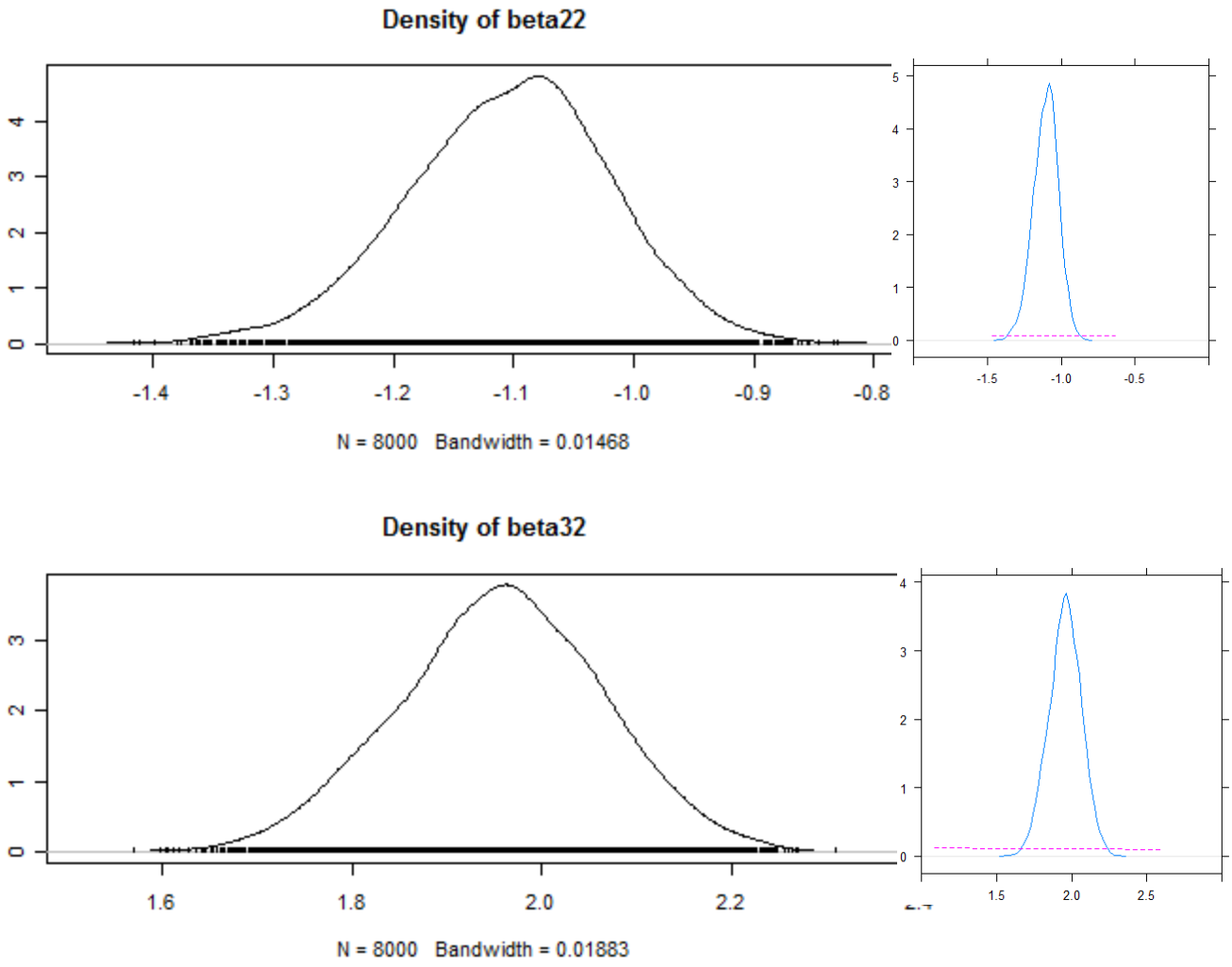


Figure 5.2: Density Plots for MCMC Draws of the Small Sample Example. (Note: Dotted pink lines show prior densities, while solid blue lines indicate posterior densities.)

5.2 Results of Simulated Data Test: Large-Sample Example with Three Response Levels

This section describes the simulation study for a trivariate MCAR model, using a data-generating process similar to that expected for the sample of 1,316 U.S. counties (as described in Chapter 3).

The true values of all slope parameters ($\boldsymbol{\beta}$) for the three response types were set to $\boldsymbol{\beta}_{\cdot 1} = (\beta_{01}, \beta_{11}, \beta_{21}, \beta_{31})' = (0.5, 1, -1.2, 1.5)'$, $\boldsymbol{\beta}_{\cdot 2} = (\beta_{02}, \beta_{12}, \beta_{22}, \beta_{32})' = (1, 1.5, -1, 2)'$, and $\boldsymbol{\beta}_{\cdot 3} = (\beta_{03}, \beta_{13}, \beta_{23}, \beta_{33})' = (2, 2.5, -1.3, 0.2)'$. The spatial autocorrelation coefficients ρ's for the three response types were set to 0.75, 0.6, and 0.3. Parameters η_{012}, η_{013}, and η_{023} capture

the cross-correlations among the three responses, while η_{112}, η_{113}, and η_{123} reveal the spatially lagged cross-correlations. Note that as the number of response levels increases, the dimension of these cross-correlation parameters also rises (so a situation with *k*=4 response types involves C_4^2 =6 cross-correlation parameters and 6 spatially lagged cross-correlation parameters). The variances of the covariance matrix for the random spatial terms of the three response types are described by parameters τ_1, τ_2, and τ_3, respectively, while τ_v represents the inverse of the variance for the heterogeneity term v_i shared by the three response levels. The true parameter values are summarized in Table 5.3.

Table 5.3: True Parameter Values of the Three-Level Response Example.

Parameter	True Value	Parameter	True Value	Parameter	True Value
β_{01}	0.5	β_{02}	1	β_{03}	2
β_{11}	1	β_{12}	1.5	β_{13}	2.5
β_{21}	-1.2	β_{22}	-1	β_{23}	-1.3
β_{31}	1.5	β_{32}	2	β_{33}	0.2
ρ_1	0.75	ρ_2	0.6	ρ_3	0.3
τ_1	1.5	τ_2	2	τ_3	0.8
η_{012}	0.8	η_{112}	0.5	η_{013}	0.7
η_{113}	0.4	η_{023}	0.1	η_{123}	0.2
τ_{v1}	0.5	τ_{v2}	1.0	τ_{v3}	2.5

The first stage includes three simultaneous Poisson processes: $y_{ik}\sim$ Poisson(λ_{ik}), where y_{ik} are the observed counts by response levels; *k*=3 denotes the number of response levels/outcomes; and *i* indicates the i^{th} geographic unit. The mean crash rates, λ_{ik}, of the second stage represent the expected counts, where $\ln(\lambda_{ik}) = \ln({E_i}^{\alpha}) + \beta_0 + x_{i1}\cdot\beta_{1k} + x_{i2}\cdot\beta_{2k} + x_{i3}\cdot\beta_{3k} + \phi_{ik} + v_i$, with parameters defined as in the previous section. The exposure measure, E_i, is generated from a uniform distribution, unif$(0,\ 50)$, and the covariates x_i are random draws from the standard normal distribution. The spatial error term, ϕ_{ik}, was simulated in a way such that $\boldsymbol{\phi}_{\cdot 3}\sim N(\mathbf{0},\ [(\mathbf{D}-\rho_3\mathbf{W})\tau_3]^{-1})$, $\boldsymbol{\phi}_{\cdot 2}|\boldsymbol{\phi}_{\cdot 3}\sim N(\mathbf{A}_{23}\boldsymbol{\phi}_{\cdot 3},\ [(\mathbf{D}-\rho_2\mathbf{W})\tau_2]^{-1})$, and $\boldsymbol{\phi}_{\cdot 1}|\boldsymbol{\phi}_{\cdot 2},\boldsymbol{\phi}_{\cdot 3}\sim N(\mathbf{A}_{12}\boldsymbol{\phi}_{\cdot 2}+\mathbf{A}_{13}\boldsymbol{\phi}_{\cdot 3},\ [(\mathbf{D}-\rho_1\mathbf{W})\tau_1]^{-1})$, where $\mathbf{W}$ is an unnormalized square

weight matrix defined via first-order contiguity or interzonal distances, and **D** is a diagonal matrix containing the *n* row-sums of **W**, as in the bivariate example. Square matrices $\mathbf{A}_{23}$, $\mathbf{A}_{12}$, and $\mathbf{A}_{13}$ capture the aspatial and spatially-lagged cross-correlations and are parameterized as $\mathbf{A}_{23} = \eta_{023}\mathbf{I} + \eta_{123}\mathbf{W}$, $\mathbf{A}_{12} = \eta_{012}\mathbf{I} + \eta_{112}\mathbf{W}$, and $\mathbf{A}_{13} = \eta_{013}\mathbf{I} + \eta_{113}\mathbf{W}$. The heterogeneity error term, v_i, is shared across the three outcomes, and is assumed to follow a log-normal prior: $v_{ik} \sim N\left(0, \frac{1}{\tau_{vk}}\right)$ with τ_{vk} assigned a gamma prior, leading to a lognormal MCAR model. Figure 5.3 illustrates the trace plots of two MCMC chains generated using two sets of different starting values and Table 5.3 summarizes parameter estimates. Table 5.3: Estimation Results of the Three-Response Example.

Parameter	True Values	Estimated Mean	STD	2.50%	Median	97.50%	MC error	MC Error/STD
ρ_1	0.75	0.773	0.10	0.546	0.785	0.933	0.002	1.59%
ρ_2	0.6	0.616	0.10	0.415	0.619	0.791	0.001	1.48%
ρ_3	0.3	0.359	0.168	0.11	0.356	0.573	0.002	0.95%
β_{01}	0.5	0.884	0.18	0.521	0.888	1.234	0.007	3.82%
β_{02}	1	0.933	0.10	0.729	0.937	1.125	0.007	6.84%
β_{03}	2	2.126	0.09	1.951	2.125	2.304	0.006	6.74%
β_{11}	1	1.118	0.08	0.969	1.115	1.276	0.003	4.11%
β_{12}	1.5	1.584	0.07	1.458	1.583	1.717	0.004	6.22%
β_{13}	2.5	2.428	0.07	2.298	2.428	2.566	0.005	6.47%
β_{21}	-1.2	-1.097	0.09	-1.277	-1.097	-0.918	0.004	4.28%
β_{22}	-1	-0.957	0.07	-1.086	-0.957	-0.831	0.004	5.86%
β_{23}	-1.3	-1.241	0.05	-1.350	-1.238	-1.149	0.003	6.03%
β_{31}	1.5	1.548	0.07	1.416	1.548	1.684	0.003	3.86%
β_{32}	2	2.047	0.07	1.915	2.048	2.181	0.004	6.07%
β_{33}	0.2	0.069	0.07	-0.061	0.066	0.211	0.005	6.61%
η_{012}	0.8	0.834	0.07	0.669	0.844	0.950	0.001	1.57%
η_{013}	0.7	0.721	0.09	0.534	0.727	0.876	0.003	2.92%
η_{023}	0.1	0.108	0.06	0.030	0.096	0.242	0.004	6.62%
η_{112}	0.5	0.502	0.10	0.305	0.504	0.702	0.003	2.99%

η_{113}	0.4	0.407	0.10	0.225	0.405	0.597	0.003	2.84%
η_{123}	0.2	0.180	0.06	0.056	0.181	0.302	0.004	6.75%
τ_{v1}	0.5	0.684	0.090	0.520	0.677	0.890	0.003	3.33%
τ_{v2}	1	0.981	0.185	0.653	0.971	1.299	0.004	2.17%
τ_{v3}	2.5	2.610	0.500	1.925	2.580	3.265	0.008	1.60%
τ_1	1.5	1.779	0.48	1.004	1.729	2.840	0.015	3.09%
τ_2	2	2.017	0.25	1.568	2.004	2.540	0.007	2.87%
τ_3	0.8	0.852	0.30	0.370	0.814	1.539	0.020	6.56%
n.chain = 2; n.iter = 10,000; n.burn-in = 2000; sample = 16,000								
DIC = 2,498.06, Run times = 7,800 seconds								

Notes: n.chain = number of chains; n.iter = number of iterations; n.burn-in = number of burn-in; sample = total draws (excluding burn-in period) generated = (n.iter−n.burn-in)×n.chain. STD and MC stand for standard deviation and Monte Carlo, respectively.

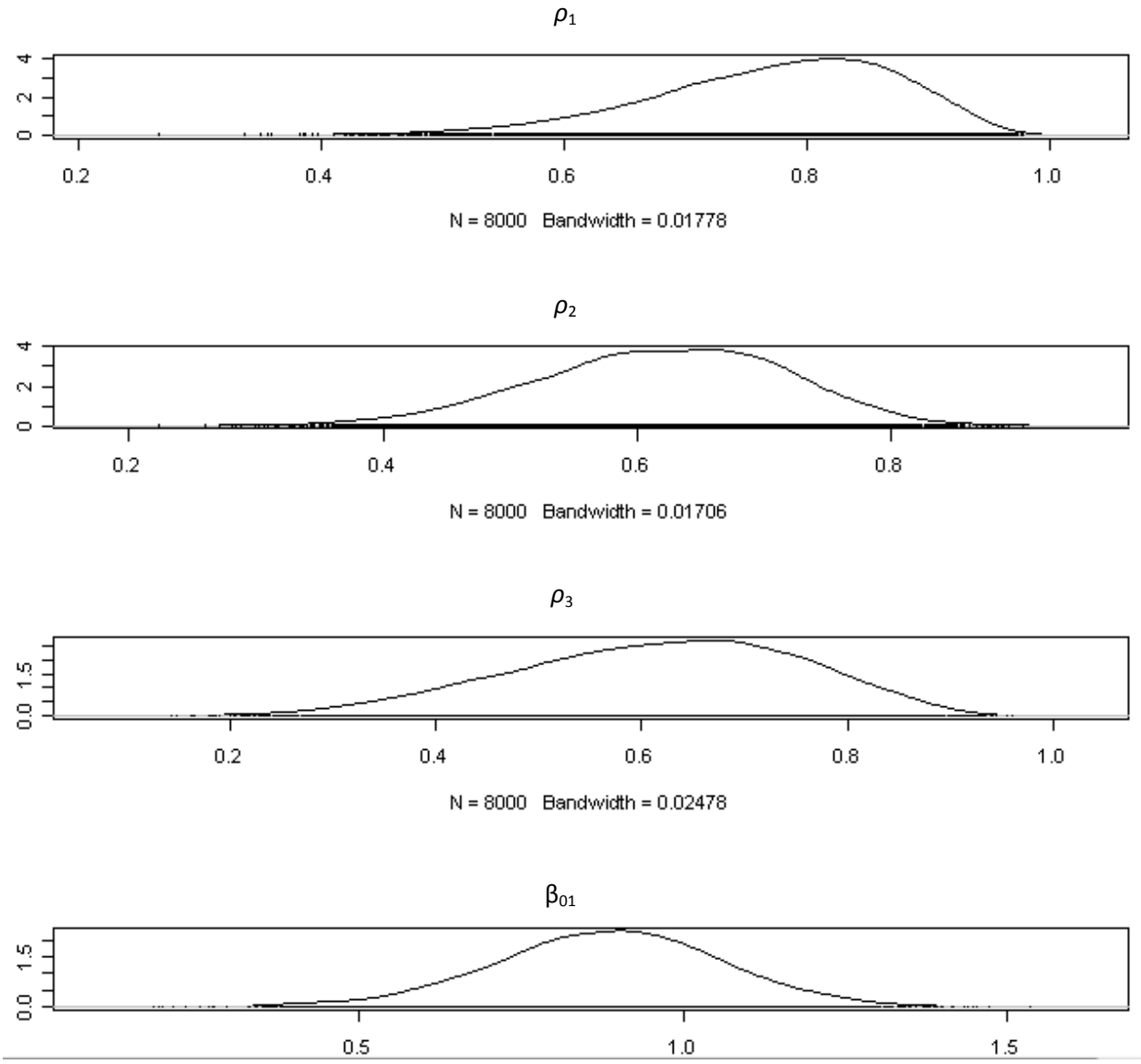


Figure 5.3 a) Density Plots for MCMC Draws of the Large Sample Example.

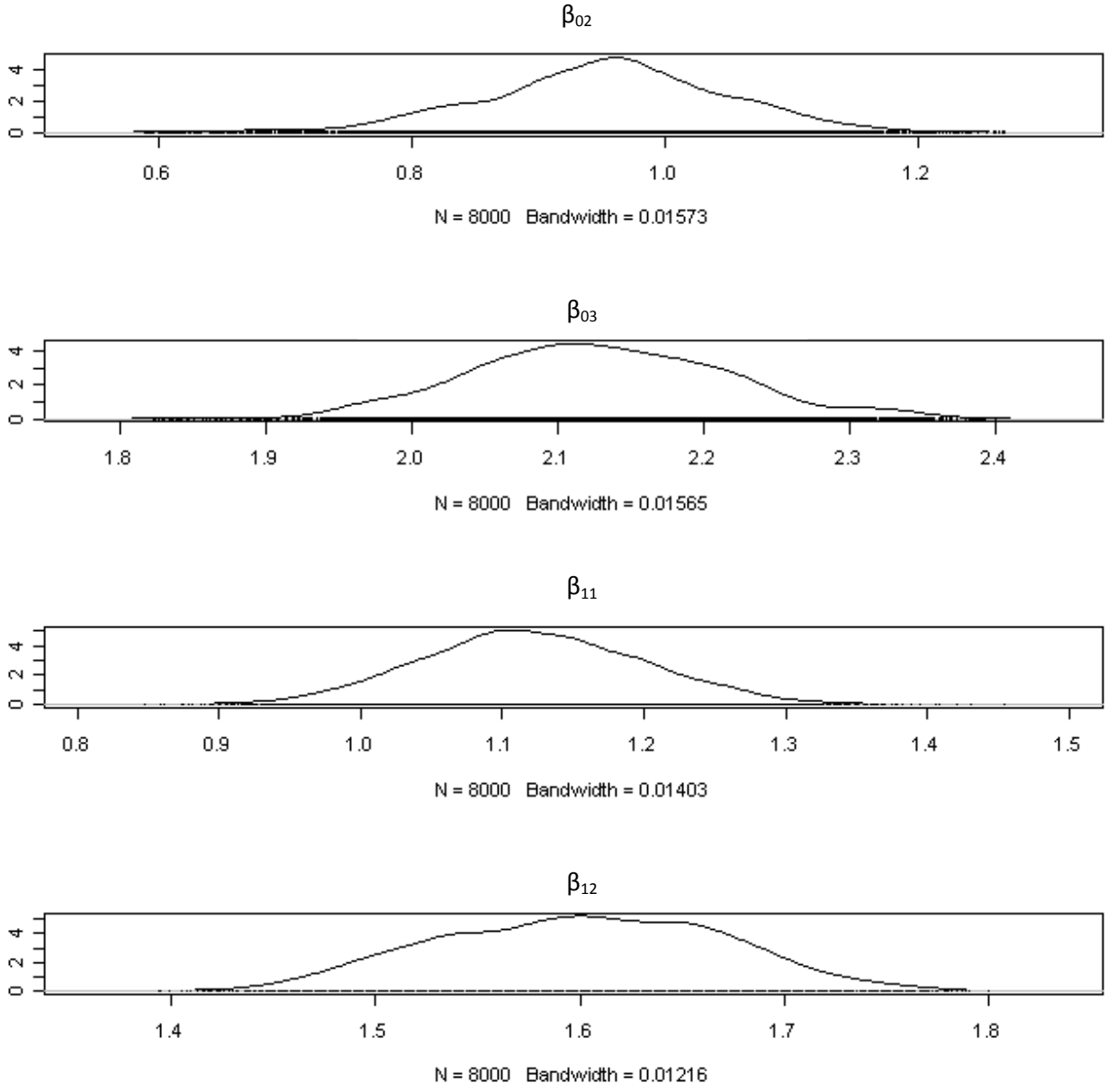


Figure 5.3 b) Density Plots for MCMC Draws of the Large Sample Example.

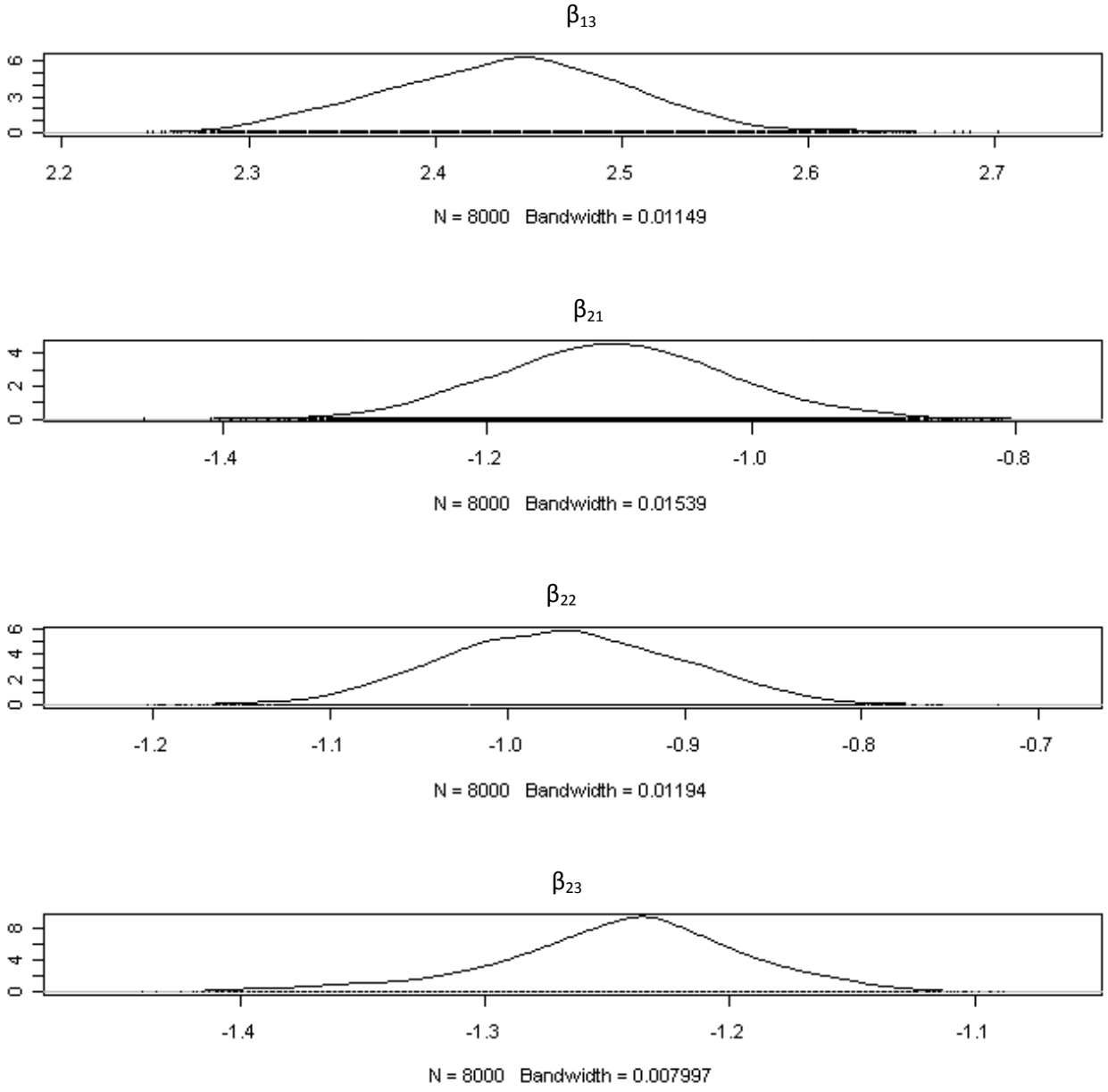


Figure 5.3 c) Density Plots for MCMC Draws of the Large Sample Example.

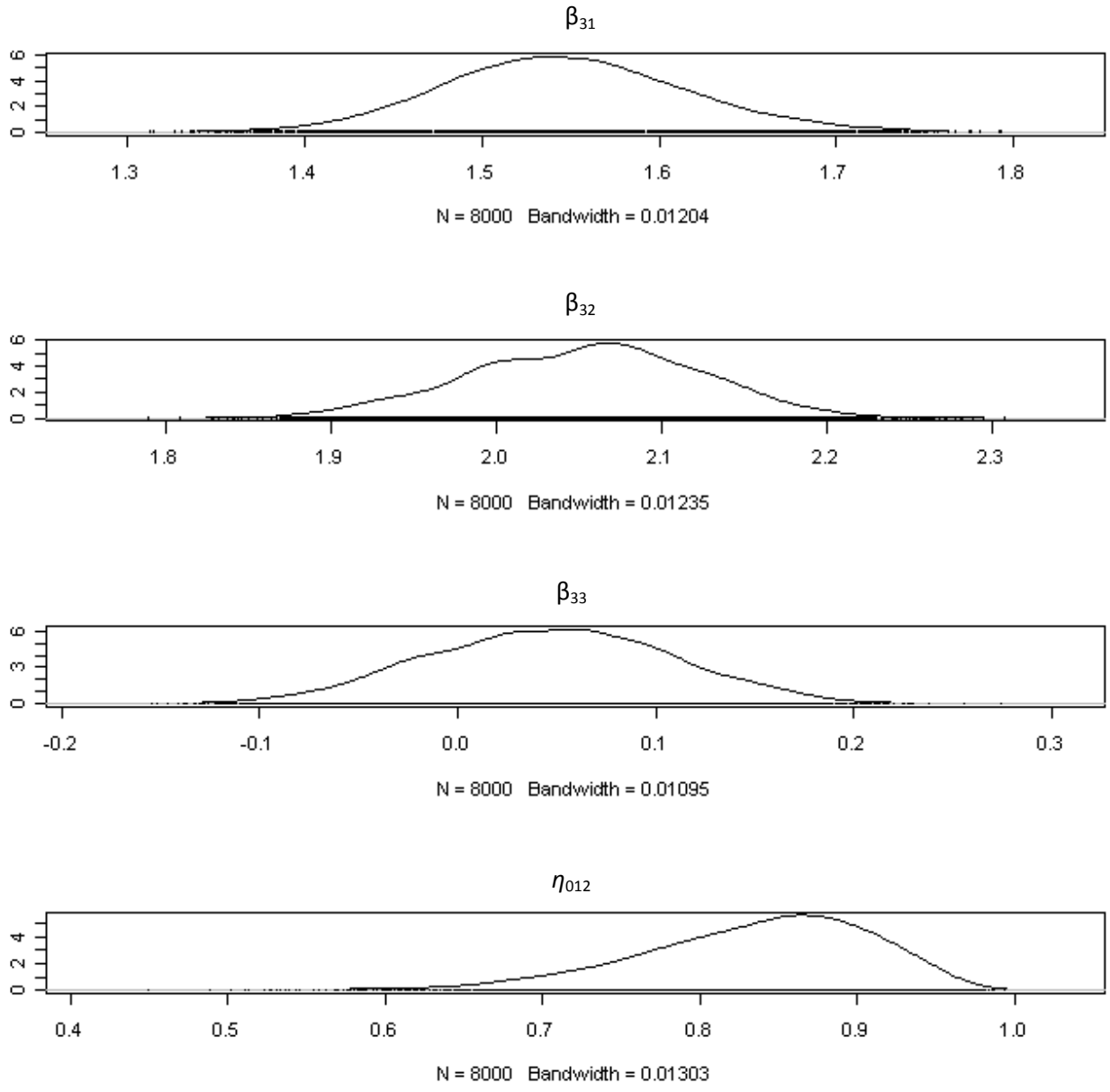


Figure 5.3 d) Density Plots for MCMC Draws of the Large Sample Example.

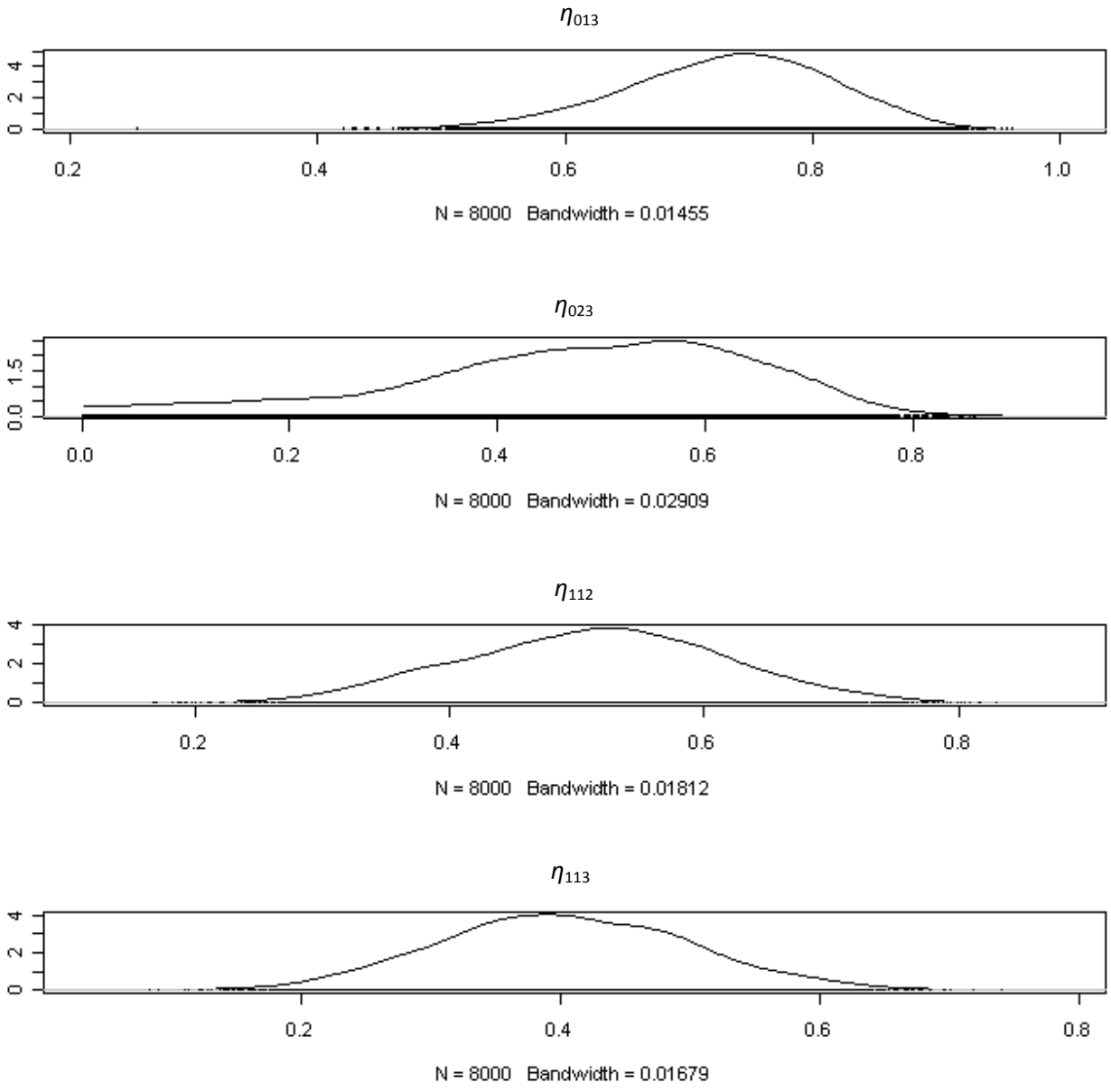


Figure 5.3 e) Density Plots for MCMC Draws of the Large Sample Example.

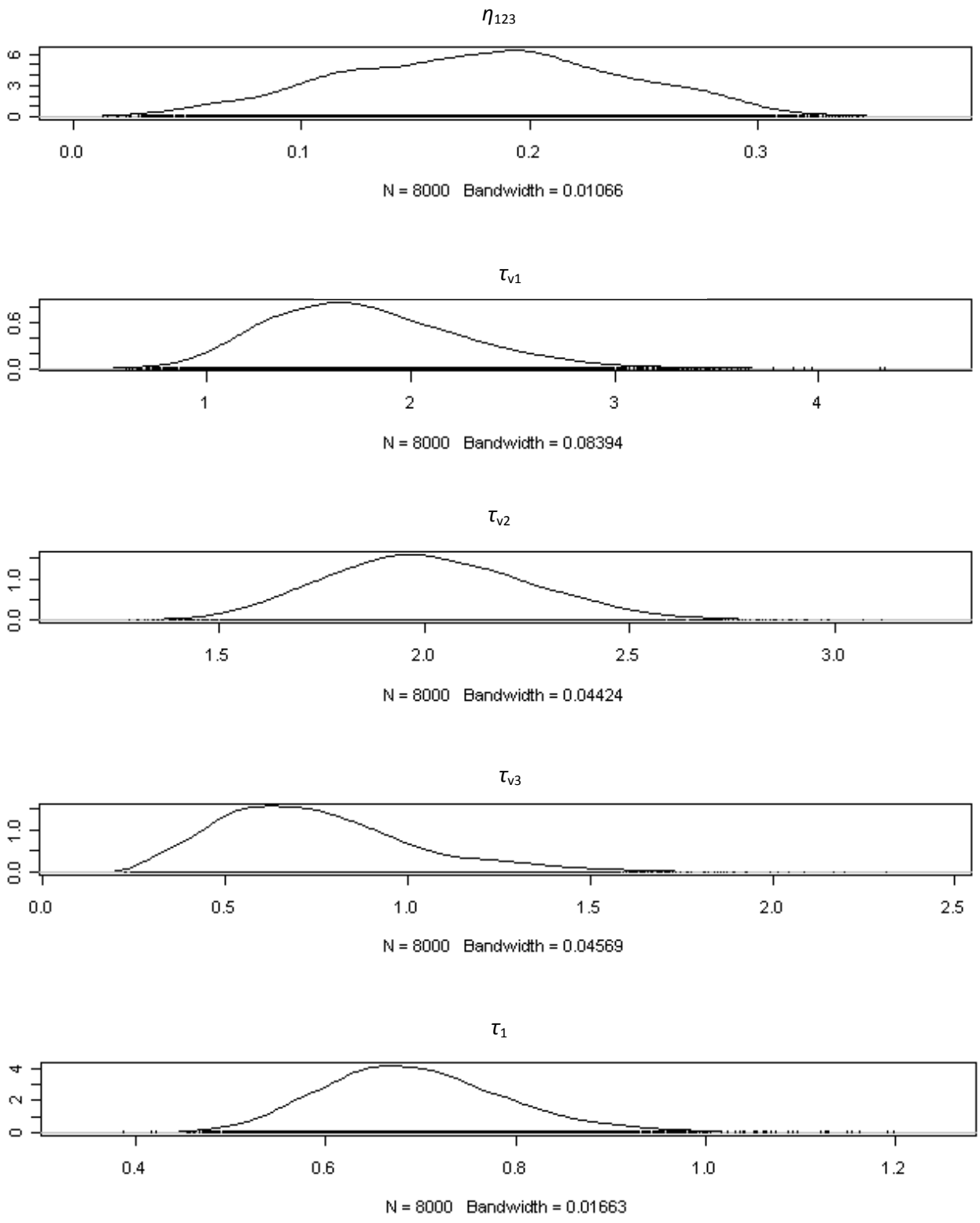


Figure 5.3 f): Density Plots for MCMC Draws of the Large Sample Example.

5.3 Results of Zone-Level Pedestrian-Crash Model

This section presents the results of the proposed MCAR model as applied to a 3-year total of pedestrian crash counts at the zone level in Austin, Texas. Walk-miles traveled (WMT) are used as a proxy for crash exposure here, since a key prerequisite for pedestrian-vehicular crashes is that pedestrians exist and are exposed by walking. Presumably, the more walking done, the higher the expected number of crashes. Since household travel surveys have relatively few respondents per zone, and walk trips are relatively rare, this study takes all respondents' walk trips per zone, scales them up to zone-level population, and builds a least-squares regression model to estimate total two-weekday WMT per zone, based on zone-level transportation, land use, and demographic variables. These estimated/smoothed WMT values were then used as the exposure measure in the pedestrian crash count model, for two count types simultaneously: severe and non-severe (as discussed in Chapter 4). This two-level model exhibited better goodness-of-fit than the other three models tested (each involving two or three different classes of crash severity [e.g., fatal crash counts versus all other crash counts]); so only results of this two-level model are presented.

5.3.1 Model for Walk-Miles Traveled (WMT)

The 2006 Austin Travel Survey (ATS) provides a glimpse of the TAZ-level walk-miles based on 569 walk trips (out of a total of 14,113 trips surveyed over a 2-day period). These walk trips occur across 217 TAZs, among which 154 zones are within the Travis County boundary and can be linked to this county's 2005 TAZ map (from CAMPO). Covariates that may influence walk-miles include zone size (in square miles), population, employment by types (i.e., basic, retail, and service), and coded lane-miles by road classes (freeway, arterial, and local streets). These covariates' summary statistics are shown in Table 5.4. The surveyed walk trips were scaled up by the ratio of zone population to zone sample size (to reflect the zone's population share) and then used as the response variable in the TAZ-based WMT model, described below. Parameters estimated from the TAZ-level WMT model were then used to impute walk-miles for each Thiessen polygon.

Table 5.4: Summary Statistics of Covariates for the Walk-Miles Traveled (WMT) Model.

		Mean	*Std Dev*	*Min*	*Max*
Response-Related Variables					
WMT (miles per zone per two-weekday period)		2,753	8,124	0	71,531
WMT per capita (miles per zone per two-weekday period divided by zone population)		0.686	0.992	0	7.200
Network					
LnMiDenFWY		4.228	6.435	0.000	44.430
LnMiDenART		8.836	6.783	0.104	51.207
LnMiDenLOC		2.435	3.770	0.000	18.932
Sidewalk (total length, in miles)		13.511	12.328	0.000	67.397
Land Use					
Entropy		0.399	0.243	0.000	0.918
# Resid. parcels near Bus Stops		304	389	0	2,255
Zone Size					
Area (sq. mi)		1.67	7.53	0.04	87
Demographics					
Population (of zone)		2,652	2,473	5	12,532
Employment Counts	Base	377	851	0	7,084
	Retail	250	270	0	1,493
	Service	791	1,252	0	8,891
N_{obs} = 154 TAZs					

Notes: WMT = total walk-miles traveled = $wmt \cdot \frac{\text{Zone Population}}{\text{Num of Residents Sampled}}$, where wmt indicates walk-miles traveled by the ATS sample population. LnMile = lane-miles, FWY = Freeway, ART = Arterial streets, & LOCAL = Local streets.

A weighted least squares (WLS) regression model for predicting total WMT in the zone yielded the best fit (R^2_{adj} = 0.51) and parameter estimates among the four model specifications attempted (ordinary least squares [OLS], WLS, Tobit, and Heckit). Table 5.5 provides a summary of these

WLS results (with statistically insignificant covariates removed). Weights were set at $\sqrt{n_i/N_i^2}$ to assure homoscedasticity of the error term, where n_i represents the number of respondents who were sampled for the i^{th} TAZ and N_i denotes the population counts for the corresponding TAZ. This weight was used because the total WMT (i.e., the response variable) is computed as the sample average times total population, with variance $N_i^2\sigma^2/n_i$.

Table 5.5: Weighted Least Squares Regression Results for Walk-Miles Traveled (WMT) Model, with Y= ln(Total WMT per zone).

Parameters	*Coef.*	*Std. Error*	*T-Statistic*
Constant	1.887	0.272	6.94
LnMiLOC	-0.068	0.027	-2.51
Area (sq. mi.)	0.344	0.123	2.80
Population	1.61E-03	2.58E-4	6.25
Sidewalk (mi.)	0.062	0.040	1.55
R^2	0.53		
Adj. R^2	0.51		
n_{obs}	154 zones		

Note: The weights are set at $\sqrt{n_i}/N_i$ for each zone.

Lane miles by road class are shown to be a significant factor in explaining walk distances per zone. So are zone size (in square miles), population counts, and sidewalk lengths. The response variable here is zone-level WMT, which is imputed by the average WMT per ATS respondent multiplied by zone's population. A WLS scheme applies because WMT values are imputed using the population scaling factor described earlier, which introduces heteroskedasticity (in error term variances). The weights are set at $\sqrt{n_i}/N_i$ for each zone.

Parameter estimates from Table 5.5 were then used to estimate WMT for each Thiessen zone, which served as the exposure measure in the MCAR model for pedestrian crash counts, as discussed in the next section.

5.3.2 Two-Response Pedestrian Crash Count Model

The parameter coefficients obtained through the WLS model (Table 5.5 values) were used to estimate walk-miles traveled in each tract-based Thiessen polygon. Note that the WLS model estimated in the previous section pertains to the TAZ level, whereas pedestrian crashes and associated explanatory variables are computed for each Thiessen polygon using *Analysis* toolbox in ArcGIS. Recall that the expected crash count, λ_{ik}, is as follows:

$$\ln(\lambda_{ik}) = \ln(WMT_i^{\alpha}) + x_i'\beta_k + \phi_{ik} + v_{ik}$$

where WMT is (an estimate of) all walk-miles traveled in the polygon (over a 2-workday period), as imputed by the Table 5.5's equation: ln(WMT) = 1.887 – 0.068 × LnMiLOC + 0.344 × Area + 0.002 × Population + 0.062 × Sidewalk.

The pedestrian crash model has two distinct crash counts, and therefore two distinctive crash-rate equations, both with the same set of (starting) covariates. Covariates include shares of residential parcels (including both single-family dwelling units and apartments) within 0.5 mile of bus stops, bus stop density, land use entropy, percentages of residential parcels within 0.5 mile of schools and commercial parcels, network density values (by roadway class and for sidewalk provision), VMT by roadway type, and demographic information (such as local population and employment densities), with summary statistics shown in Table 5.6.

Table 5.6: Summary Statistics of Covariates for the Pedestrian Crash Count Model.

	Mean	*Std Dev*	*Min*	*Max*
Transit Access				
% SFDU[a] near Transit in zone (within 1/2 mi.)	0.628	0.433	0	1
% APT[b] near Transit (1/2 mi.)	0.655	0.432	0	1
Transit Density (# of bus stops per sq. mile)	13.66	17.57	0	98.6
Land Use				
Land Use Entropy	0.647	0.229	0.037	0.989
% Resid. Parcels near Commercial (1/2 mi.)	0.759	0.304	0	1
Network Intensity				
LnMiDenFWY	4.228	6.435	0.000	44.43
LnMiDenART	8.836	6.783	0.104	51.20

LnMiDenLOC	2.435	3.770	0.000	18.93
Sidewalk Density	6.718	6.076	0.000	28.85
Vehicle Miles Traveled *(per day in 2010)*				
VMTFWY	1.59E+05	2.93E+05	0	1.52E+06
VMTART	3.22E+05	3.32E+05	937	3.61E+06
VMTLOC	1.37E+04	2.93E+04	0	2.45E+05
Demographics & Employment *(2007)* [c]				
Population Density	2,470	2,611	5	1.563E4
Basic Emp. Density	356	653	0	5,137
Retail Emp. Density	235	279	0	1,842
Service Emp. Density	598	762	1	5,308
Access to School				
% SFDU near school (within 1/2 mi.)	0.514	0.352	0	1
% APT near school (within 1/2 mi.)	0.487	0.386	0	1
Exposure Measure				
Walk-Miles Traveled (WMT[d]) (in miles over a two-weekday period)	68.80	41.26	4.79	291.3
Response Variable				
Severe Crash Counts (Fatal & Incapacitating Crashes, 2007–2009)	0.89	1.53	0	15
Non-Severe Crash Counts (Incapacitating, Possible Injury, & No Injury Crash Counts, 2007–2009)	3.23	7.4	0	100

Notes: [a.] SFDU stands for single-family dwelling units, including single family and large-lot single family dwelling units; [b.] APT denotes apartment parcels (e.g., group quarters, duplexes, apartment buildings, and condos, as defined by the City of Austin's land use archive); [c.] population and employment densities are computed as the estimated counts (by overlaying TAZ-level count information obtained from CAMPO) divided by polygon size; [d] WMT is the crash exposure measure, estimated using household travel survey data and WLS regression.

Table 5.7 summarizes parameter estimates and inferences of the pedestrian crash count model with two response levels (severe [including fatal and incapacitating-injury crashes] and non-severe [including non-incapacitating, light, possible, and no-injury crashes]), with trace plots and density plots for all parameter estimates provided in Appendix B. This dichotomous grouping of distinct crash severities was adopted (over the other dichotomous and trichotomous groupings of

the 5 crash types) because it provides the best Deviance Information Criterion (DIC[2]) value (where lower values are associated with better fit [Carlin and Louis 2009]) and highest pseudo t-statistics among the four models attempted.

Table 5.7: Parameter Estimates and Inferences of the Zone-Level Pedestrian Crash Model.

[2] The deviance information criterion (DIC) indicates the goodness-of-fit for hierarchical models, especially those where the posterior distributions are obtained by Markov chain Monte Carlo (MCMC) simulation. DIC is expressed as: $DIC = p_D + \overline{D}$, where$\overline{D} = E^{\theta}(D(\theta))$ measures how well the model fits the data (with larger value indicating worse fit), and $p_D = \overline{D} - D(\bar{\theta})$ is the effective number of parameters. The deviance function is defined as: $D(\theta) = -2\log\big(p(y|\theta)\big) + C$ where $p(y|\theta)$ is the likelihood function, y are the data, θ the unknown parameters, and C a constant that cancels out when comparing different models.

	Severe or Not?	Mean Estimate	Std Dev	Pseudo T-stat.	MC error	2.5% Estimate	Median	97.5% Estimate	Elasticity
Constant	1 (yes)	-0.652	0.169	-3.87	0.002	-0.844	-0.570	-0.463	
	2 (no)	0.462	0.142	3.25	0.002	0.300	0.458	0.621	
Transit Density	1								
	2	0.482	0.137	3.51	0.002	0.325	0.393	0.635	0.03
Land Use Entropy	1	-0.595	0.278	-2.14	0.001	-0.912	-0.432	-0.284	-0.05
	2								
% Resi. Parcels near Commercial	1	1.158	0.480	2.41	0.003	0.611	0.689	1.696	0.04
	2	0.950	0.497	1.91	0.002	0.383	0.581	1.507	0.06
LnMileDen FWY	1	0.225	0.089	2.52	0.001	0.123	0.223	0.326	0.03
	2	0.111	0.070	1.59	0.001	0.031	0.123	0.189	0.04
LnMileDen ART	1	0.607	0.189	3.21	0.002	0.392	0.574	0.819	0.47
	2	0.830	0.306	2.71	0.002	0.481	0.565	1.173	0.52
LnMileDen LOC	1	-0.259	0.089	-2.91	0.003	-0.360	-0.092	-0.159	-0.41
	2	-0.033	0.014	-2.31	0.000	-0.050	0.155	-0.017	-0.21
Population Density	1	0.208	0.136	1.54	0.001	0.054	0.203	0.360	0.04
	2	0.213	0.190	1.12	0.002	-0.004	0.234	0.425	0.08
% Resi. Parcels near Schools	1	-0.323	0.107	-3.01	0.001	-0.445	-0.270	-0.203	-0.03
	2								
Sidewalk Density	1	-0.374	0.104	-3.61	0.003	-0.492	-0.238	-0.258	-0.13
	2	-0.571	0.164	-3.48	0.003	-0.756	-0.569	-0.382	-0.22
ln(VMTART)	1	0.008	0.004	1.87	0.006	0.003	1.010	0.013	0.01
	2	0.024	0.010	2.51	0.008	0.013	1.515	0.035	0.05
ρ_1		0.728	0.127	5.71	0.002	0.575	0.724	0.873	
ρ_2		0.612	0.102	5.99	0.002	0.496	0.612	0.728	
α		0.131	0.057	2.31	0.001	0.051	0.123	0.196	
η_0		0.712	0.134	5.31	0.001	0.563	0.714	0.865	
η_1		0.312	0.076	4.13	0.002	0.226	0.312	0.398	
τ_{v1}		1.352	0.348	3.886	0.009	0.788	1.310	2.138	
τ_{v2}		2.677	0.476	5.623	0.007	1.863	2.640	3.716	
τ_1		1.653	0.495	3.342	0.007	0.852	1.615	2.707	
τ_2		2.113	0.261	8.083	0.004	1.635	2.115	2.655	
DIC	3200.5								

Mean of LogLik	-2568.1
RMSE	2.41
Run times = 59 mins; # of Iteration=15,000; Burn-in period=5,000; # of chains = 3;	

Note: "1" rows denote values for fatal and incapacitating-injury crash count prediction, and "2" rows denote parameter values for predicting other (non-severe) crash counts.

Table 5.7's elasticities were computed as the average percentage change (over the entire sample) in the mean crash counts (or expected value, λ_i) following a one-percent change in the k^{th} covariate (for each zone, *i*). These mean crash rates incorporate Eq. 3.4.2's unknown/latent error terms, as simulated for the region-specific errors, spatial autocorrelation, and correlations across various response types.

Table 5.7's results reveal noticeable spatial clustering patterns of zone-based crash counts. Severe (i.e., fatal and incapacitating) counts are estimated to have a statistically (and practically) significant spatial autocorrelation coefficient of 0.73, whereas non-severe (i.e., non-incapacitating, light, possible, or no injury) counts yield a slightly lower, but still significant coefficient of 0.61. Apart from these within-category spatial autocorrelations, statistically (and practically) significant spatial dependence emerges across the two crash-type categories: η_1 is estimated to be 0.31 and measures spatially lagged effects of cross-correlation across the two categories.

Area-level crash counts also exhibit strong correlation between the two severity levels, as measured by a statistically (and practically) significant η_0 value of 0.71. This value implies that severe and less severe pedestrian crash rates correlate in a very positive way, even *after* controlling for exposure and various other zone-level attributes. Such aspatial cross-correlation is expected, and attributable to omitted variables shared by crash types within a zone (but not across zones, as reflected via the η_1 term estimate). Examples of such missing-variables correlation (across response types) include presence of unusual site conditions (like heavy industry or entertainment zones), distinctive local lighting conditions (affecting night-time crash rates), and sight obstructions (affecting pedestrian and motorist visibility at all times). In contrast, the spatially lagged effects of cross-correlation capture missing variables that are spatially clustered but wider spread, thus affecting many nearby zones, and are shared across

crash-severity levels—such as terrain features, weather conditions, and various socio-economic variables.

The relationship between crash exposure (WMT per zone) and crash rates is estimated to be highly non-linear (with an average exponent, α, of 0.131, rather than 1 [for the linear case]), with rates (per mile walked) falling off *dramatically* as walk levels rise, presumably thanks to drivers expecting more pedestrians in high-WMT zones and responding accordingly and/or safer pedestrian environments encouraging more walking. This is a salient result: *crash rates fall substantially (per WMT) as pedestrian exposure (WMT) rises*, ceteris paribus, as shown in Figure 5.4. Also, this parameter was assumed to be identical across severity levels, based on DIC values not changing when distinctive exponents were permitted. ln(VMTs) by roadway class failed to show significance and were removed from the model.

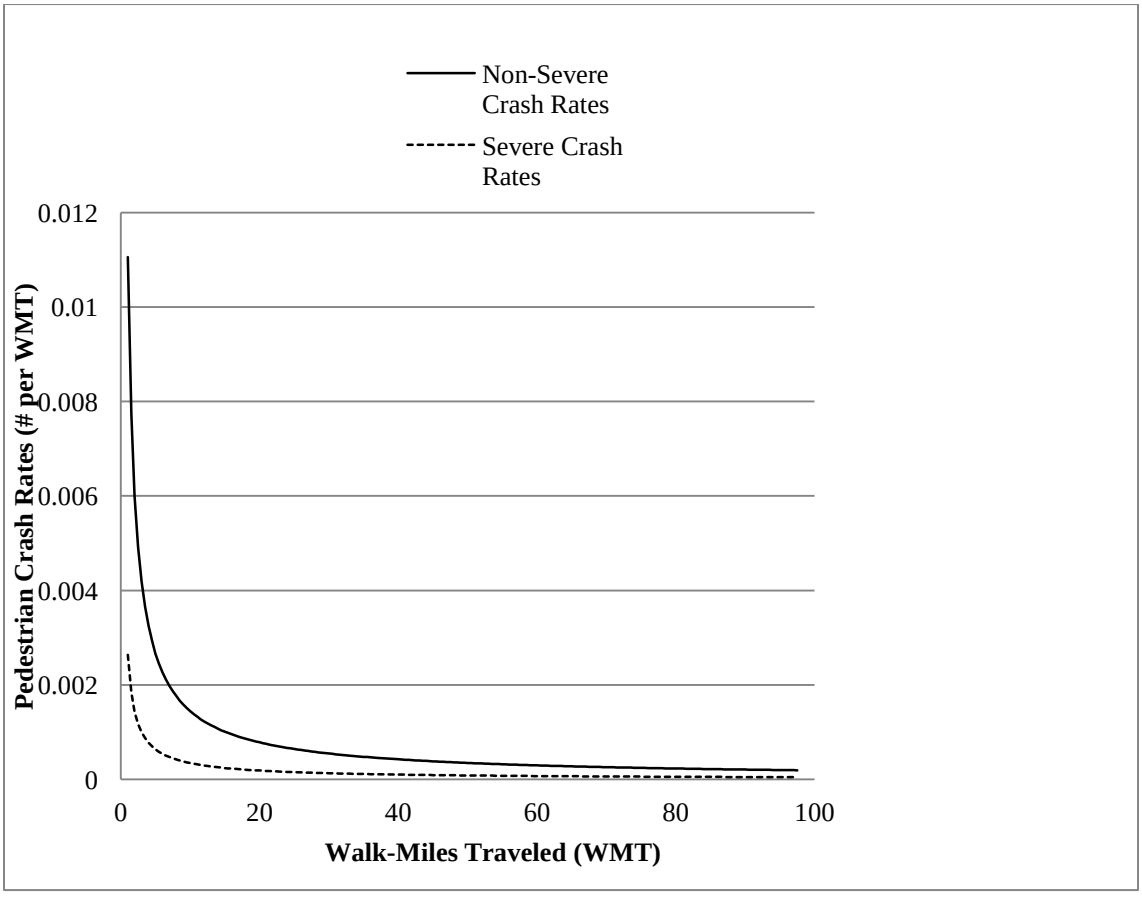


Figure 5.4: Relationship between Pedestrian Crash Rates (# per WMT) and Walk-Miles Traveled over a Two-Workday Period.

After controlling for exposure (WMT), greater land use balance was estimated to lower severe crash rates, as reflected by a negative coefficient estimate on the entropy measure. But entropy's effect on less-injurious crash rates was not statistically (nor practically) significant, so it was not included in that piece of the final specification (as reflected in Table 5.7). Shares of apartment and single-family parcels near commercial parcels showed very similar parameter estimates and so were merged to form a single covariate (the share of residential parcels within ½ mile of commercial parcels). In contrast, an increase in pedestrian crashes across *both* crash types is predicted (everything else constant) when a higher share of a zone's residential parcels are near commercial land uses, as reflected by (modest) elasticities of +0.04 and +0.06 (for the two crash types, respectively). The variance term for the non-severe crash rates is estimated to be 2.7, surpassing the variance for the severe crash rates by 1.4, as expected, since non-severe crash rates are generally higher than the rates for severe crash rates and permit greater variation in the error term.

Higher bus-stop density appears to contribute somewhat to less-injurious crash rates (after controlling for pedestrian exposure), but its effect on severe pedestrian crash rates was found to be minimal (and so was removed from the final model for that crash type). Residents' proximity to schools was found to have almost no practical effect on either crash rate (after controlling for WMT estimates in each zone), but its coefficient was statistically significant in the case of severe crash rates.

Network intensity covariates yielded mixed effects: A higher density of arterial streets is predicted to notably contribute to both severe and less-severe crashes (with elasticities of +0.47 and +0.52, respectively), whereas freeway intensity had little practical effect. Interestingly, a higher local-street density is estimated to significantly ***lower*** severe crash rates, and, to a lesser extent, non-severe crash rates. It would be useful to be able to control for traffic levels, instead of simply centerline miles, to get a better sense of how these network-design effects (arterials vs. locals) play out, in order to better anticipate an optimal balance in serving all travelers while protecting pedestrians.

Residual Spatial Autocorrelation

Residuals were computed as the difference between estimated and observed pedestrian crash counts by severity levels, shown in Figure 5.5. Both maps show negligible, positive spatial dependence, as measured by Moran's I value and the corresponding p-value to measure the statistical significance of spatial strength.

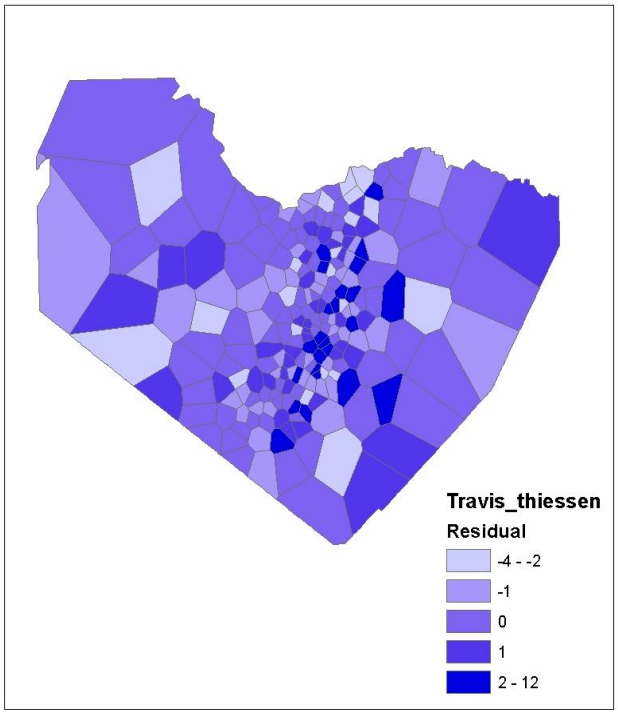


Figure 5.5 a): Spatial Distribution of Residuals for Severe Crash Counts.

Note: Moran's I = 0.013 (with p value = 0.70)

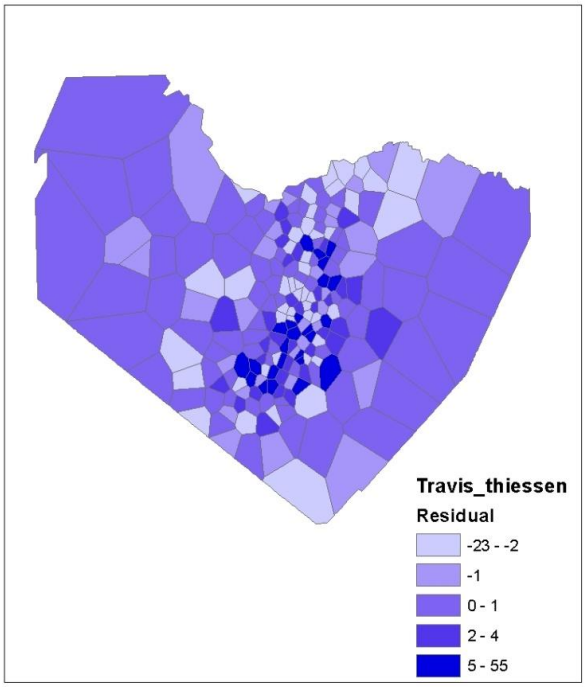


Figure 5.5 b): Spatial Distribution of Residuals for Non-Severe Crash Counts.

Note: Moran's I = 0.028 (with p-value = 0.03)

The Poisson log-normal MCAR model was also compared with an aspatial multivariate Poisson-lognormal model and a spatial Poisson-lognormal model (without correlations across different severity levels), with results shown in Table 5.8. The Poisson log-normal MCAR model yields the lowest DIC value and Moran's I of residuals, among the three models tested. Including the spatial autocorrelation effect has proved to greatly improve fit statistics, as reflected by the marked increase in the mean log-likelihood (and decrease in DIC values) after convergence was achieved. The Poisson log-normal CAR model (i.e., Model II, which incorporates spatial autocorrelation within each severity level but omits cross-severity correlation) reduces the DIC value by 10% from a pure multivariate Poisson log-normal model (Model III). Another decrease of 34% in DIC value resulted from Model I's incorporating aspatial and spatially lagged cross correlation into Model II. Similar observations were also made when comparing the root mean squared errors (RMSE) across the three models.

Table 5.8: Comparison of Full Model Results (I) to Aspatial Model (II) and Spatial Model without Cross-Correlation (III) Results.

	Poisson Log-Normal MCAR	*Poisson Log-Normal CAR*	*Poisson Log-Normal Multivariate*
Model No.	I	II	III
Parameter Constraints	-	η_0 & $\eta_1 = 0$	ρ_1, ρ_2, & $\eta_1 = 0$
DIC	3200.5	4852.31	5061.41
Mean LogLik	-2568.1	-3731.13	-3999.12
RMSE	2.41	4.21	6.70
Moran's I of Residuals for Severe Crash Counts	0.013 (p-value = 0.70)	0.132 (0.06)	0.651 (0.04)
Moran's I of Residuals for Non-Severe Counts	0.028 (p-value = 0.03)	0.192 (0.09)	0.581 (0.01)

It is worth noting that hold-out samples are often used to compare goodness-of-fit among different *aspatial* models. In the context of spatial regression models, however, it is rather unnatural to assume that spatial influences will remain the same across large geographic regions. Therefore, deviance information criteria (DIC) are used for model comparison among different *spatial* models in a Bayesian estimation setting.

5.4 Model Results for Firm Birth Counts across Counties

The study of firm births and deaths falls within the literature of organization ecology, a branch of study introduced by Hannan and Freeman (1977), who identified the four strands associated with "the population ecology of organizations": inertia and resistance to change (deemed a by-product of the need of a firm to be reliable and accountable), age dependence (i.e., the risk of mortality tends to decline as an organization grows but then shoots up once the organization has depleted its initial resources), niche theory (which explains those varying organizational structures in different industries given the two distinct sets of generalists and specialists within organizations), and density dependence (which predicts that the rates of births and deaths depend on the density of the organizations in the market).

Reynolds et al. (1995) examined firm births and deaths over a 12-year period (from 1976 to 1988) using a pooled OLS regression across U.S. counties. They concluded that higher shares of mid-career, educated adults are associated with higher firm birth rates (computed as the number of new firms divided by the total number of firms) across all time periods, and that economic diversity (measured as firm count per employee plus an occupational diversity index) also tends to contribute to higher levels of firm births and deaths. Higher personal wealth also exerted a positive impact on firm births, presumably due to an increase in consumer demand on service and retail activities. The absence of work unions and the presence of work laws were also associated with an elevated birth rate of firms in a significant way, as is higher population growth. By contrast, the following factors were shown to yield marginal or no effects on firm births: unemployment rates, input and production costs (of the economic system), national transportation access (measured by distance to the nearest airports), the size of economic systems (i.e., total population, total work force, and total number of establishments), and a host of research and development (R&D) variables (measured by numbers of post-college, professional, and technical employees; patents granted; and doctorates granted per 1000 square miles).

Alsaaty (2012) investigated the birth and death cycle of micro firms (defined as employer firms with less than 20 people) in the U.S. and maintained that the survival rate of micro firms in the U.S. is approximately 10 percent and depends on states, industries, and entrepreneurs. He also suggested several influencing factors, including the size of the state (measured by population), the budget situation of the state, the level of economic development (gauged by indicators such as the output of goods and services), and the intensity of the state's industrial (or service) base.

Brown et al. (2009) examined the birth and death of manufacturing firms for U.S. counties in the lower 48 states, using a standard negative binomial model with lagged spatial effects. In other words, spatial ripple effects enter the mean birth (or death) rates (μ_i) by $\mu_i = \exp(\beta' x_i) \prod_{j \neq i} y_j^{\rho w_{ij}} = \exp\left(\beta' x_i + \rho \sum_{j \neq i} w_{ij} \ln(y_j)\right)$, where y_j is the observed birth (or death) counts for location j. Results suggested that agglomeration economies, including both local agglomeration (measured by the percentage of manufacturing establishments with less than 10 employees and shares of employment in each county) and economies of scale (captured by the percentage of manufacturing establishments with more than 100 employees and the total

number of establishments divided by county area) tended to be associated with increases in firm births. Higher median household income was also estimated to contribute to firm births.

This dissertation applied the MCAR model to analyze the number of new firms by industries (i.e., base, service, and retail) across counties in the lower 48 states. To group new firms by industries is not only of theoretical interest (so that the research problem can be cast in a multivariate setting) but also a study in behavioral realism, as different industries tends to exhibit different organizational structures and growth patterns (Hanna and Freeman, 1977).

The following sections discuss the results of the county-level firm births model for a selected sample of the 3,109 U.S. counties in the lower 48 states.

5.4.1 Modeling Results of the Firm-Birth Model

A trivariate extension of the new MCAR model was applied to analyze firm births during the 2008–2009 period for a sample of 1,316 counties in the midwestern and western regions of the U.S., as shown in Figure 5.6.

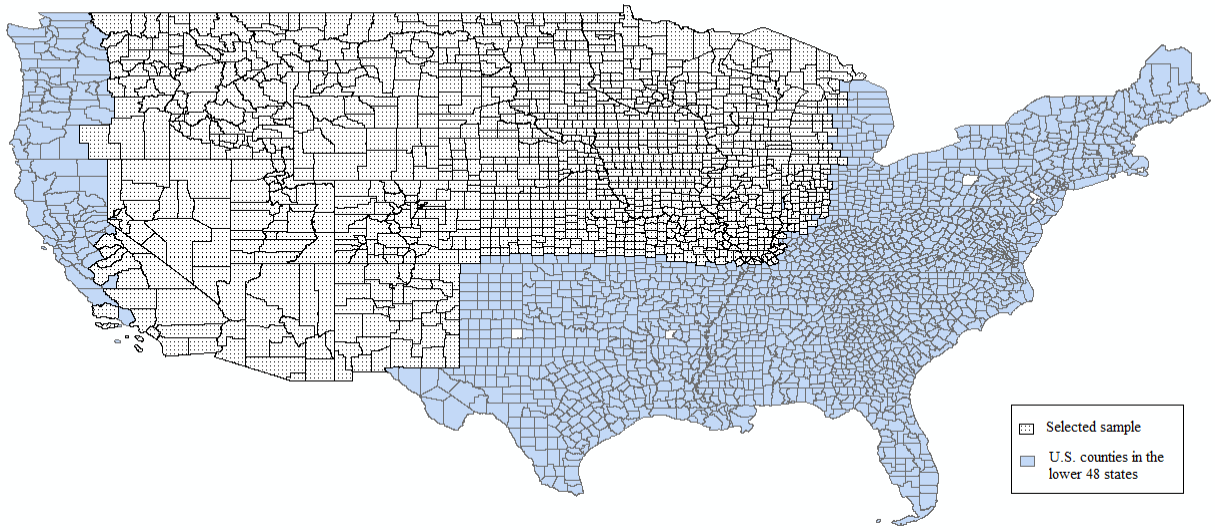


Figure 5.6: Selected Sample for the Firm Birth Model.

As with the model for pedestrian crash counts, a size or exposure measure (like county population or land area) can be used in the firm birth model. Alternatively, simply using the natural logarithm of size-variable candidates as covariates is feasible, since one does not

normally expect or find simple proportionality (between firm births or crash counts and population, for example). Here, the exposure measure used in the trivariate firm birth model is the total number of business establishments in the same industry category (as the response variable of interest) at the end of 2008. The unknown parameters (i.e., the exponents γ) associated with these three size or exposure measures were estimated simultaneously with all other model parameters, using Bayesian techniques.

Summary statistics for all covariates (including exposure metrics) use are summarized in Table 5.9. Covariates include shares of vacant housing units (VacntHous), median age of county residents (MedAge), shares of families living below the poverty line (PrvtyFamil), median annual income of working people older than 16 years, and population density (PopDen). As mentioned earlier, the three response variables (per county) are the number of new firms in basic, retail, and service industries in 2009. Table 5.10 presents the model's results

Table 5.9: Summary Statistics of Covariates for the Firm-Birth Model.

	Mean	Stdev	Min	Max	Median
Potential Exposure Measures					
Land Area (sq. miles)	950.84	1304.02	2.00	20,057.11	610.31
Total Establishments	2,180	7,248	1	218,787	517
Total Jobs	37,476	138,060	0	3,909,096	6,741
Covariates					
VacntHous (share)	0.162	0.098	0.038	0.074	0.013
MedAge (years)	40.37	5.03	22.40	62.70	40.30
PvrtyFamily (share)	0.11	0.06	0.00	0.45	0.10
MedAnnIncome per Worker	25,466	5,047	5,559	59,672	24,934
MetroIndex (indicator)	0.35	0.48	0.00	1.00	0.00
PopDensity (persons per acre)	0.40	2.70	0.00	108.51	0.07
Response Variables					
New Basic Estab.	23	99.25	0	3798	6
New Retail Estab.	25	88.44	0	2402	6
New Service Estab.	153	574.19	0	16,931	28

Table 5.10: Model Results for the Firm Birth Model with Three-Level Responses.

	Parameter	Estimates	Pseudo-T-Stats	2.5%	Median	97.5%	Elasticity
Constant	Basic	-7.441	-8.09	-8.934	-7.333	-5.777	
	Retail	-3.792	-7.00	-4.712	-3.832	-2.816	
	Service	-2.833	-2.66	-4.708	-2.998	-0.756	
Exposure term (New Firms): y	Basic	0.884	9.59	0.596	0.823	0.943	
	Retail	1.068	28.5	0.990	1.076	1.124	
	Service	1.015	27.8	0.940	1.019	1.080	
VacntHous	Basic	-0.041	-1.98	-0.131	-0.052	0.003	-0.05
	Retail	0.011	2.06	0.001	0.011	0.021	0.02
	Service	0.010	2.18	0.001	0.010	0.019	0.02
MedAge	Basic	-0.006	-1.95	-0.024	-0.006	0.014	-0.003
	Retail	-0.016	-2.37	-0.029	-0.015	-0.004	-0.03
	Service	-0.031	-3.16	-0.059	-0.024	-0.005	-0.05
PvrtyFamily	Basic	1.861	1.71	0.772	2.171	4.338	0.07
	Retail	1.481	1.65	-0.299	1.527	3.093	0.06
	Service	-0.875	-1.08	-2.423	-0.893	0.740	-0.02
Metro Index (1=Metro.)	Basic	0.051	2.64	0.028	0.043	0.104	0.07
	Retail						
	Service	0.108	1.75	-0.011	0.107	0.232	0.05
PopDen	Basic	0.174	3.30	0.071	0.174	0.279	0.07
	Retail	-0.114	-2.07	-0.222	-0.114	-0.010	-0.02
	Service						
ρ_1		0.664	6.53	0.452	0.671	0.843	
ρ_2		0.499	5.05	0.309	0.499	0.694	
ρ_3		0.717	7.10	0.500	0.725	0.891	
η_{012}		0.380	3.74	0.196	0.376	0.582	
η_{013}		0.320	3.73	0.166	0.315	0.502	
η_{023}		0.348	3.87	0.181	0.345	0.534	
η_{112}		0.230	4.13	0.128	0.229	0.347	
η_{113}		0.127	3.19	0.059	0.124	0.215	
η_{123}		0.128	3.16	0.060	0.124	0.214	

τ_{v1}	12.510	11.1	10.410	12.470	14.790	
τ_{v2}	6.835	8.92	5.434	6.798	8.449	
τ_{v3}	7.051	9.05	5.632	7.014	8.669	
τ_1	4.555	6.26	3.305	4.490	6.153	
τ_2	4.813	6.21	3.466	4.755	6.487	
τ_3	5.180	7.31	3.947	5.135	6.696	
DIC	7687.3					
Run Time: 10.1 hrs	# of chains=2, burn-in=5000, # of iterations=10,000					

Firm growth seems to exhibit a strong clustering pattern across all three industries. This result is as expected, thanks to lowered production costs (as more firms in related industries cluster together) and increased attractiveness (as competing firms in the same industry cluster, which invites more suppliers and customers than a single firm could). Shares of vacant housing appear to contribute to an elevated number of new firms in the retail and service industries while exerting a negative effect on firm births in the basic industry. A plausible explanation is that owners of vacant housing may solicit retail and service business and thus create mixed-use development; however, more vacant housing units indicate a smaller population (and presumably a smaller worker population), meaning firms in the basic industry category lack the luxury of abundant workers. Older median age is associated with fewer new firms across all three types of industries, which suggests that a workforce's vitality (measured by age) is also a strong factor in firm births. MetroIndex indicator variable tends to be positively associated with higher birth rates of service and basic establishments (with 1 denoting a metro area). Population densities are negatively correlated with retail firm births, probably due to competition between housing and retail as more housing units are occupied by residents (rather than being converted into mixed-use developments). As expected, the three industries exhibit positive cross-correlations and positive, spatially lagged cross-correlations, supporting the notion of agglomeration economies.

This application of the MCAR model for a trivariate response variable with a relatively large sample size achieved success. Application to the *full set* of U.S. counties (n = 3,109 counties) is left for future work, due to the daunting computing efforts that this multi-faceted model for discrete responses requires (relating to inversion of the large covariance matrices when simulating the MCMC chains).

5.5 Chapter Summary

This chapter first described the results of two initial simulation studies, using the MCAR model developed in Chapter 3, and then presented in greater depth the empirical results for prediction of pedestrian crash counts across Austin neighborhoods and firm births across much of the U.S. The initial simulation studies include a bivariate example built on 218 Thiessens polygons (created for the 3-year pedestrian crash count totals, to avoid missing crash points that occurred along census tracts' boundaries and at tract corners) and a trivariate example using a larger data set (n = 1,316). After achieving coding success with the simulations, two empirical applications illuminated the presence of much positive spatial correlation and clustering across Austin neighborhoods and U.S. counties for the very different data contexts (crashes versus firm starts), along with many interesting interpretations of parameter estimates. The next chapter revisits key findings and summarizes the contributions and limitations of this dissertation.

The two empirical studies indicates positive spatial clustering patterns for both pedestrian crashes and firm births, along with many interesting interpretation of parameter estimates. The next chapter will revisit several key findings and summarize the contributions and limitations of this dissertation.

CHAPTER 6: CONCLUSIONS

This chapter summarizes the major findings and contributions of this dissertation, including benefits and limitations of the proposed MCAR model, and identification of future pathways for related contributions.

Existing literature for spatial count models in crash prediction, disease mapping, and other geo-coded count-data contexts (such as species distribution)is still quite limited and tends to rely on an improper CAR prior, leaving out a spatial autocorrelation coefficient term in the covariance matrix (Wang et al. 2009 and Song et al., 2006). There are numerous drawbacks associated with this prior structure, including an improper joint posterior distribution (Gelfand and Vounatsou 2003), which can be mitigated by imposing a linear constraint on the spatial random terms at each iteration of the Gibbs sampler, as discussed in Chapter 3 of this dissertation. A more serious concern that the functional form of the joint distribution of those spatial random terms is not identified when the CAR model's "precision" parameter (the inverse of the variance term, σ^2, as discussed in Chapter 3) is unknown (which is almost always the case). Another concern is that this type of CAR structure provides no information about the overall spatial autocorrelation, due to the omission of the spatial autocorrelation coefficient.

Only a few studies have attempted to incorporate the spatial autocorrelation coefficient in the multivariate count model framework. Mardia (1988) cast the question as a series of multivariate conditional distributions but was hindered by computational difficulties (at that time). Gelfand and Vounatsou (2003) revisited Mardia's specification but still encountered substantial computing times. Their model assumed that the spatial random terms followed a multivariate normal distribution: $\boldsymbol{\phi} \sim MVN(\mathbf{0}, [(D - \rho W) \otimes \Lambda]^{-1})$, where the *K* by *K* matrix Λ describes the aspatial cross-correlations among the *K* response types. This specification comes with significant computing-time costs, due to the required matrix inversion at each iteration of the MCMC draws. Building from the covariance matrix (rather than the precision matrix), Jin et al. (2005) devised an MCAR model that serves as a more general form of the model studied by Mardia (1998) and Gelfand and Vounatsou (2003) and improves computing times.

This dissertation proposed, developed, coded and then estimated a Poisson log-normal multivariate CAR model, along the thread of Jin et al.'s model (2005), for count data over space,

to reflect not only site-specific heterogeneity, but also correlations across response types (e.g., severe and non-severe crash rates) and spatial dependence associated with latent heterogeneity (i.e., missing variables). In this way, it is more general and thus flexible than Jin et al.'s model, while also proving itself estimable/computationally practical, at least for reasonable sample sizes and response-vector lengths (as evidenced here) on a desktop computer.

6.1 The Austin Application

For the first application, with Austin's pedestrian crash counts (by severity), Thiessen polygons – rather than Census tract boundaries – were used to aggregate the count data. The GIS-based approach helped ensure that high-crash locations could be uniquely assigned to polygon zones (rather than arbitrarily assigned to or split across adjacent tracts). In contrast, the dissertation's county-level firm birth example was able to adhere to original U.S.-county boundaries, since firms rarely operate on the edge of a county.

This dissertation's new spatial model was able to analyze the relationships between zone-level pedestrian crash counts and various land use, network, and demographic factors, including residents' proximity to schools, land use balance (measured by entropy), transit access, network densities (by roadway class), sidewalk provision, and resident demographics. Interestingly, after controlling for these other variables, local job-count variables did not come out as relevant, presumably due to their reflection already in the model's exposure term and the various covariates used.

Walk-miles traveled (WMT) per zone were used as the sole exposure measure (since VMT estimates on CAMPO-coded links were not helpful to prediction) and imputed using the 2005–2006 Austin Travel Survey's walk trips. Parameter estimates suggest that crash rates fall dramatically at first, as WMT levels rise: from roughly one reported (pedestrian-vehicle) crash every 3 years with daily WMT values of just 0.002 miles per resident (i.e., just one person out of every 500 present in the zone walking just 1 mile a day [and others logging zero miles that day]), to one crash every 100 years with daily WMT values of 0.008 mile per person (e.g., one person out of every 125 people logging a mile of walking that day, and others logging zero). Higher pedestrian crash risks across both severity levels (after controlling for WMT) were found to be associated with higher shares of residential parcels within one-half mile of commercial parcels (presumably due to such mixtures creating more conflict opportunities across modes) and with

higher lane-mile densities of Austin's arterials and freeways. A better balance (higher entropy) of land use appears to reduce severe-crash rates, ceteris paribus (including WMT or pedestrian exposure), but its effect is not practically significant in these data.

Pure, positive spatial autocorrelation (indicating clustering patterns) appears present across Austin neighborhoods, as expected (due to measurement errors that trend in space and the spatial clustering patterns of crash counts). The spatially lagged effects of cross-response correlation (estimated to be statistically and practically significant) capture missing variables that are both spatially clustered and shared across crash types, such as socio-economic variables (like ethnicity and poverty). In contrast, the model's aspatial cross-correlation ($\eta_0 = 0.712$) represents omitted variables that are meaningful for both crash-severity levels but apply within zones, more locally (like relatively poor lighting conditions and the presence of unusual sight obstructions).

From a planning and policy perspective, this paper's results reinforced the importance of advocating walking in order to reduce crash rates, as reflected by the drastic decrease in crash rates as walk miles traveled increase. Providing walking facilities (such as sidewalks and other pedestrian paths) and greater local street intensity for all road users may also reduce crash rates, per walk-mile traveled, as suggested by the conspicuous elasticity estimates for sidewalk and local-street provision in the pedestrian crash model's results. In addition, balanced land development offers a mild, positive impact in reducing severe crashes and could serve as a countermeasure to curb pedestrian fatalities. Other countermeasures may include providing pedestrian signals that count down (to warn walkers of time remaining), pedestrian (and cyclist) overpasses/underpasses, walk beacons at popular mid-block crossings, pedestrian phases that turn on before the green signal for vehicles (crossing in the same direction), and more safety programs for vulnerable road users (like school children and disabled pedestrians), while restricting parking near intersections, as suggested in Zegeer and Bushell (2011).

6.2 The U.S. Firm Birth Application

The main purpose of this dissertation's firm-birth example was to showcase the model's applicability for larger spatial data sets (N_{obs} = 1,316 U.S. counties) with a higher dimension of response (three firm-type counts, rather than just two crash counts). The firm-birth model's results also help describe firm growth patterns (in basic, retail, and service industries), across the nation's midwestern and western regions. The model's exposure tern was the total number of

establishments in the same industry category as the response variable of interest at the end of 2008.

Like many economic and other phenomena, higher counts of new firms tend to be spatially clustered. Such patterns may reflect the underlying mechanisms that govern firm growth: related firms may cluster thanks to lowered production costs (agglomeration economies) arising from such intensity; they may also arise in the presence of inter-firm competition, attracting more patrons while sharing suppliers.

Interestingly, county higher shares of vacant housing appear to be associated with more new firms in the retail and service industries, but fewer starts of basic-industry firms. After controlling for the natural log of existing-firm counts (in 2008), a younger (and possibly more vital) work force and/or clientele (as quantified by each county's median-age values) was associated with more firm births (in 2009) in all three categories (with elasticities of -0.003, -0.03, -and -0.05 on median age, for basic, retail, and service firm births, respectively). Rising population densities appear to contribute new basic firms, while reducing retail-firm starts. As with the pedestrian crash count applications, new firm counts exhibit positive aspatial cross correlations (across industries) and positive, spatially lagged cross correlations, suggesting the presence of missing variables at both local (within-county) and cross-county scales, as expected.

6.3 Opportunities for Model Enhancements

The new model's incorporation of spatial and cross-response effects for count data noticeably improved inference and fit statistics, as shown in Table 5.8 (for the pedestrian crash count data), where comparisons were made among a full-blown MCAR model, an aspatial multivariate model, and a spatial model with no cross correlations. The model developed here starts with Jin et al.'s (2005) specification and presents a novel alternative to the Poisson MCAR model proposed by Gelfand and Vounatsou (2003) and Song et al. (2006), thanks to its more intuitive parameterization of the spatial influence (by focusing on the covariance matrix rather than the precision matrix), its ability to distinguish aspatial cross correlations from spatially lagged relationships in the data, and its faster computation/parameter-estimation times.

As with most any model in use, several enhancements can be pursued here – in both the theoretical and empirical domains. In terms of empirics, more network variables may prove quite

useful for crash-count predictions; these may include at-grade intersection density values and more accurate estimates of VMT values (by roadway type) for zone-level exposure values. (Such flow values, and thus VMT estimates, are currently missing for nearly all local streets and many segments with higher functional classifications in the Austin network.) In terms of the firm birth model, the nature of a county's transportation transport access, as measured by the presence of interstate lane-miles and proximity of international airports, for example, would be helpful, along with more information on existing firms (e.g., their size and profitability).

In terms of theory, several improvements can be pursued in future research. First, spatial autocorrelation can occur in multiple ways, through correlated error terms, response variables, or covariates. Here, the proposed MCAR model assumes that spatial interactions are carried by the latent/unobserved error terms, a situation that alludes to subtle spatial interactions (mainly due to missing variables that happen to be spatially correlated, and in some cases due to measurement errors). Other spatial structures may also prove meaningful for various data sets, such as one that incorporates spatially lagged covariate terms (e.g., the spatial Durbin model [LeSage and Pace 2009] and possibly a CAR variation with spatially lagged covariates). This line of work may offer practical benefits for planning and policy perspectives, since it is often helpful to know what types of variables generate what kind of spatial autocorrelation: a pure within-severity-level dependence, a spatially lagged cross-severity correlation, or perhaps only an aspatial cross-response correlation.

Second, a temporal extension should be pursued to identify basic time trends and the presence of temporal autocorrelation in the data (such as declining, autoregressive rates of pedestrian crash counts and firm births). Temporal effects can enter the specification through an autoregressive (AR) component, such as an AR(1) structure (with a one-period time lag in the response values or the error terms). Alternatively, one could adopt the time-space filter introduced by Parent and LeSage (2010) for their model of continuous commute times, mathematically expressed as the following:

$$Q\varepsilon = v \text{ and } Q = (I_T - \phi L) \otimes (I_N - \rho W) = I_{NT} - \rho I_T \otimes W - \phi L \otimes I_N + (\phi\rho) L \otimes W$$

where $\varepsilon = (\varepsilon'_{\cdot 1}, \varepsilon'_{\cdot 2}, \ldots, \varepsilon'_{\cdot T})'$ is an NT×1 vector, $\varepsilon_{\cdot t} = (\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{Nt})'$ is an N×1 vector of time-space error terms for a total of N geographic units at time *t*, I_T is a T by T identity matrix (with

T indicating the number of time periods in the data set), L represents a first-order time lag operator -- expressed as a T by T matrix: $\begin{bmatrix} \varphi & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & ... & 1 & 0 \end{bmatrix}$, W denotes a row-standardized N by N time-invariant weight matrix (defined by contiguity or distances and treated as known/exogenous in the model), v is an NT by 1 uncorrelated error term assumed to follow a normal distribution: $N(0, \sigma_v^2 I_{NT})$, and $\otimes$ represents the Kronecker product of two matrices. The parameter φ denotes the initial period value and can be treated as endogenous or assumed known (Parent and LeSage 2008). Other parameters to estimate include ρ, ϕ, σ_u^2, and σ_v^2, where ρ measures the strength of spatial autocorrelation in the error term, ϕ denotes the strength of temporal correlation, σ_u^2 represents the magnitude of heterogeneity (which leads to overdispersion), and σ_v^2 is the variance of the uncorrelated white noise after both spatial and temporal correlations are filtered out.

Some fundamental questions also exist here, relating to the propriety of the posterior distribution. As Nobile (2000) noted, in generalimproper posteriors (or posteriors that are not bounded in the real domain) may result even when the conditional posterior distributions are well defined, a problem exacerbated by the fact that the trace plots of the Gibbs sampling output may not be able to hint at posterior impropriety (see Hobert and Casella [1996] for more details). Even though the estimation results seem successful, a theoretical proof of the propriety (boundedness) of the joint posterior would be helpful, to further validate the MCAR specification. This is a recurring question for many) complex model estimates using Bayesian methods , but deemed dismissible given the fact that the model can successfully uncover true parameter values in simulation tests.

6.4 Final Thoughts

This dissertation successfully estimated a spatial multivariate model for count data that allows for zone-specific heterogeneity, spatial autocorrelation, and aspatial and spatially-lagged cross-correlations across response types, with case-study applications to zone-level pedestrian crash counts (by severity) and tract-level firm births (by industry). This model contributes to the relevant literature by introducing a flexible correlation structure that permits both spatial and aspatial cross correlations among different response types for the first time, while imposing less computational burden than past model proposals (Mardia 1988, Gelfand and Vounatsou 2003).

For example, when specifying the covariance structure of the multivariate normal distribution (for the spatial random terms), it is useful to start with the covariance matrix rather than the precision matrix, since this provides more behavioral realism and reduces run times (thanks to avoiding matrix inversions at each iteration of the MCMC chains).

This dissertation opens several doors for extensions. For example, it will be useful to analyze panel data (over space), with temporal effects entering the model via an autoregressive component or other forms of interaction, over time and space (as discussed in 6.3).

This dissertation's contributions are applicable to any setting that involving physical space or other forms of connected data (like social networks), with count response data. Examples include vehicle ownership levels (by vehicle type) or activity counts by households and/or firms across parcels or Census tracts, trip making by individuals or households or firms, instances of disease or plant species by zone, and so forth. The world is full of count data, and nearly every data set has spatial features to it, since observational units (largely) exist in space. Data points that are reasonably close in space tend to share unobserved qualities, a key feature of this new model.

APPENDIX A.
R AND WINBUGS CODES FOR THE SIMULATED AND EMPIRICAL MODELS

This appendix provides the core codes (written in R and WinBUGS) for the new multivariate CAR model. Combining the use of R and WinBUGS represents an effective way to estimate relatively complex Bayesian models, helping speed adoption of such models across a variety of research areas, including economics, epidemiology, ecology, and transportation.

Only core codes that represent the major contribution of this dissertation are presented here. Codes for graphic outputs and diagnostic tests can be found in R functions and packages (e.g., Coda). Here, R codes contain the procedures to compute contiguity-based weight matrices using a shapefile and other input parameters for the proper.car function (which specifies the proper CAR structure) in WinBUGS, whereas WinBUGS codes are used for parameter estimation.

Pedestrian Crash Count Model Code

```
library(spdep)
library(R2WinBUGS)
library(MASS) # in order to call the mvrnorm() function

#read shapefile
shape <- readShapePoly("C:/projcted.shp", IDvar="IDn")
shape_nb <- poly2nb(shape, queen=FALSE)
N= length(shape_nb)  #total number of geographic units
num=sapply(shape_nb,length) # number of neighbors for each unit
adj=unlist(shape_nb) # neighbor IDs for each unit
sumNumNeigh=length(unlist(shape_nb))
W = nb2mat(shape_nb,style="B")  # style= B-binary coding. W- row-standardized
W1=nb2mat(shape_nb,style="W")

s=1
C=rep(0, sumNumNeigh)
for (i in 1:N) {
    for (j in 1:N) {
         if (W1[i,j] !=0) {C[s]= W1[i,j]
         s=s+1}
}
}

ped=read.csv("C:/peddata.csv", header = TRUE)
Y10=ped[,13]   # fatal & incapacitating crash counts
Y20=ped[,15]   # non-incapacitating, possible, no injury crash counts
x1=ped[,1] #AptBusShare
x2=ped[,2] #SingHomBusShare
x3=ped[,3] #wSpdLimit
x4=ped[,4]  #JxnDen
```

```
x5=ped[,5]   #SWDen
x6=ped[,6]  #RWDen
x7=ped[,7]  #LessThan19Ys
x8=ped[,8]  #OlderTh65Ys
x9=ped[,9]  #ResiPerc
x10=ped[,10] # Log of tract area

# run a simple Poisson regression to get the initial values of beta's
poi1.data <- data.frame(Y10, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10)
poi1<- glm(formula=Y10~ x1 + x2 +x3 + x4 +x5 + x6 + x7 +x8 + x9 +x10,
family="poisson")
beta1.init <- poi1$coefficients

poi2.data <- data.frame(Y20, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10)
poi2<- glm(formula=Y20~ x1 + x2 +x3 + x4 +x5 + x6 + x7 +x8 + x9 +x10,
family="poisson")
beta2.init <- poi2$coefficients

data <- list(Y10=Y10,Y20=Y20, N=N,adj=adj, sumNumNeigh=sumNumNeigh, num=num,
W=W, x1=x1, x2=x2, x4=x4, x5=x5, x6=x6,x8=x8, x9=x9, x10=x10,
beta02=beta2.init[1], beta12=beta2.init[2], beta22=beta2.init[3],
beta42=beta2.init[5], beta52=beta2.init[6], beta62=beta2.init[7],
beta82=beta2.init[9], beta92=beta2.init[10], beta102=beta2.init[11],   C=C)
inits <- function() {
list(beta01=0.1, beta11=-0.2, beta21=0, beta41=0.1, beta51=-0.2, beta61=0,
beta81=0, beta91=-0.2, beta101=0, tau1=1, tau2=2, tau.v=0.5, eta_0=0.8,
eta_1=0.6, alpha=c(0.8,0.6), phi1=rep(0,N), phi2=rep(0,N), V=rep(0, N) )}
parameters <- c("eta_0", "eta_1", "beta01", "beta11", "beta21", "beta41",
"beta51", "beta61", "beta81", "beta91", "beta101", "tau2", "tau1","alpha",
"tau.v")
result <- bugs(data, inits, parameters,
model.file="C:/codes_heterogeneity_ped/model_hetero_3.txt", n.chains=1,
n.iter=5, n.burnin=0,n.thin=1, codaPkg=FALSE, DIC=FALSE, debug=TRUE,
bugs.directory="C:/Users/yw3534/Desktop/winbugs14/WinBUGS14")
```

Firm Birth Model Code

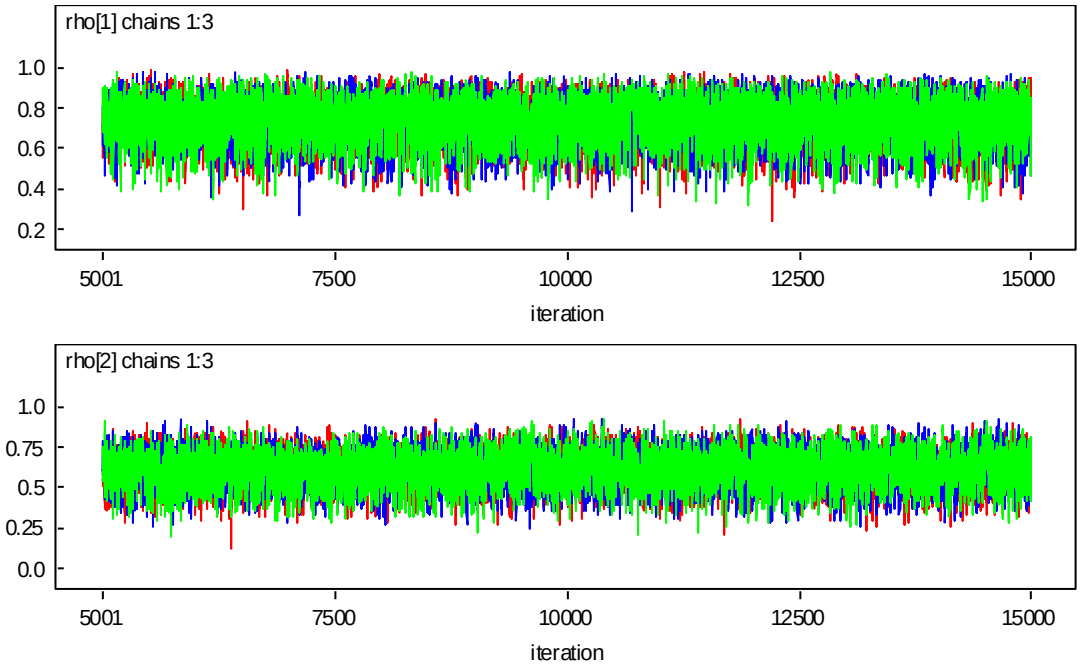
```
library(spdep)
library(R2WinBUGS)

#read shapefile
shape <- readShapePoly("C:/Users/yw3534/Desktop/New
folder/Graduates_soon/tl_2012_us_county/uscounty467.shp", IDvar="nID")
shape_nb <- poly2nb(shape, queen=FALSE)
N= length(shape_nb)  #total number of geographic units
num=sapply(shape_nb,length) # number of neighbors for each unit
adj=unlist(shape_nb) # neighbor IDs for each unit
sumNumNeigh=length(unlist(shape_nb))
W = nb2mat(shape_nb,style="B")  # style= B-binary coding. W- row-standardized
W1=nb2mat(shape_nb,style="W")

s=1
C=rep(0, sumNumNeigh)
for (i in 1:N) {
    for (j in 1:N) {
         if (W1[i,j] !=0) {C[s]= W1[i,j]
```

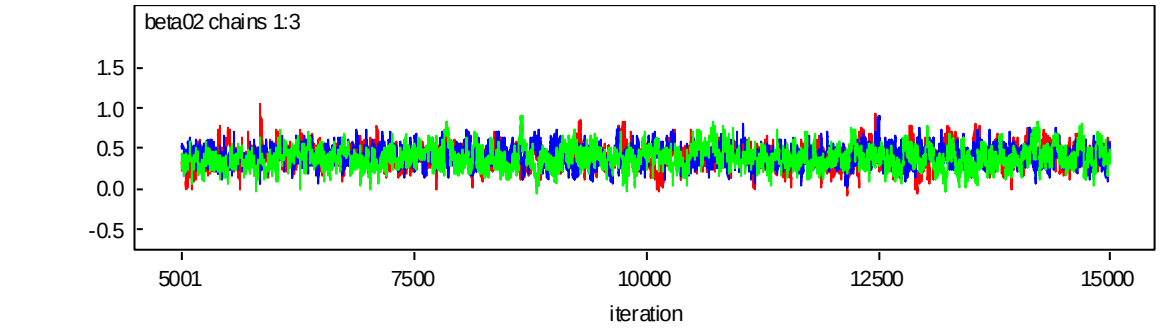
```
        s=s+1}
}
}

firm=read.csv("C:/Users/yw3534/Desktop/New
folder/Graduates_soon/tl_2012_us_county/uscounty_selected.csv", header =
TRUE) #Column1 is nID; #column16 is polygon size

Y10=firm[,19] #Base Est. Birth
Y20=firm[,23] #Retail Est. Birth
Y30=firm[,27] #Service Est. Birth
data=firm[,3:14]
            #1=ALAND(Sq. Meter); 2=pop2010(counts); 3=popden;4=Black
Percent; #5=VacantHousingPercent; 6=AvgHHSize; 7=Med Age; 8=Poverty Family
Percent; 9=Med Person #Income (16 years & older)            #9=LNMileLOC; #
10=Metro INdex; 11=Est. Total; 12=Emply. Total
attach(data)
x1=log(ALAND/4046.8564224) # area land in acres
x2=log(EstTot)
x9=log(EmpTot)
x3=popden #persons per acres
x4=BlckPct #black percentage
x5=VctHousPct #vacant housing units percentage
x6=MedAge # median age
x7=PvtyFmlPct #family below poverty line percentage
x8=MedPrsnInc/10000 #median person income for 16 years and older/10000
x10=MetroIndex

# run a simple Poisson regression to get the initial values of beta's
poi1.data <- data.frame(Y10, x1, x2, x9, x3, x4, x5, x6, x7, x8, x10)
poi1<- glm(formula=Y10~ x1 + x2 + x9 + x3 + x4 +x5 + x6 + x7 + x8 + x10,
family="poisson")
beta1.init <- poi1$coefficients

poi2.data <- data.frame(Y20, x1, x2, x9, x3, x4, x5, x6, x7, x8, x10)
poi2<- glm(formula=Y20~ x1 + x2 + x9 + x3 + x4 +x5 + x6 + x7 + x8 + x10,
family="poisson")
beta2.init <- poi2$coefficients

poi3.data <- data.frame(Y30, x1, x2, x9, x3, x4, x5, x6, x7, x8, x10)
poi3<- glm(formula=Y30~ x1 + x2 + x9 + x3 + x4 +x5 + x6 + x7 + x8 + x10,
family="poisson")
beta3.init <- poi3$coefficients
# estimate beta1 and beta3
data <- list(Y10=Y10,Y20=Y20,Y30=Y30, N=N,adj=adj, sumNumNeigh=sumNumNeigh,
num=num, W=W, x1=x1, x2=x2, x9=x9, x3=x3, x4=x4, x5=x5,x6=x6, x7=x7, x8=x8,
x10=x10, C=C)
inits <- function() {
list(beta01=beta1.init[1], beta11=beta1.init[2], beta21=beta1.init[3],
beta91=beta1.init[4], beta31=beta1.init[5],
beta41=beta1.init[6],beta51=beta1.init[7], beta61=beta1.init[8],
beta71=beta1.init[9], beta81=beta1.init[10], beta101=beta1.init[11],
beta02=beta2.init[1], beta12=beta2.init[2], beta22=beta2.init[3],
beta92=beta2.init[4], beta32=beta2.init[5],
beta42=beta2.init[6],beta52=beta2.init[7], beta62=beta2.init[8],
beta72=beta2.init[9], beta82=beta2.init[10], beta102=beta2.init[11],
beta03=beta3.init[1], beta13=beta3.init[2], beta23=beta3.init[3],
```

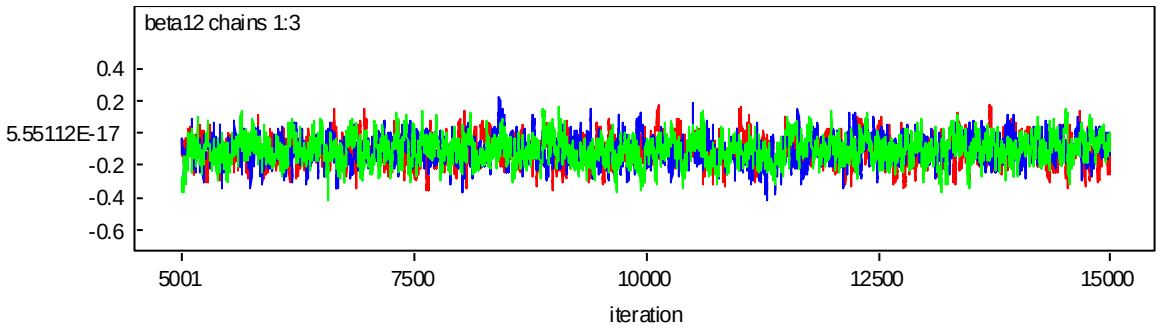
```
beta93=beta3.init[4], beta33=beta3.init[5],
beta43=beta3.init[6],beta53=beta3.init[7], beta63=beta3.init[8],
beta73=beta3.init[9], beta83=beta3.init[10], beta103=beta3.init[11], tau1=1,
tau2=1, tau3=1, tau.v=0.5, eta_012=0.5, eta_112=0.5, eta_013=0.5,
eta_113=0.5, eta_023=0.5, eta_123=0.5, alpha=c(0.7,0.7, 0.7), phi1=rep(0,N),
phi2=rep(0,N), phi3=rep(0,N), V=rep(0, N) )}
parameters <- c("eta_012", "eta_112", "eta_013", "eta_113", "eta_023",
"eta_123",  "beta01", "beta11", "beta21", "beta31","beta41", "beta51",
"beta61", "beta71","beta81", "beta91", "beta101", "beta03", "beta13",
"beta23", "beta33","beta43", "beta53", "beta63", "beta73","beta83", "beta93",
"beta103", "tau2", "tau1","tau3","alpha", "tau.v")
model <- bugs(data, inits, parameters,
model.file="C:/trivariate/firm/trivar3.txt", n.chains=1, n.iter=10000,
n.burnin=2000,n.thin=1, codaPkg=TRUE, DIC=TRUE, debug=TRUE,
bugs.directory="C:/Users/yw3534/Desktop/winbugs14/WinBUGS14")
```

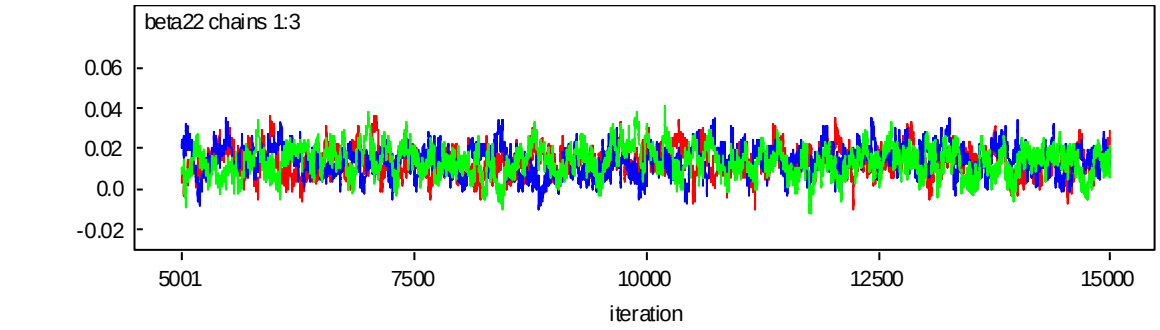
APPENDIX B.
TRACE PLOTS AND DENSITY PLOTS FOR THE BIVARIATE PEDESTRIAN CRASH MODEL

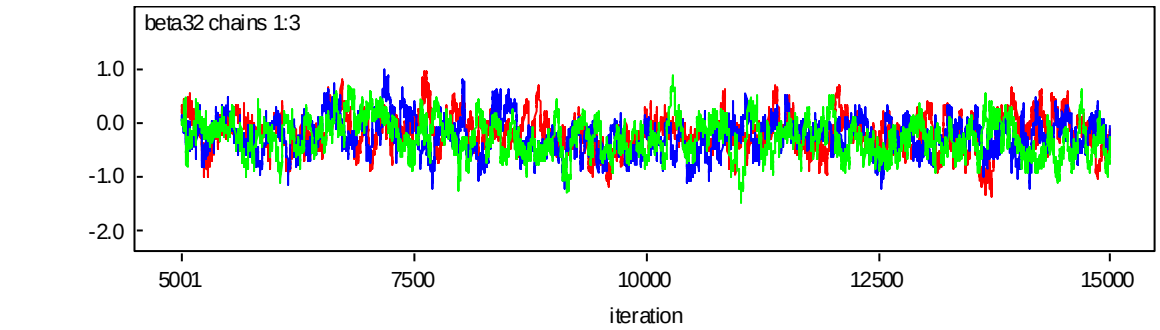
Appendix B contains the plot outputs for the bivariate pedestrian crash model. Trace plots are used to visually examine convergence, as a complement to diagnostic tests (such as Geweke's convergence test). Density plots are also presented here to provide details about the sample space of those unknown parameters.

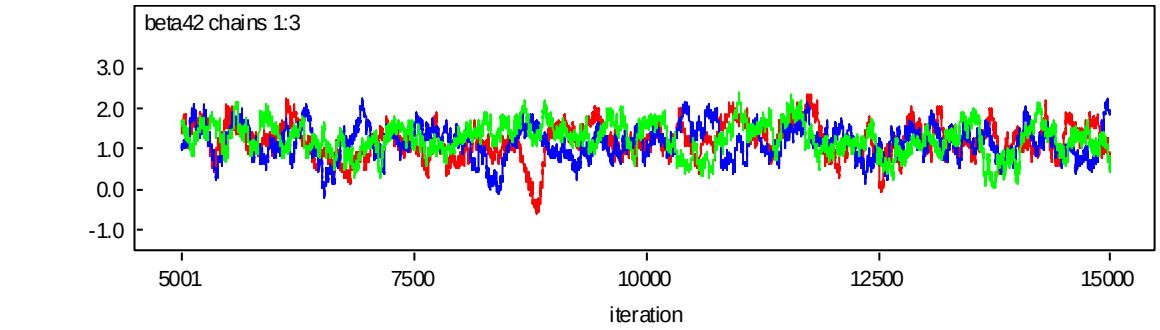


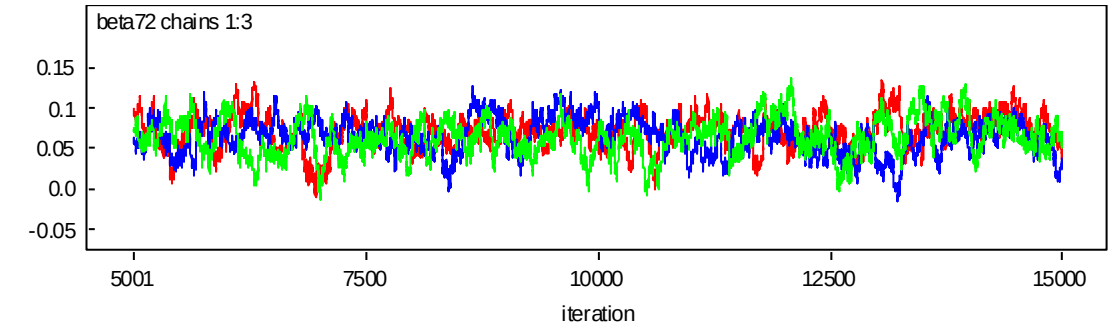

beta02 chains 1:3
1.5
1.0
0.5
0.0
-0.5
5001
7500
10000
12500
15000
iteration

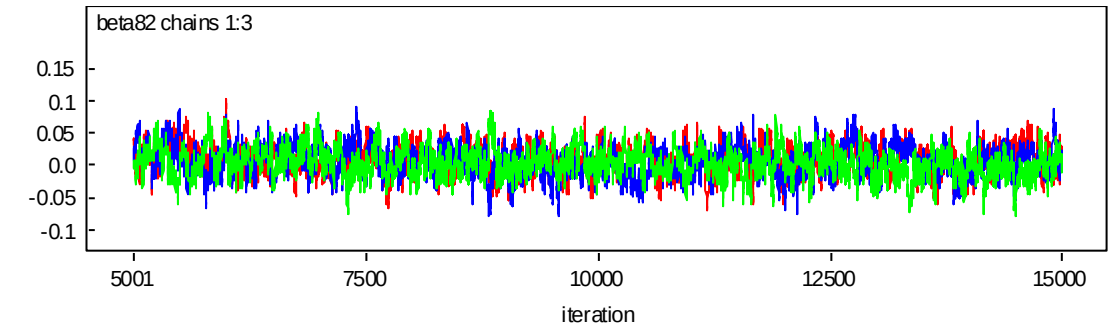

beta12 chains 1:3
0.4
0.2
5.55112E-17
-0.2
-0.4
-0.6
5001
7500
10000
12500
15000
iteration

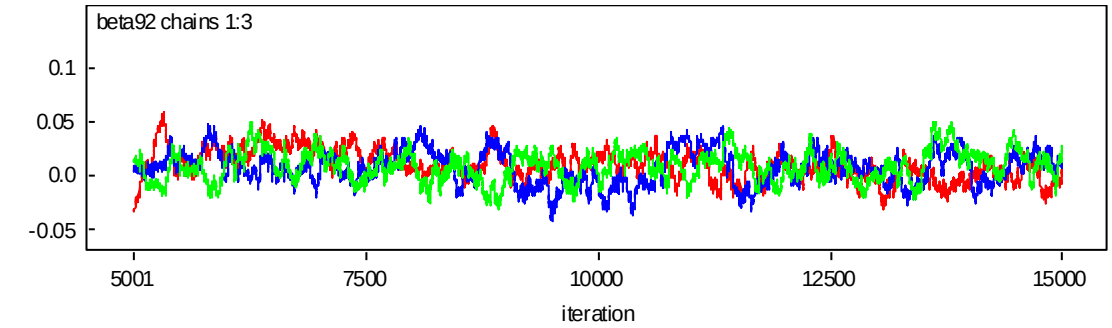

beta22 chains 1:3
0.06
0.04
0.02
0.0
-0.02
5001
7500
10000
12500
15000
iteration

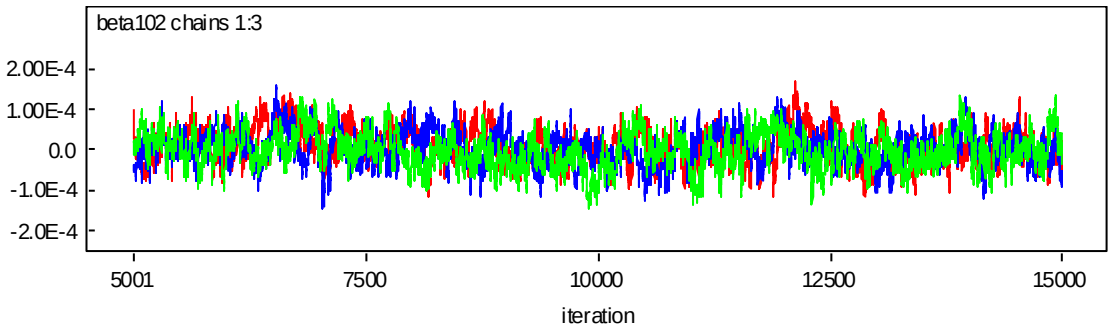

beta32 chains 1:3
1.0
0.0
-1.0
-2.0
5001
7500
10000
12500
15000
iteration

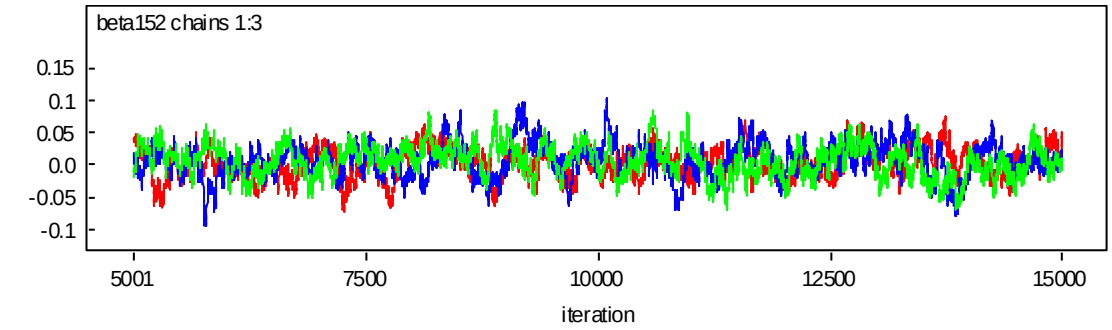

beta42 chains 1:3
3.0
2.0
1.0
0.0
-1.0
5001
7500
10000
12500
15000
iteration

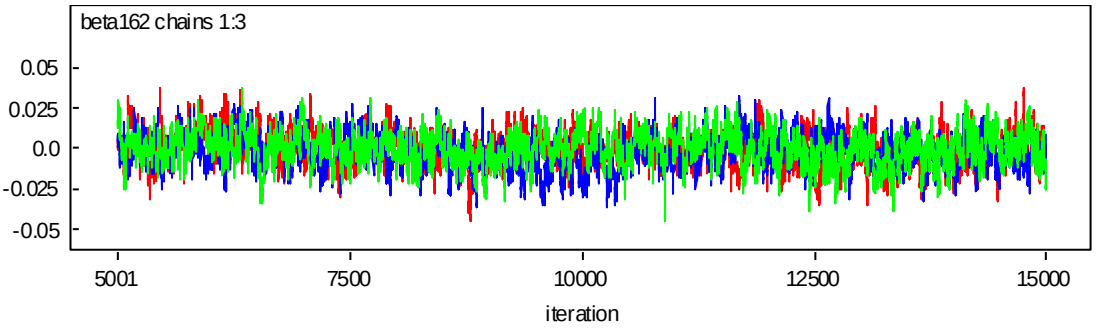

beta72 chains 1:3
0.15
0.1
0.05
0.0
-0.05
5001
7500
10000
12500
15000
iteration

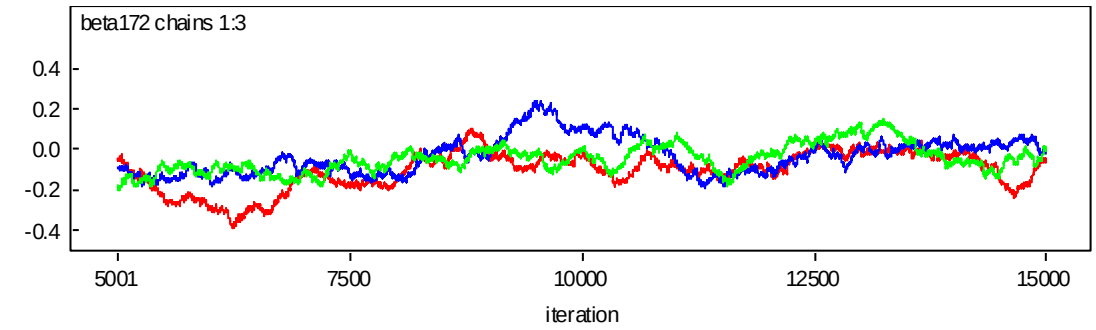

beta82 chains 1:3
0.15
0.1
0.05
0.0
-0.05
-0.1
5001
7500
10000
12500
15000
iteration

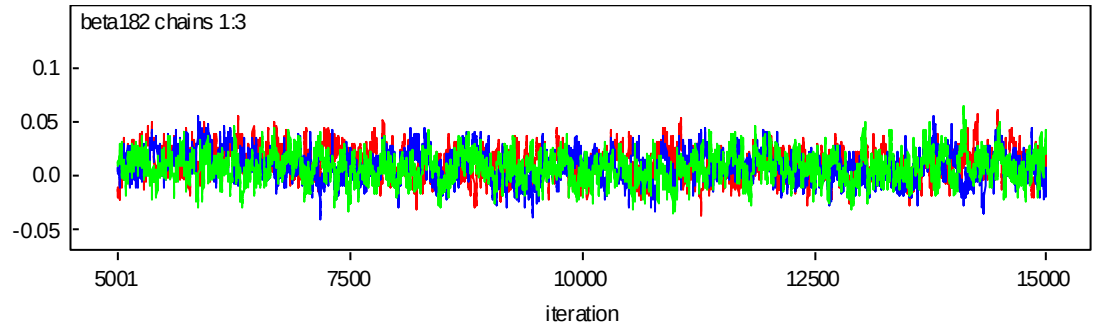

beta92 chains 1:3
0.1
0.05
0.0
-0.05
5001
7500
10000
12500
15000
iteration

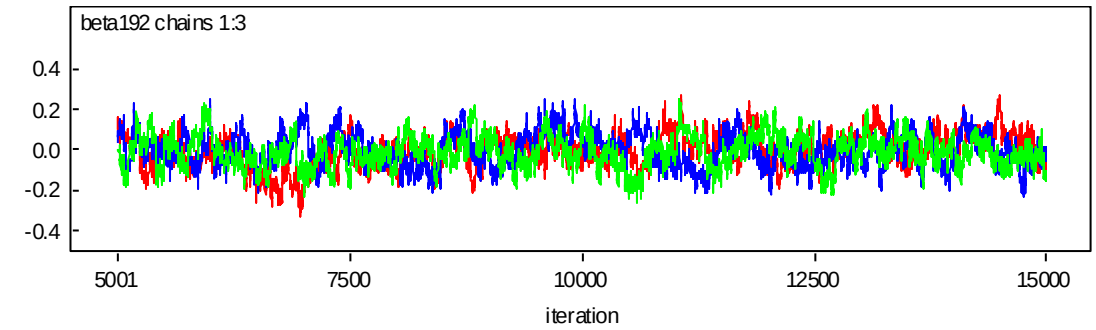

beta102 chains 1:3
2.00E-4
1.00E-4
0.0
-1.0E-4
-2.0E-4
5001
7500
10000
12500
15000
iteration

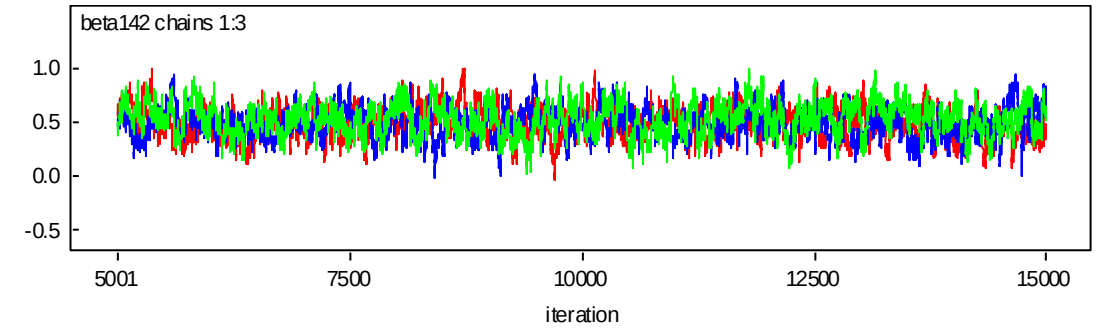

beta152 chains 1:3
0.15
0.1
0.05
0.0
-0.05
-0.1
5001
7500
10000
12500
15000
iteration

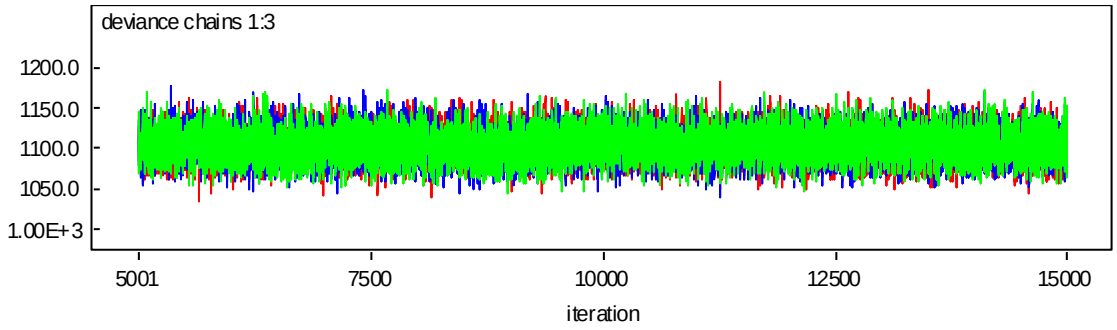

beta162 chains 1:3
0.05
0.025
0.0
-0.025
-0.05
5001
7500
10000
12500
15000
iteration

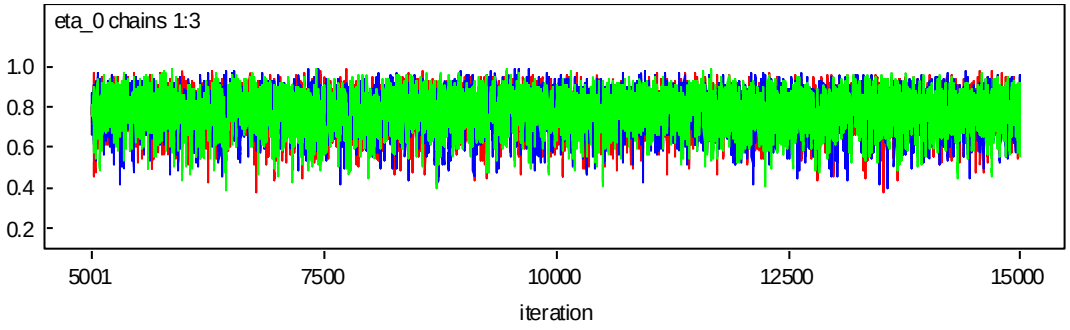

beta172 chains 1:3
0.4
0.2
0.0
-0.2
-0.4
5001
7500
10000
12500
15000
iteration

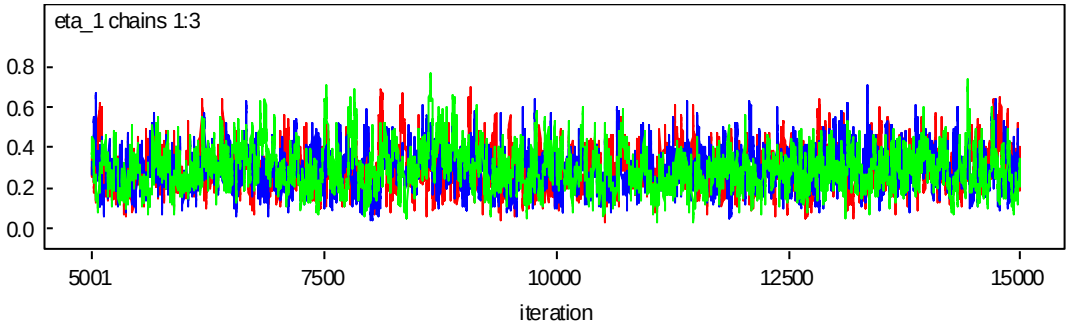

beta182 chains 1:3
0.1
0.05
0.0
-0.05
5001
7500
10000
12500
15000
iteration

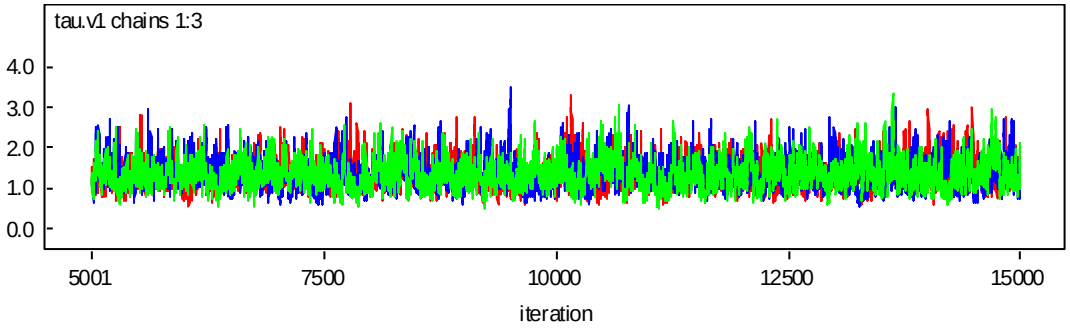

beta192 chains 1:3
0.4
0.2
0.0
-0.2
-0.4
5001
7500
10000
12500
15000
iteration

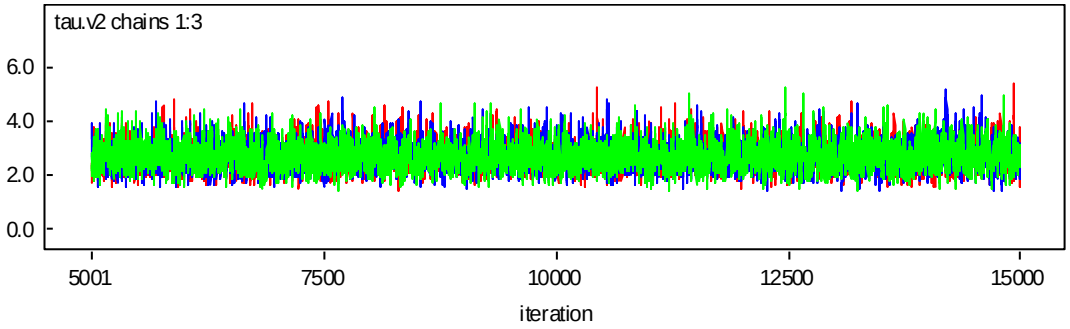

beta142 chains 1:3
1.0
0.5
0.0
-0.5
5001
7500
10000
12500
15000
iteration

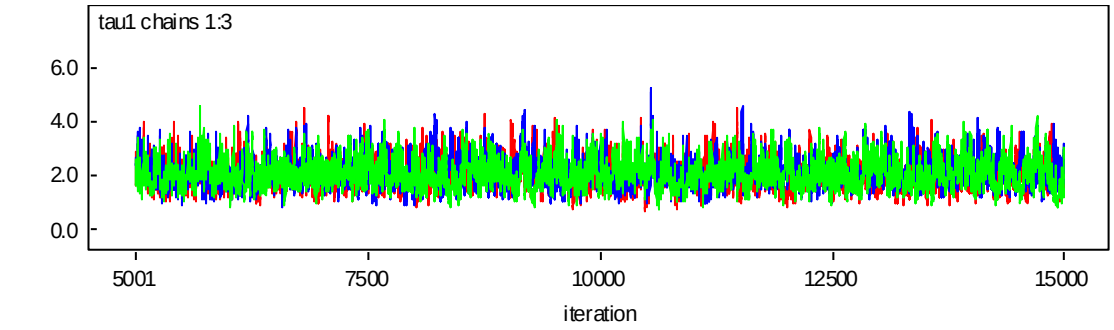

deviance chains 1:3
1200.0
1150.0
1100.0
1050.0
1.00E+3
5001
7500
10000
12500
15000
iteration

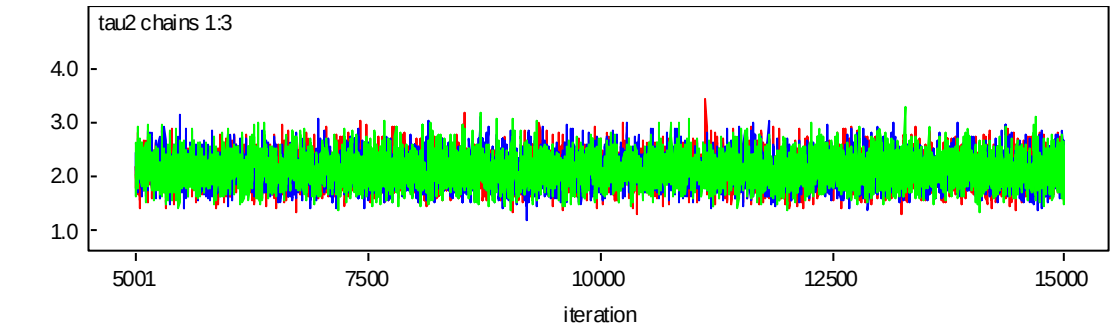

eta_0 chains 1:3
1.0
0.8
0.6
0.4
0.2
5001
7500
10000
12500
15000
iteration

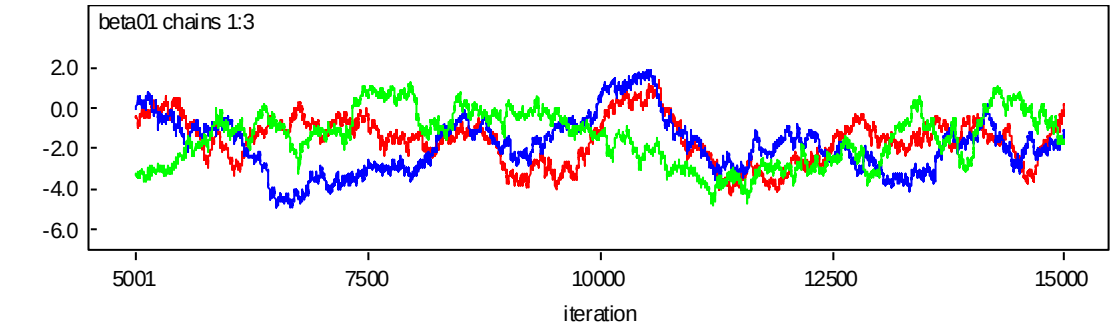

eta_1 chains 1:3
0.8
0.6
0.4
0.2
0.0
5001
7500
10000
12500
15000
iteration

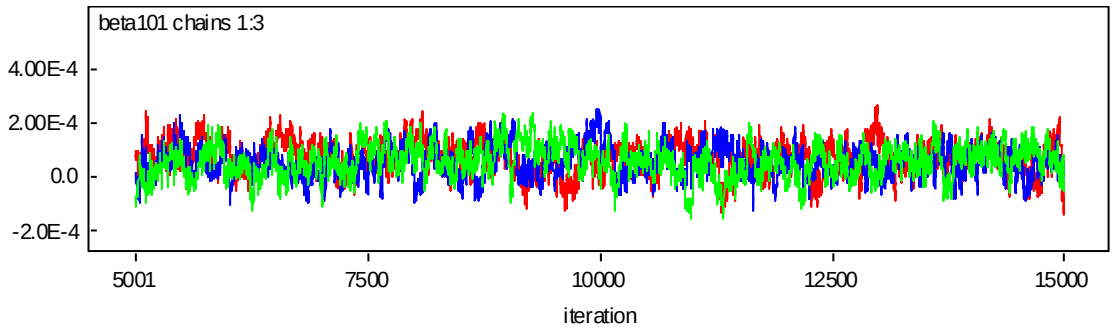

tau.v1 chains 1:3
4.0
3.0
2.0
1.0
0.0
5001
7500
10000
12500
15000
iteration

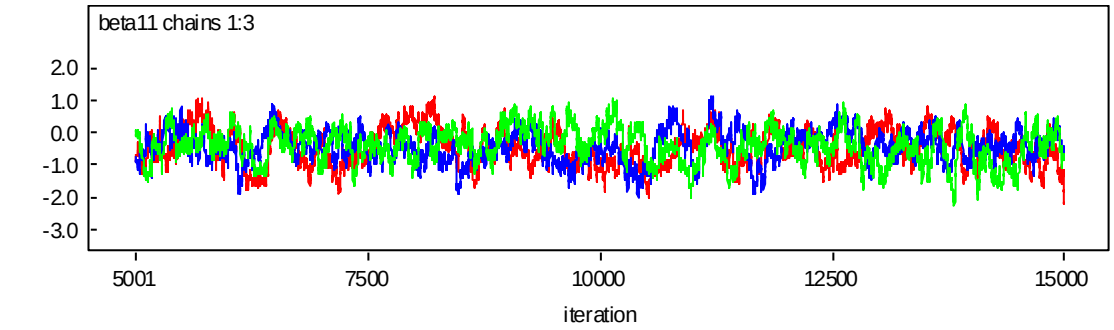

tau.v2 chains 1:3
6.0
4.0
2.0
0.0
5001
7500
10000
12500
15000
iteration

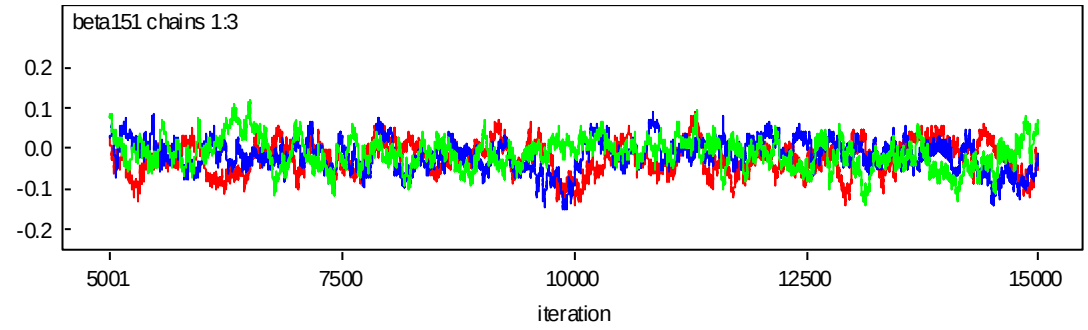

tau1 chains 1:3
6.0
4.0
2.0
0.0
5001
7500
10000
12500
15000
iteration

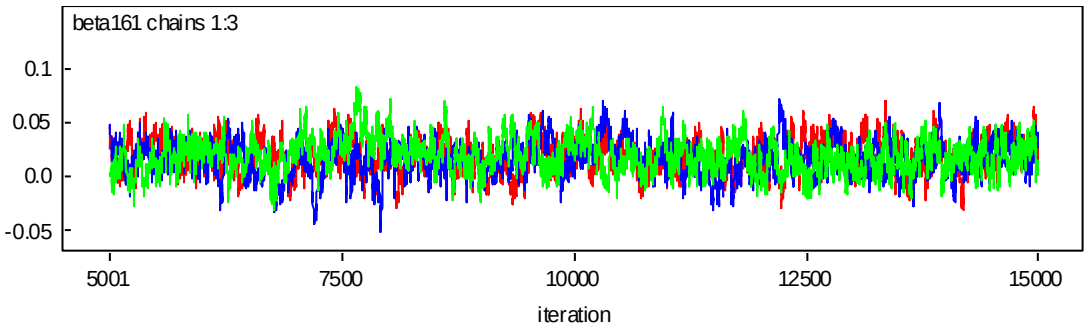

tau2 chains 1:3
4.0
3.0
2.0
1.0
5001
7500
10000
12500
15000
iteration

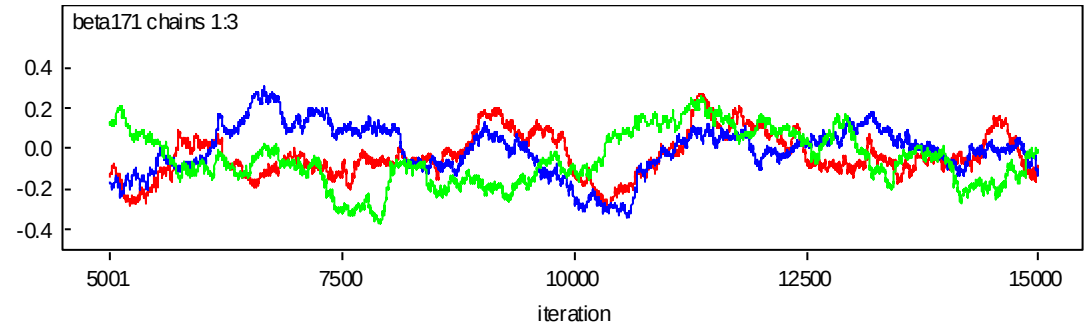

beta01 chains 1:3
2.0
0.0
-2.0
-4.0
-6.0
5001
7500
10000
12500
15000
iteration

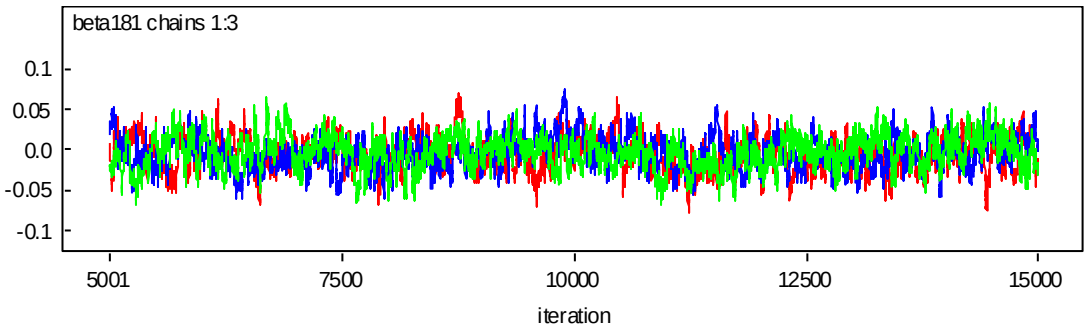

beta101 chains 1:3
4.00E-4
2.00E-4
0.0
-2.0E-4
5001
7500
10000
12500
15000
iteration

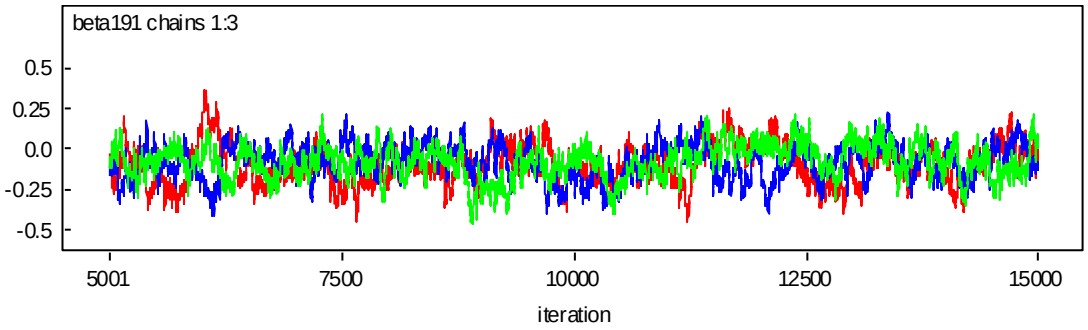

beta11 chains 1:3
2.0
1.0
0.0
-1.0
-2.0
-3.0
5001
7500
10000
12500
15000
iteration

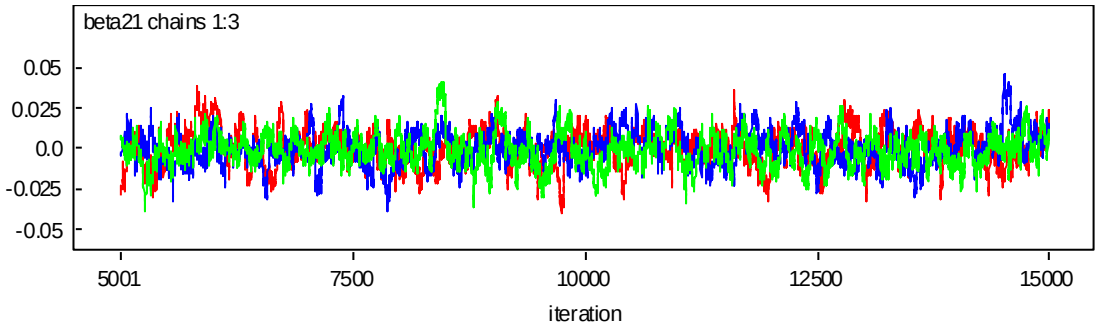

beta151 chains 1:3
0.2
0.1
0.0
-0.1
-0.2
5001
7500
10000
12500
15000
iteration

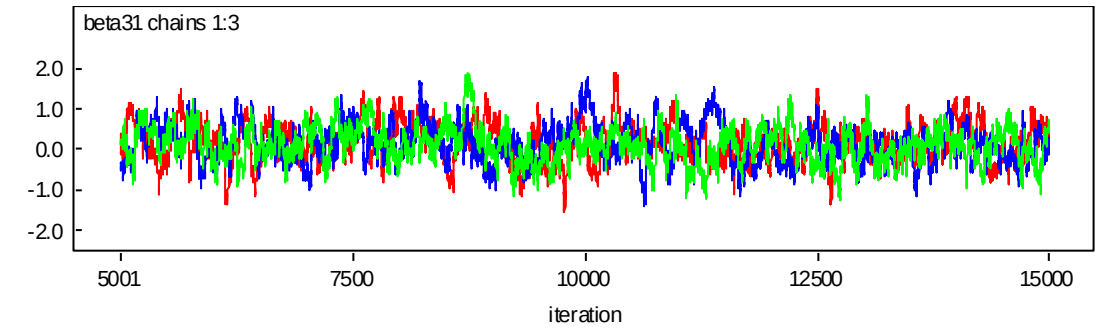

beta161 chains 1:3
0.1
0.05
0.0
-0.05
5001
7500
10000
12500
15000
iteration

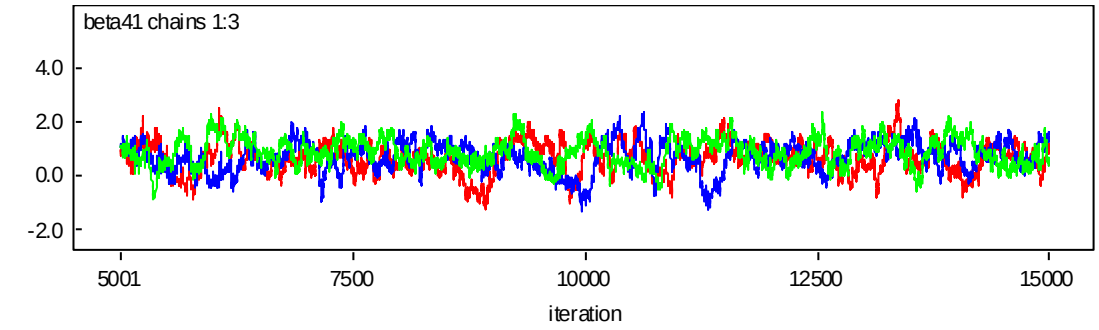

beta171 chains 1:3
0.4
0.2
0.0
-0.2
-0.4
5001
7500
10000
12500
15000
iteration

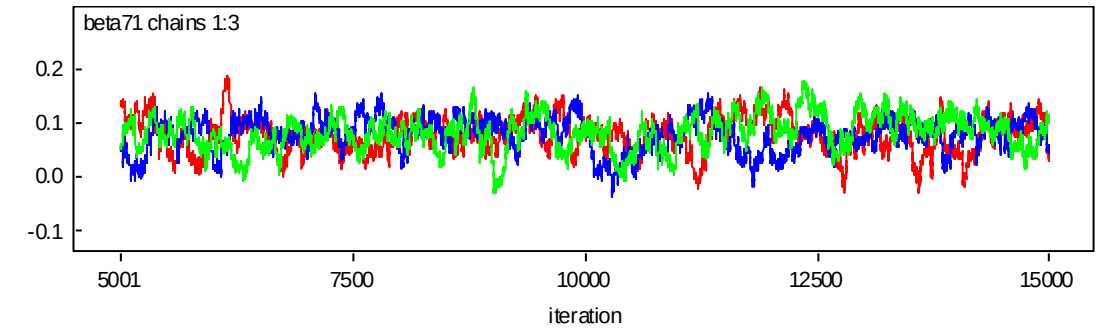

beta181 chains 1:3
0.1
0.05
0.0
-0.05
-0.1
5001
7500
10000
12500
15000
iteration


beta191 chains 1:3
0.5
0.25
0.0
-0.25
-0.5
5001
7500
10000
12500
15000
iteration


beta21 chains 1:3
0.05
0.025
0.0
-0.025
-0.05
5001
7500
10000
12500
15000
iteration


beta31 chains 1:3
2.0
1.0
0.0
-1.0
-2.0
5001
7500
10000
12500
15000
iteration


beta41 chains 1:3
4.0
2.0
0.0
-2.0
5001
7500
10000
12500
15000
iteration


beta71 chains 1:3
0.2
0.1
0.0
-0.1
5001
7500
10000
12500
15000
iteration

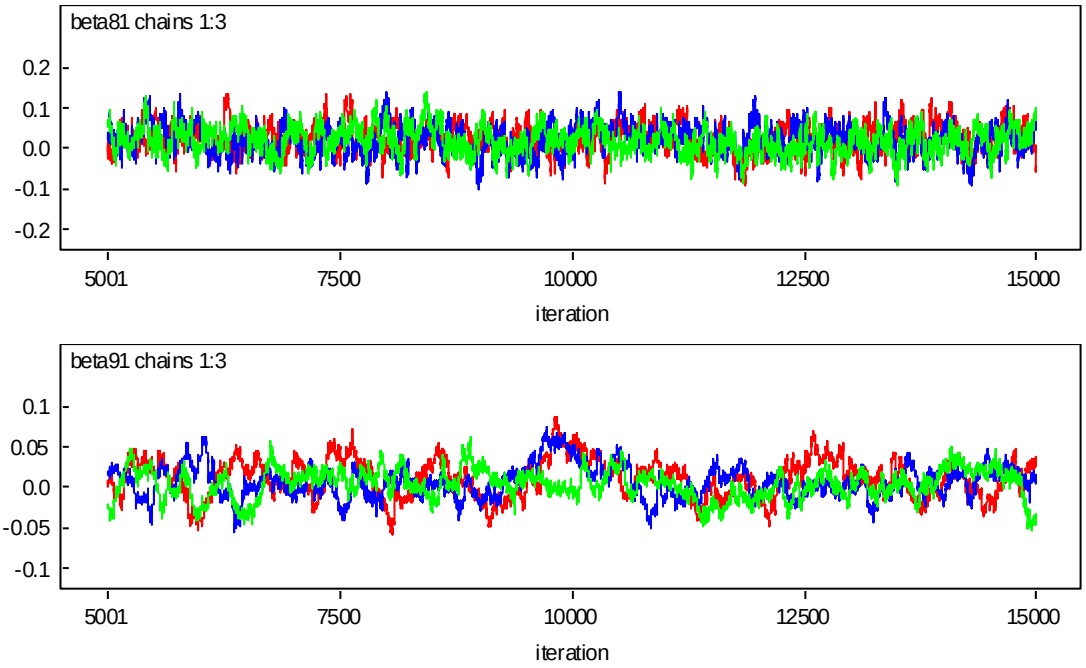
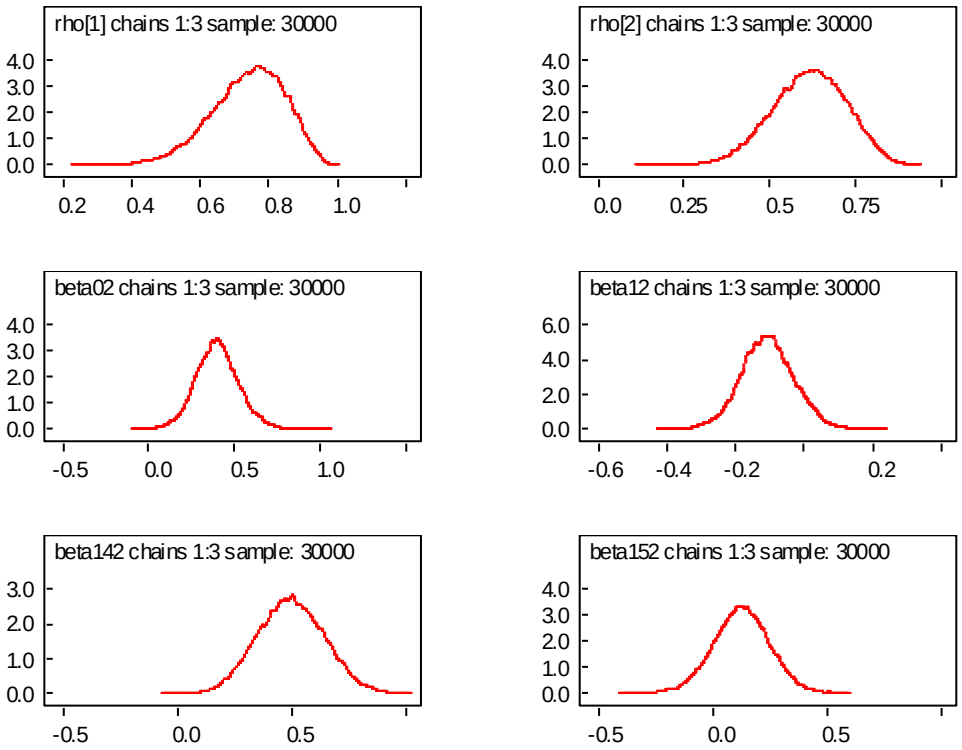
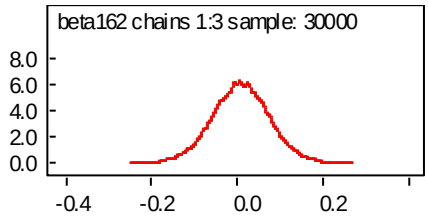
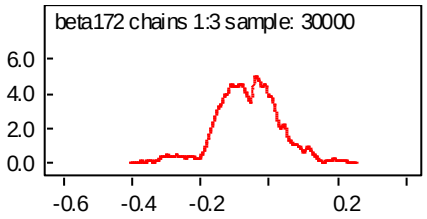
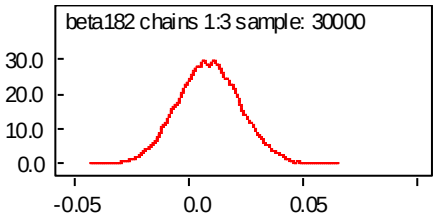


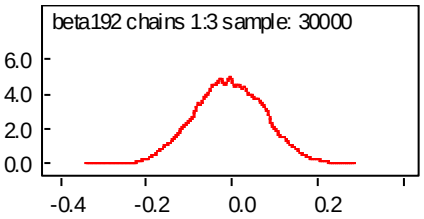
Figure B1. Trace Plots of Parameter Draws for the Pedestrian Crash Model.

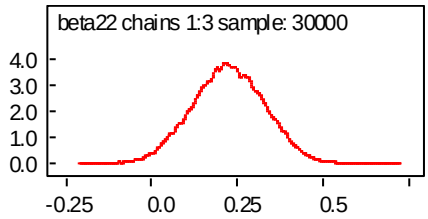


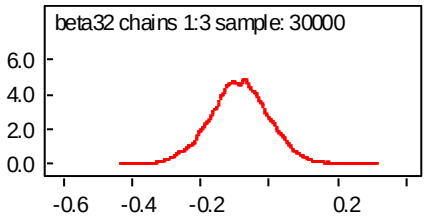

beta162 chains 1:3 sample: 30000
8.0
6.0
4.0
2.0
0.0
-0.4
-0.2
0.0
0.2

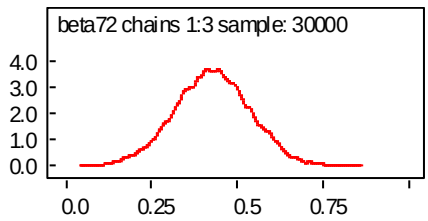

beta172 chains 1:3 sample: 30000
6.0
4.0
2.0
0.0
-0.6
-0.4
-0.2
0.2

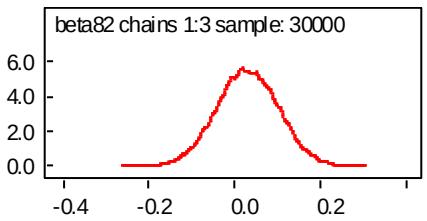

beta182 chains 1:3 sample: 30000
30.0
20.0
10.0
0.0
-0.05
0.0
0.05

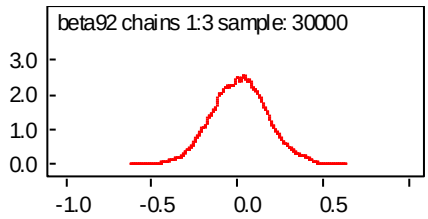

beta192 chains 1:3 sample: 30000
6.0
4.0
2.0
0.0
-0.4
-0.2
0.0
0.2

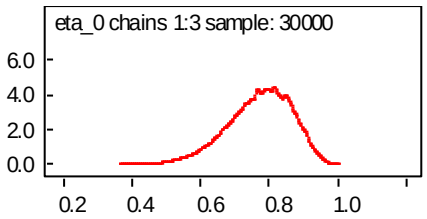

beta22 chains 1:3 sample: 30000
4.0
3.0
2.0
1.0
0.0
-0.25
0.0
0.25
0.5

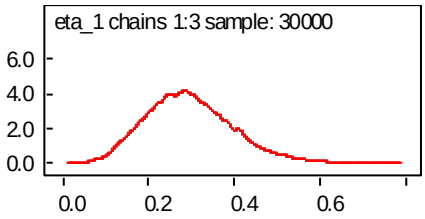

beta32 chains 1:3 sample: 30000
6.0
4.0
2.0
0.0
-0.6
-0.4
-0.2
0.2

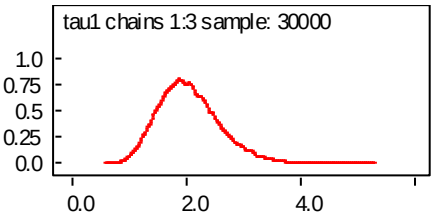

beta72 chains 1:3 sample: 30000
4.0
3.0
2.0
1.0
0.0
0.0
0.25
0.5
0.75

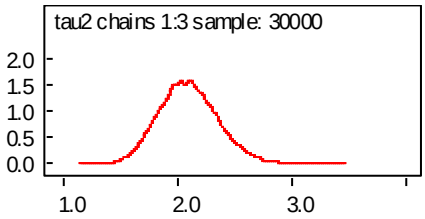

beta82 chains 1:3 sample: 30000
6.0
4.0
2.0
0.0
-0.4
-0.2
0.0
0.2

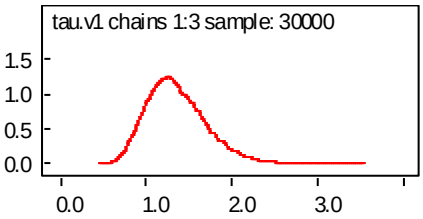

beta92 chains 1:3 sample: 30000
3.0
2.0
1.0
0.0
-1.0
-0.5
0.0
0.5

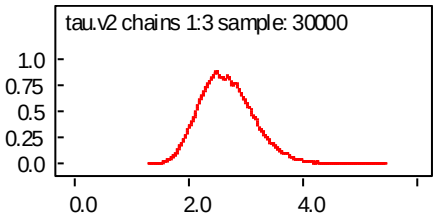

eta_0 chains 1:3 sample: 30000
6.0
4.0
2.0
0.0
0.2
0.4
0.6
0.8
1.0

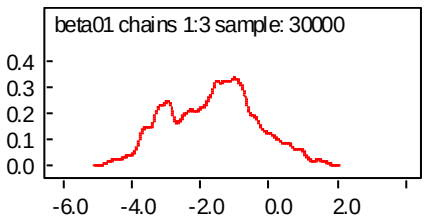

eta_1 chains 1:3 sample: 30000
6.0
4.0
2.0
0.0
0.0
0.2
0.4
0.6

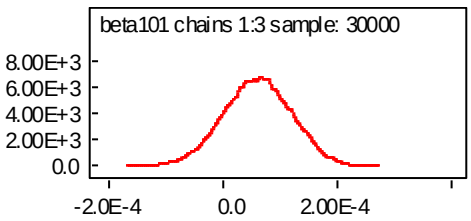

tau1 chains 1:3 sample: 30000
1.0
0.75
0.5
0.25
0.0
0.0
2.0
4.0

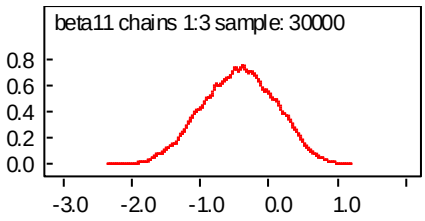

tau2 chains 1:3 sample: 30000
2.0
1.5
1.0
0.5
0.0
1.0
2.0
3.0

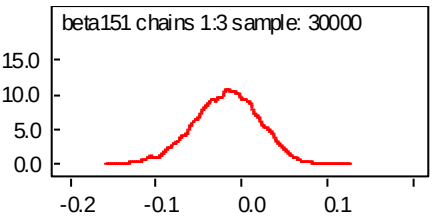

tau.v1 chains 1:3 sample: 30000
1.5
1.0
0.5
0.0
0.0
1.0
2.0
3.0

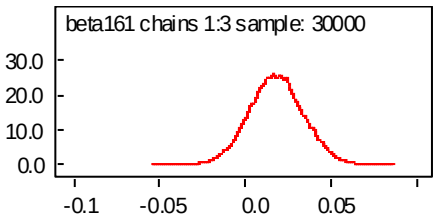

tau.v2 chains 1:3 sample: 30000
1.0
0.75
0.5
0.25
0.0
0.0
2.0
4.0

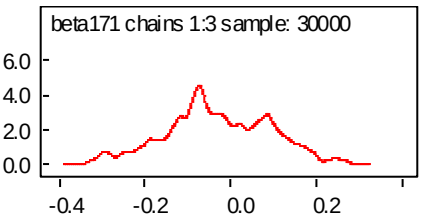

beta01 chains 1:3 sample: 30000
0.4
0.3
0.2
0.1
0.0
-6.0
-4.0
-2.0
0.0
2.0


beta101 chains 1:3 sample: 30000
8.00E+3
6.00E+3
4.00E+3
2.00E+3
0.0
-2.0E-4
0.0
2.00E-4


beta11 chains 1:3 sample: 30000
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0


beta151 chains 1:3 sample: 30000
15.0
10.0
5.0
0.0
-0.2
-0.1
0.0
0.1


beta161 chains 1:3 sample: 30000
30.0
20.0
10.0
0.0
-0.1
-0.05
0.0
0.05


beta171 chains 1:3 sample: 30000
6.0
4.0
2.0
0.0
-0.4
-0.2
0.0
0.2

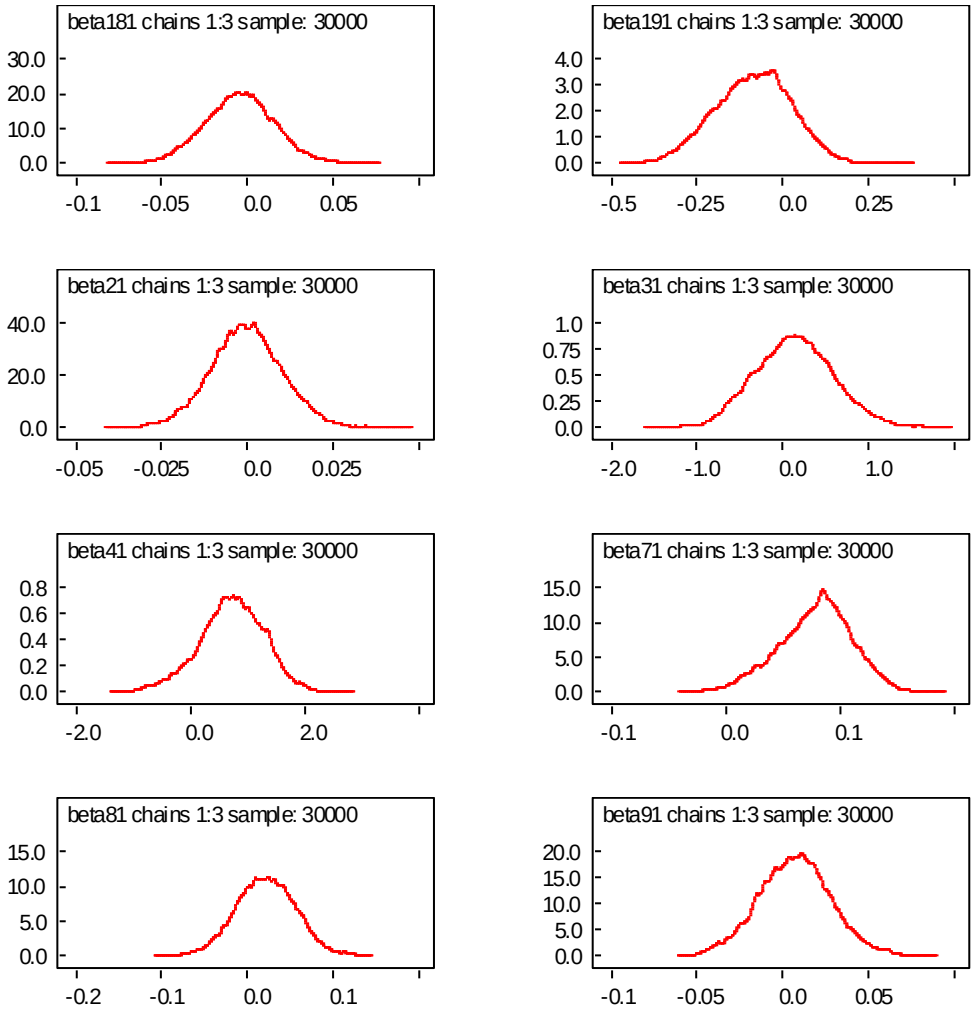
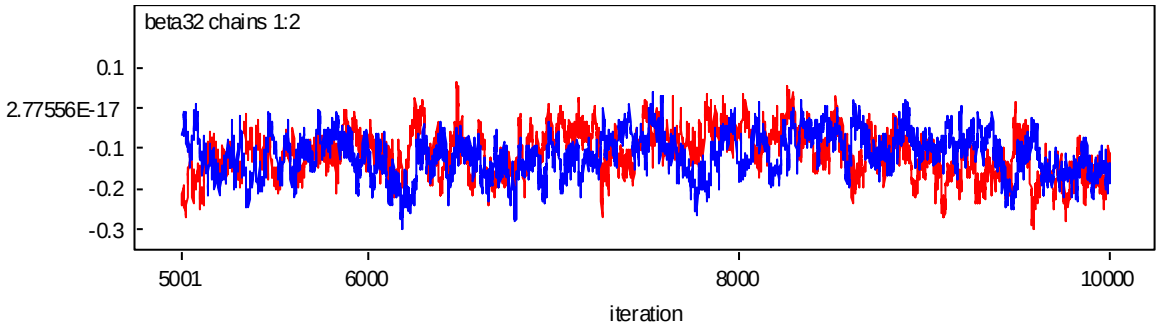
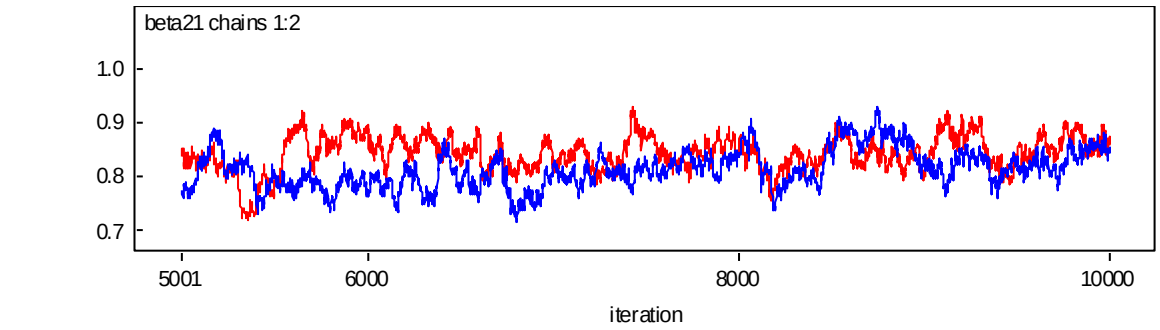
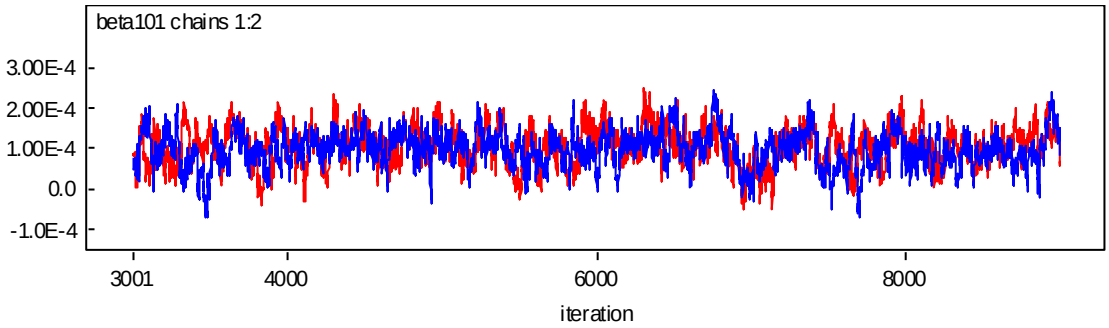
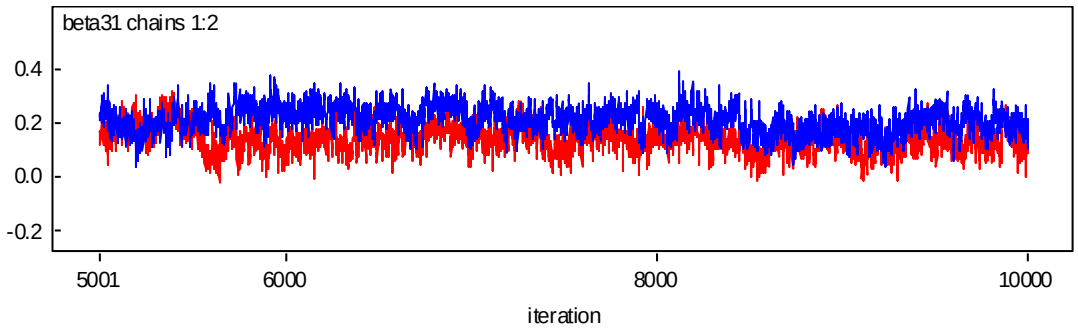


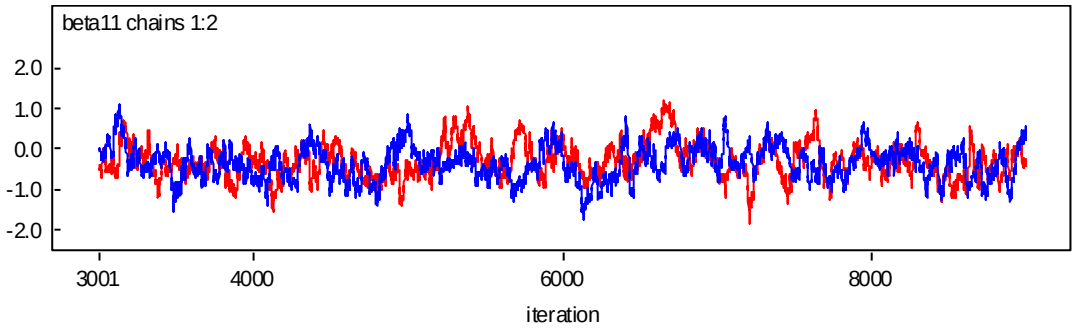
Figure A2. Density Plots of Parameter Estimates for the Pedestrian Crash Model.

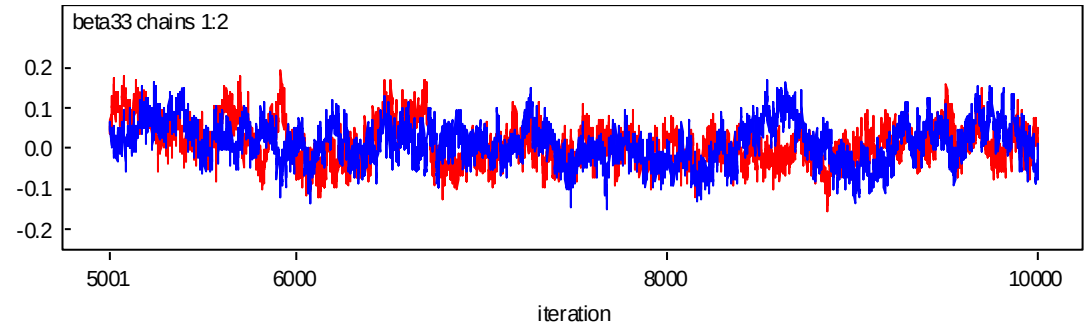

beta32 chains 1:2
0.1
2.77556E-17
-0.1
-0.2
-0.3
5001
6000
8000
10000
iteration

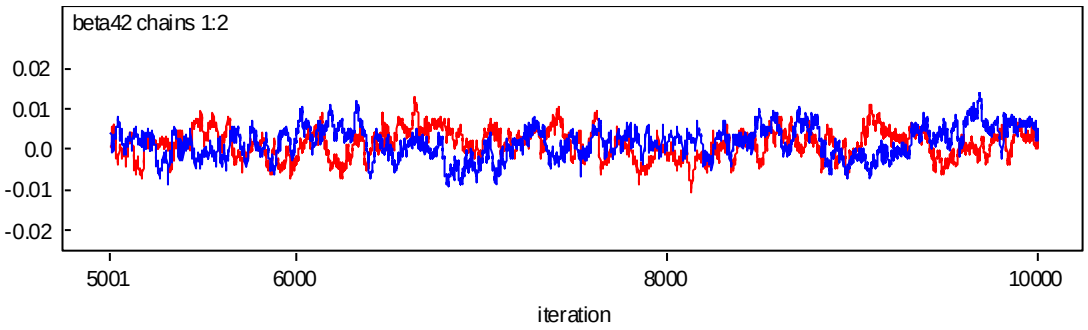

beta21 chains 1:2
1.0
0.9
0.8
0.7
5001
6000
8000
10000
iteration

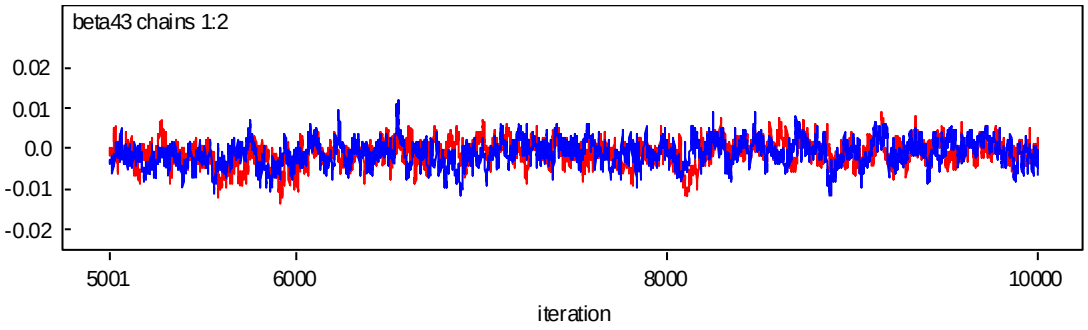

beta101 chains 1:2
3.00E-4
2.00E-4
1.00E-4
0.0
-1.0E-4
3001
4000
6000
8000
iteration

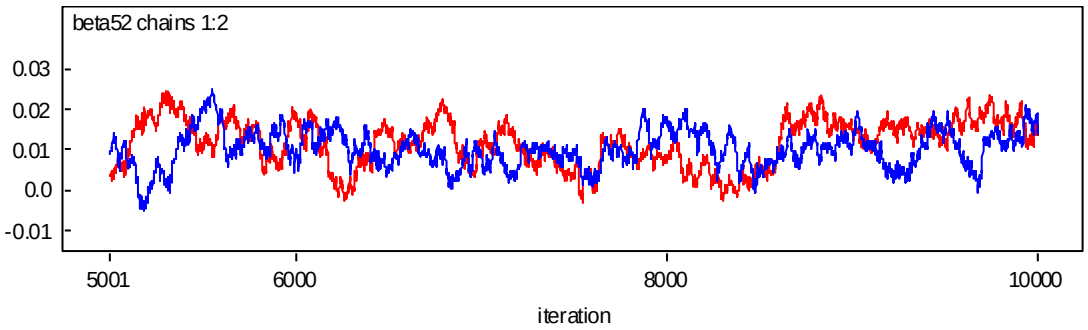

beta31 chains 1:2
0.4
0.2
0.0
-0.2
5001
6000
8000
10000
iteration

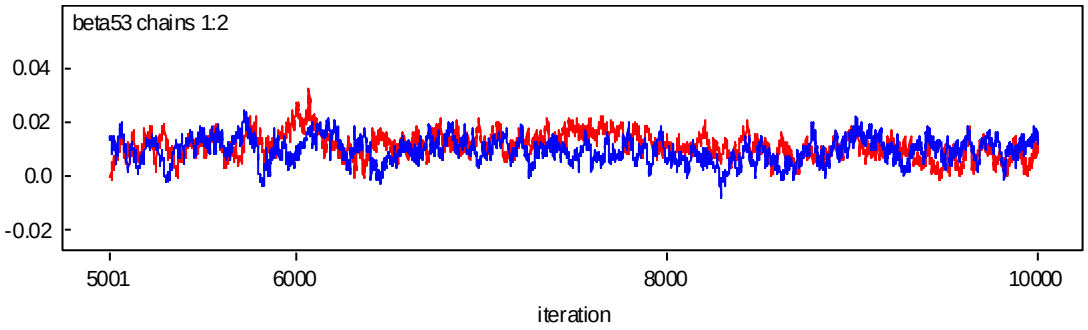

beta11 chains 1:2
2.0
1.0
0.0
-1.0
-2.0
3001
4000
6000
8000
iteration

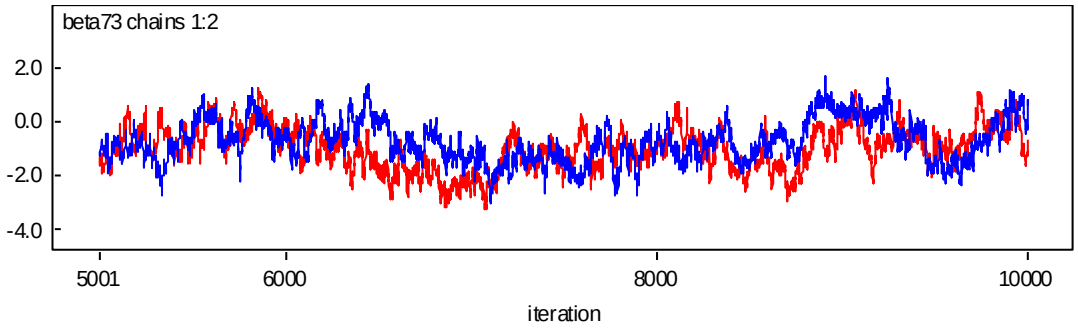

beta33 chains 1:2
0.2
0.1
0.0
-0.1
-0.2
5001
6000
8000
10000
iteration

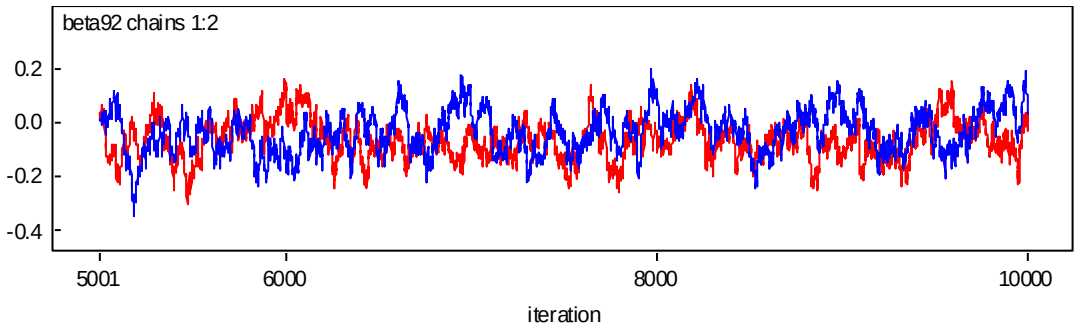

beta42 chains 1:2
0.02
0.01
0.0
-0.01
-0.02
5001
6000
8000
10000
iteration

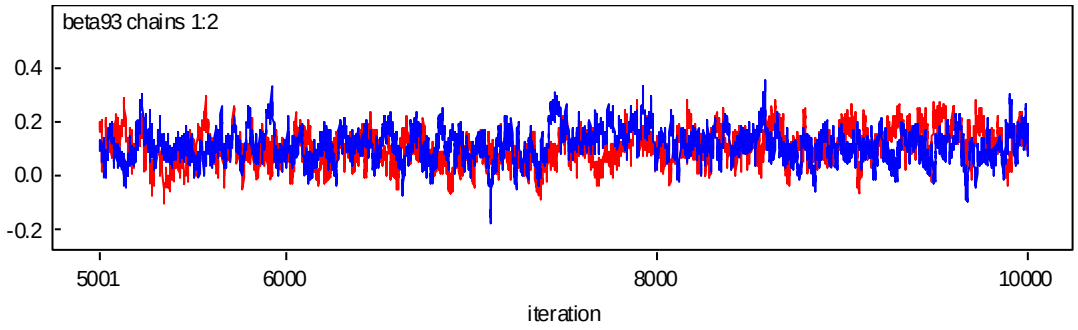

beta43 chains 1:2
0.02
0.01
0.0
-0.01
-0.02
5001
6000
8000
10000
iteration

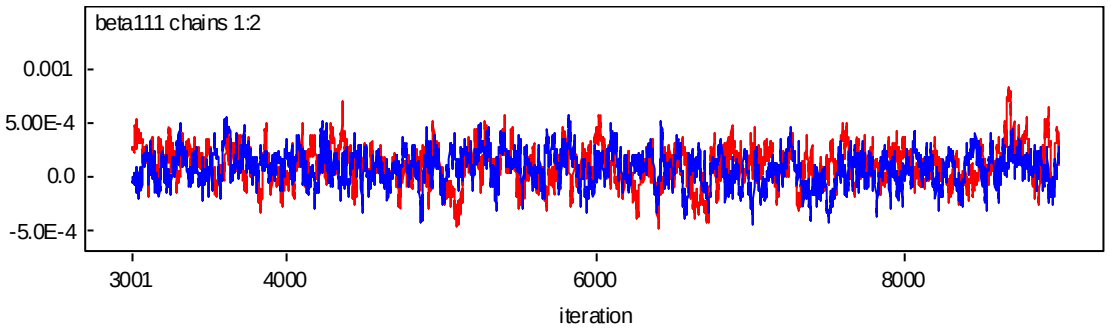

beta52 chains 1:2
0.03
0.02
0.01
0.0
-0.01
5001
6000
8000
10000
iteration

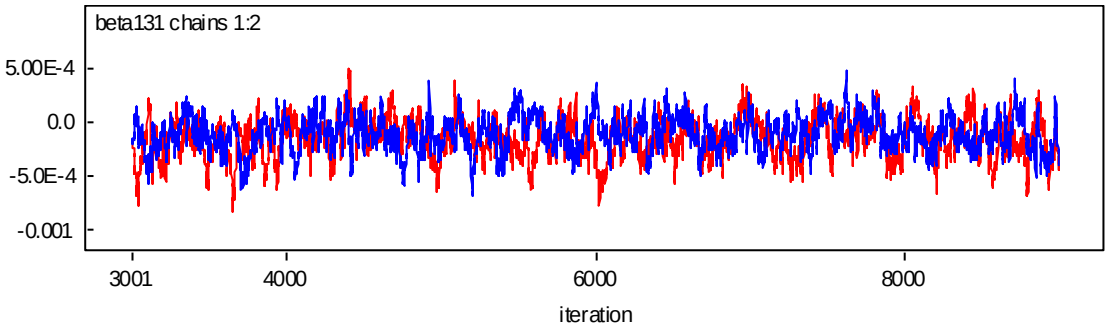

beta53 chains 1:2
0.04
0.02
0.0
-0.02
5001
6000
8000
10000
iteration

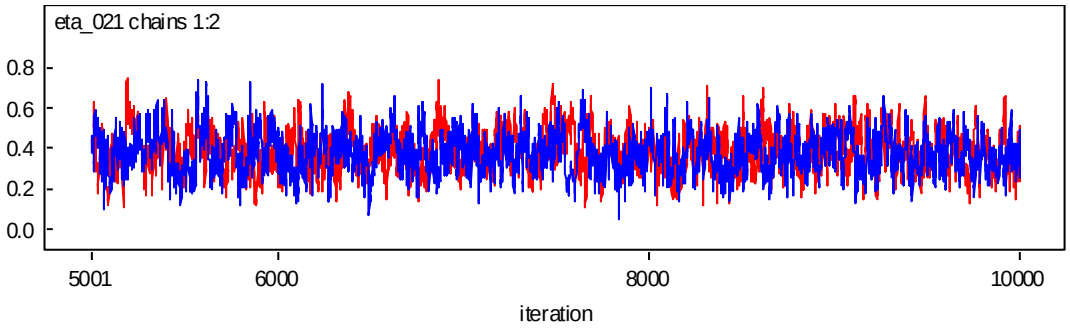

beta73 chains 1:2
2.0
0.0
-2.0
-4.0
5001
6000
8000
10000
iteration

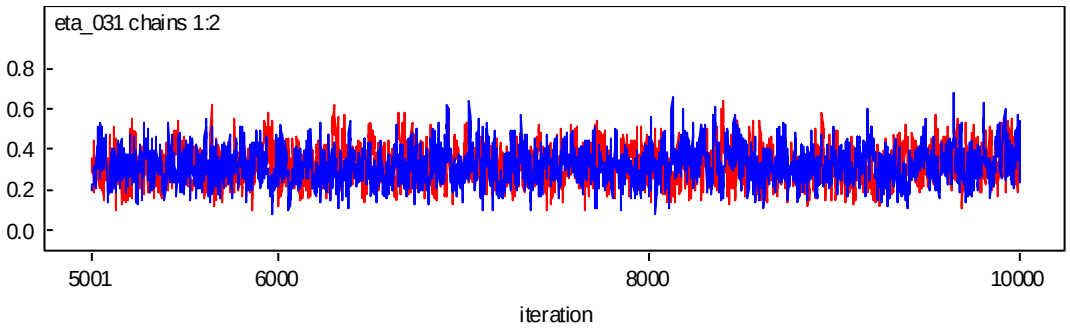

beta92 chains 1:2
0.2
0.0
-0.2
-0.4
5001
6000
8000
10000
iteration

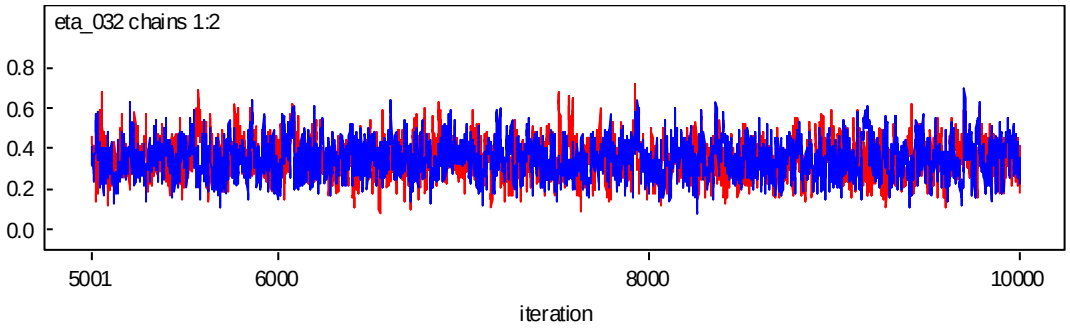

beta93 chains 1:2
0.4
0.2
0.0
-0.2
5001
6000
8000
10000
iteration

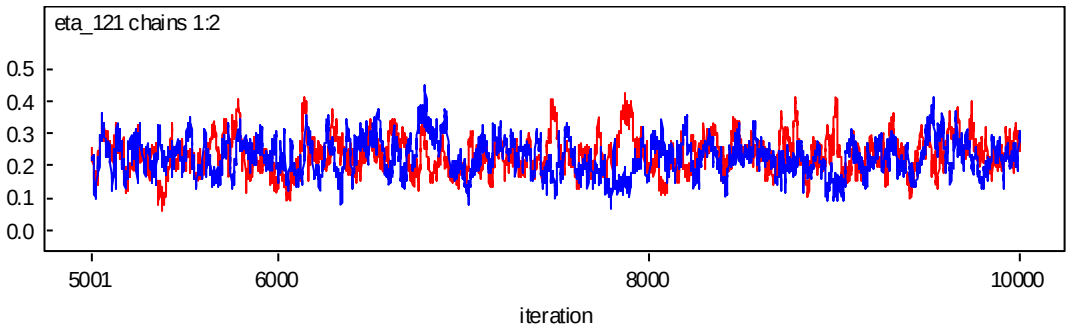

beta111 chains 1:2
0.001
5.00E-4
0.0
-5.0E-4
3001
4000
6000
8000
iteration

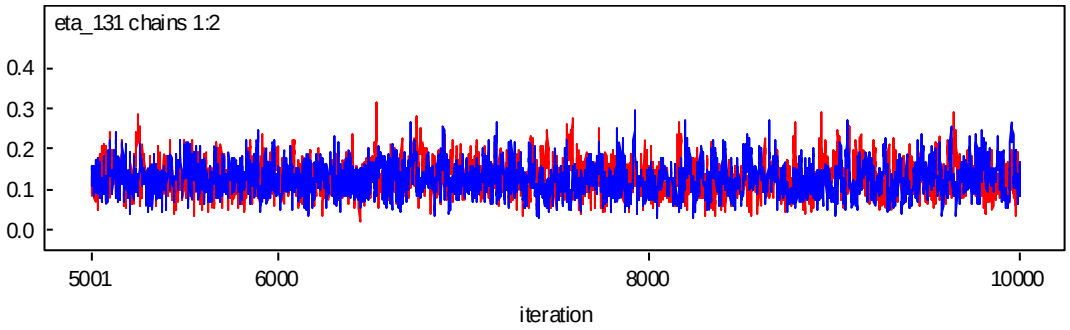

beta131 chains 1:2
5.00E-4
0.0
-5.0E-4
-0.001
3001
4000
6000
8000
iteration

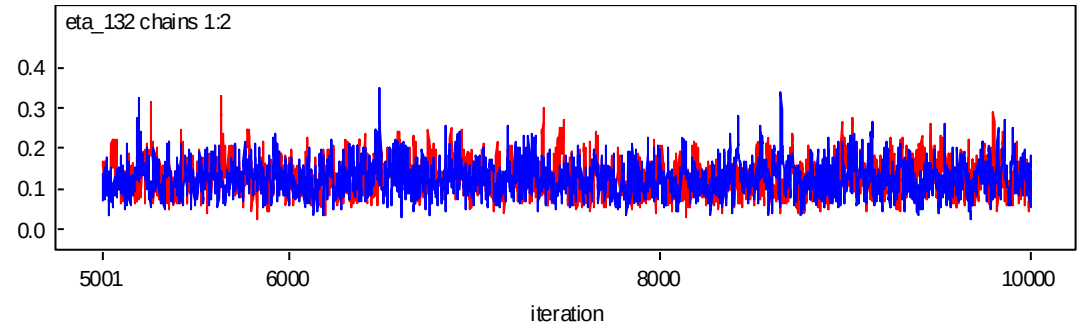

eta_021 chains 1:2
0.8
0.6
0.4
0.2
0.0
5001
6000
8000
10000
iteration

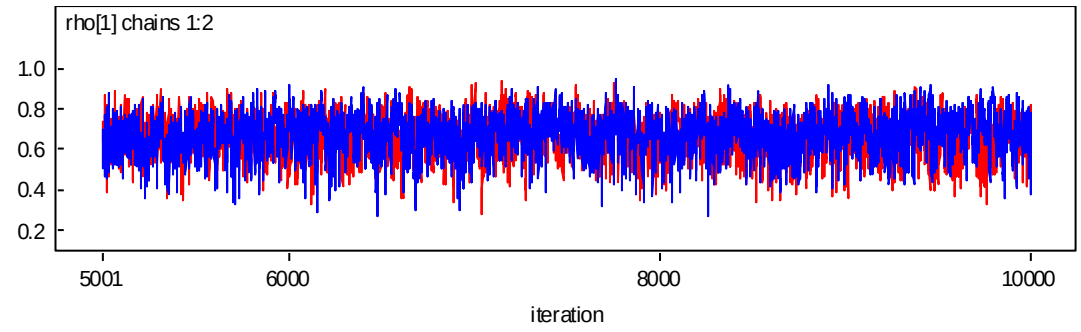

eta_031 chains 1:2
0.8
0.6
0.4
0.2
0.0
5001
6000
8000
10000
iteration

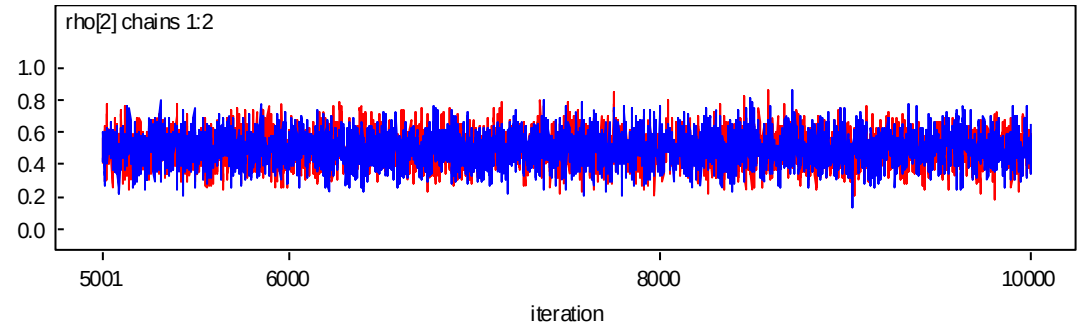

eta_032 chains 1:2
0.8
0.6
0.4
0.2
0.0
5001
6000
8000
10000
iteration

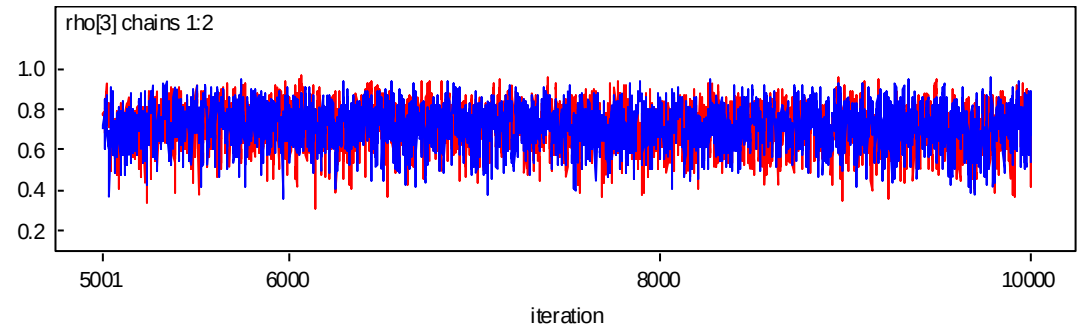

eta_121 chains 1:2
0.5
0.4
0.3
0.2
0.1
0.0
5001
6000
8000
10000
iteration

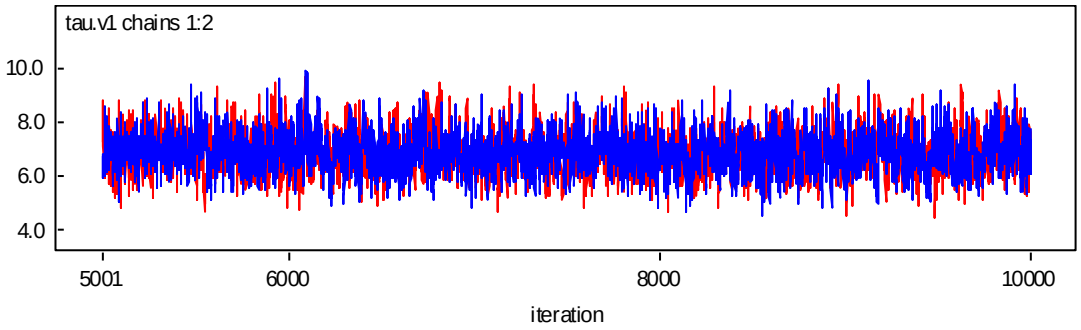

eta_131 chains 1:2
0.4
0.3
0.2
0.1
0.0
5001
6000
8000
10000
iteration

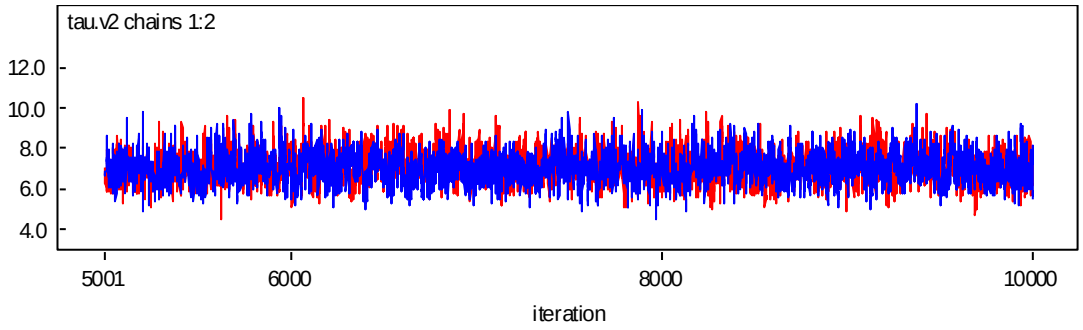

eta_132 chains 1:2
0.4
0.3
0.2
0.1
0.0
5001
6000
8000
10000
iteration

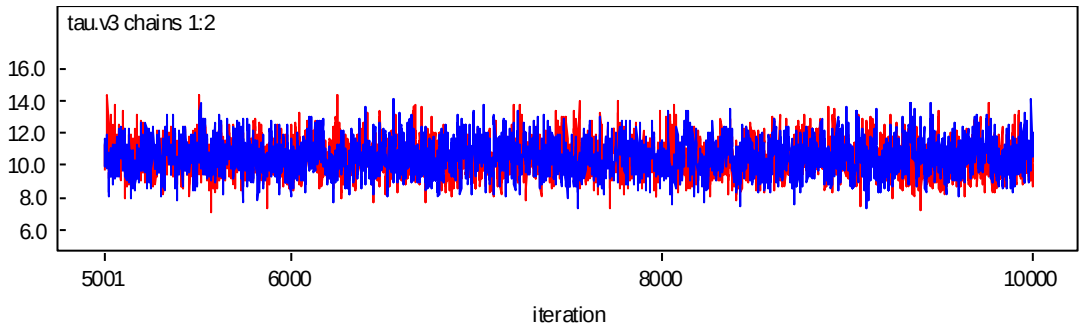

rho[1] chains 1:2
1.0
0.8
0.6
0.4
0.2
5001
6000
8000
10000
iteration

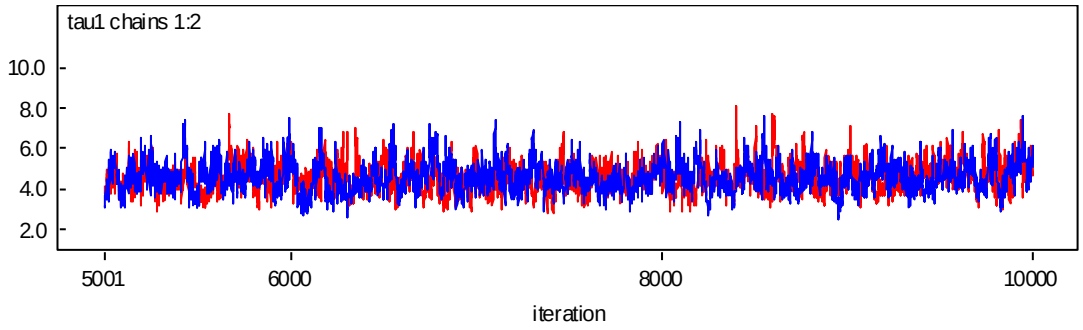

rho[2] chains 1:2
1.0
0.8
0.6
0.4
0.2
0.0
5001
6000
8000
10000
iteration

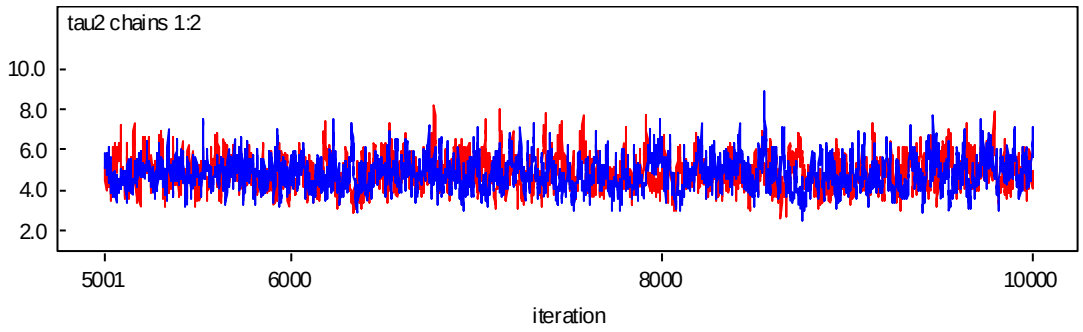

rho[3] chains 1:2
1.0
0.8
0.6
0.4
0.2
5001
6000
8000
10000
iteration


tau.v1 chains 1:2
10.0
8.0
6.0
4.0
5001
6000
8000
10000
iteration


tau.v2 chains 1:2
12.0
10.0
8.0
6.0
4.0
5001
6000
8000
10000
iteration


tau.v3 chains 1:2
16.0
14.0
12.0
10.0
8.0
6.0
5001
6000
8000
10000
iteration


tau1 chains 1:2
10.0
8.0
6.0
4.0
2.0
5001
6000
8000
10000
iteration


tau2 chains 1:2
10.0
8.0
6.0
4.0
2.0
5001
6000
8000
10000
iteration

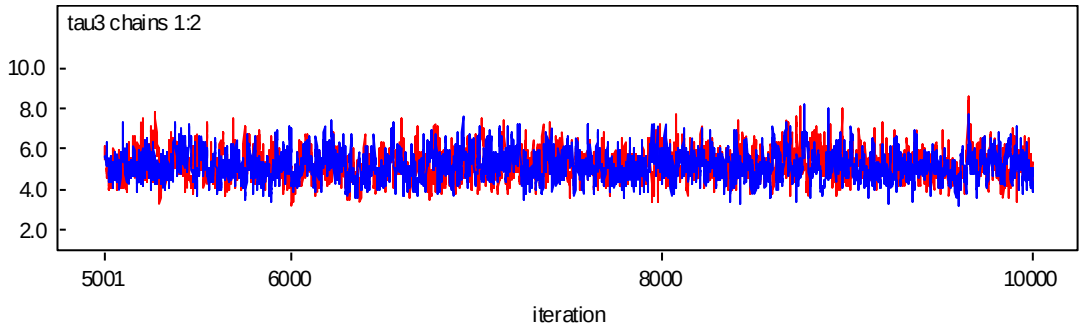
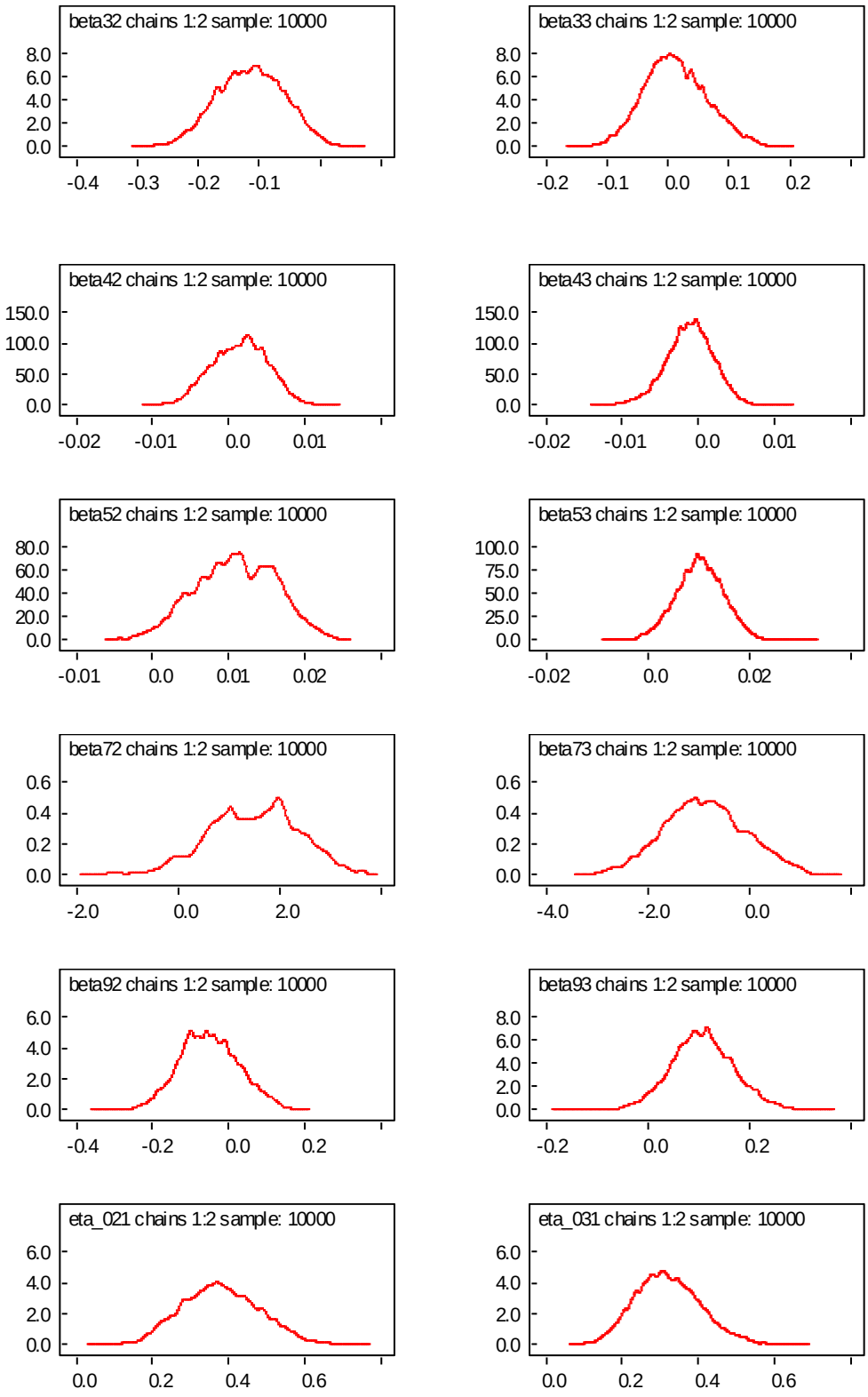
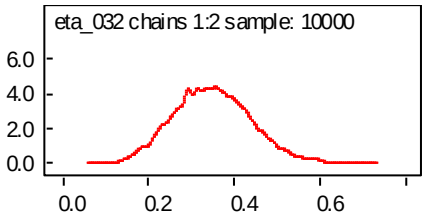
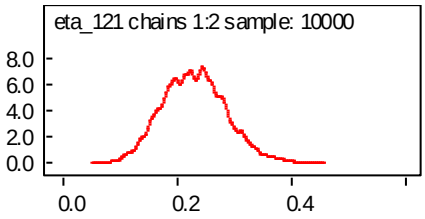
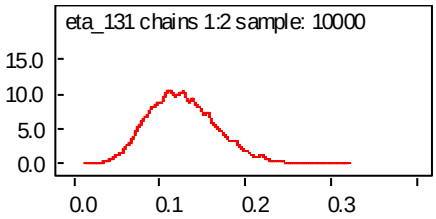

tau3 chains 1:2
10.0
8.0
6.0
4.0
2.0
5001
6000
8000
10000
iteration

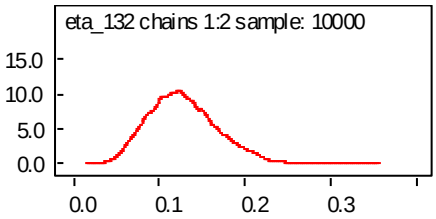
Figure B3. Trace Plots of Parameter Draws of the Firm Birth Model with Three-Level Response.

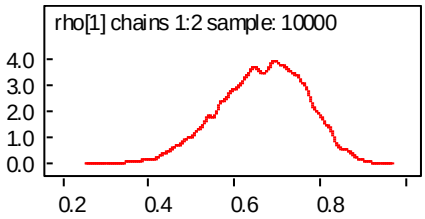


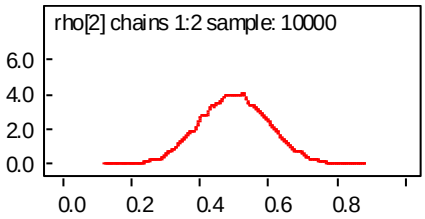

eta_032 chains 1:2 sample: 10000
6.0
4.0
2.0
0.0
0.0
0.2
0.4
0.6

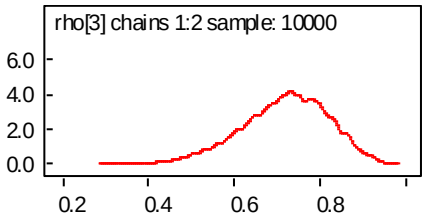

eta_121 chains 1:2 sample: 10000
8.0
6.0
4.0
2.0
0.0
0.0
0.2
0.4

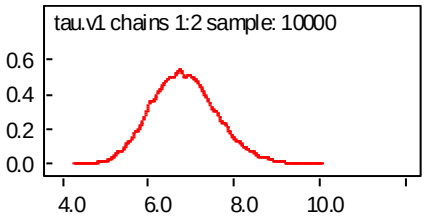

eta_131 chains 1:2 sample: 10000
15.0
10.0
5.0
0.0
0.0
0.1
0.2
0.3

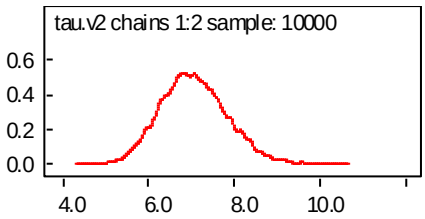

eta_132 chains 1:2 sample: 10000
15.0
10.0
5.0
0.0
0.0
0.1
0.2
0.3

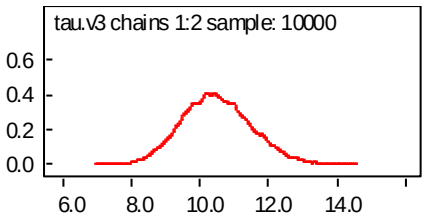

rho[1] chains 1:2 sample: 10000
4.0
3.0
2.0
1.0
0.0
0.2
0.4
0.6
0.8


rho[2] chains 1:2 sample: 10000
6.0
4.0
2.0
0.0
0.0
0.2
0.4
0.6
0.8


rho[3] chains 1:2 sample: 10000
6.0
4.0
2.0
0.0
0.2
0.4
0.6
0.8


tau.v1 chains 1:2 sample: 10000
0.6
0.4
0.2
0.0
4.0
6.0
8.0
10.0


tau.v2 chains 1:2 sample: 10000
0.6
0.4
0.2
0.0
4.0
6.0
8.0
10.0


tau.v3 chains 1:2 sample: 10000
0.6
0.4
0.2
0.0
6.0
8.0
10.0
12.0
14.0

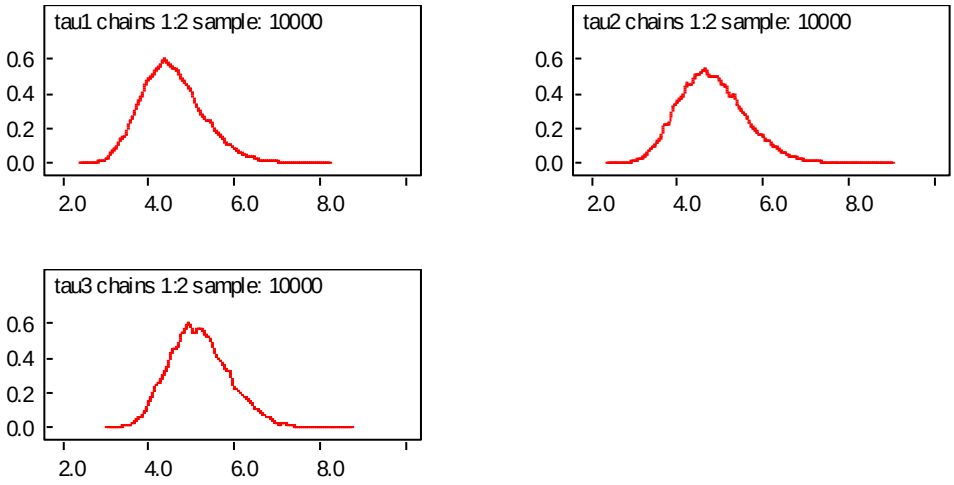


Figure B4. Density Plots of Parameter Estimates for the Firm Birth Model.

References

Abdel-Aty, M., Essam Radwan, E. (2000) Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention* 32: 633–642.

Alsaaty, F. M. (2012) The Cycle of Births and Deaths of U.S. Employer Micro Firms. *Journal of Management and Marketing Research* 11: 1-13.

Anselin, L. (1988) *Spatial Econometrics: Methods and Models.* Kluwer Academic Publisher, Norwell, MA

Besag, J. (1975) Statistical Analysis of Non-Lattice Data. *Statistician* 24 (3): 179–195.

Caliendo, C., Guida, M., Parisi, A. (2007) A crash-prediction model for multilane roads. *Accident Analysis and Prevention* 39: 657- 670.

Carlin, B. and Louis, T. (2009) *Bayesian Methods for Data Analysis. Third Edition*. Chapman & Hall/CRC, Boca Raton, FL.

Carson, J., Mannering, F., Legg, B., Nee, J., Nam, D. (1999) Are incident management programs effective? Findings from Washington State. *Transportation Research Record 1683,* 8-13.

Census (2010) http://www.census.gov/2010census/

Case, B., Clapp, J., Dubin, R., and Rodriguez, M. (2003) Modling Spatial and Temporal House Price Patterns: A Comparison of Four Models. The Journal of Real Estate Finance and Economics 29 (2): 167-191.

Chakir, R., and Parent, O. (2009) Determinants of Land Use Changes: a Spatial Multinomial Probit Approach. *Papers in Regional Science* 88 (2): 327-344.

Clifton, K., Burnier, C., Schneider, R., Huang, S., and Kang, M. (2008) Pedestrian Demand Model for Evaluating Pedestrian Risk Exposure. Technical Report. URL: http://www.kellyjclifton.com/MoPeD/SHAPedestrianModelingpresentation5_19_2008.pdf

Clifton, K., Burnier, C., and Fults, K.K. (2004) Women's involvement in pedestrian-vehicle crashes: influence of personal and environmental factors. Women's Issues in Transportation Conference Proceedings 2 (35): 155-162.

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

Cressie N. A. (1991) *Statistics for Spatial Data.* John Wiley & Sons, Inc. New York.

Davies, R.B., Cenek, P.D., Henderson, R.J. (2005) The effect of skid resistance and texture on crash risk, International Surface Friction Conference, 2005, Christchurch, New Zealand.

Ewing, R. (2006) Fatal and Non-fatal Injuries. Understanding the Relationship Between Public Health and the Built Environment: A Report Prepared for the LEED-ND Core Committee. URL: http://www.activeliving.org/files/LEED%20ND%20report.pdf

Elhorst, P. (2009) Spatial Panel Data Models. In Fischer M., and Getis A. (eds.) *Handbook of Applied Spatial Analysis*: 377-407. Springer, Berlin.

FHWA (2007) Safety at unsignalized intersections. Federal Highway Adminsitration, US Department of Transportation, Washington DC. URL: http://safety.fhwa.dot.gov/intersection/unsignalized/presentations/unsig_pps_041409/long.cfm

Gelfand, A.E., and Vounatsou, P. (2003) Proper multivariate conditional autoregressive models for spatial data analysis. *Biostat* 4 (1): 11-15.

Gelman, A. and Rubin, D. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science* 7(4): 457-472.

Geweke, J. (1992) Evaluating the accuracy of sampling based approaches to the calculation of posterior moments. *Bayesian Statistics* 4: 169-193.

Gibbs, J. P. and Poston, D. L., Jr. (1975) The Division of Labor: Conceptualization and Related Measures. *Social Forces* 53 (3): 468-476.

Goodchild, M. F. and Haining, R. P. (2004) GIS and Spatial Data Analysis: Converging Perspectives. *Papers in Regional Science* 83: 363-385.

Griffith, D. (2000) A Linear Regression Solution to the Spatial Autocorrelation Problem. *Journal of Geographical Systems* 2: 141-156.

Gschlöÿl, S. and Czado, C. (2008) Does a Gibbs sampler approach to spatial Poisson regression models outperform a single site MH sampler? *Computational Statistics and Data Analysis* 52: 4184-4202.

Hanna, M. and Freeman, J. (1977) The Population Ecology of Organizations. American Journal of Sociology 82(5): 929-964.

Jin, X., Carlin, B.P., and Banerjee, S. (2005) Generalized Hierarchical Multivariate CAR Models for Areal Data, *Biometrics* 61: 950--961.

Khan, G., Qin, X., and Noyce, D. (2008) Spatial Analysis of Weather Crash Patterns in Wisconsin. *Journal of Transportation Engineering* 134 (5): 191-202.

Lee, J. and Mannering, F. (2002) Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis. *Accident Analysis and Prevention* 34 (2): 149-161.

Leyden, K.M. (2003) Social capital and the built environment: the importance of walkable neighborhoods. *American Journal of Public Health* 93 (9): 1546–1551.

LeSage, J., and Pace, K. (2009) *Introduction to Spatial Econometrics*. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL.

Levine, N., Kim, K., and Nitz, L. (1995a) Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. *Accident Analysis and Prevention* 27 (5): 663-674.

Levine, N., Kim, K., and Nitz, L. (1995b) Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators. *Accident Analysis and Prevention* 27 (5): 675-685.

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

Ma, J., Kockelman, K.M., and Damien, P. (2008) A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis and Prevention* 40 (3): 964–975.

Mardia, K. V. (1988) Multi-dimensional Multivariate Gaussian Markov Random Field with Application to Image Processing. *Journal of Multivariate Analysis* 24: 265-284.

Miaou, S-P., Song, J., and Mallick, B. (2003) Roadway traffic crash mapping: a space-time modeling approach, *Journal of Transportation & Statistics* 6 (1): 33-58.

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

Miranda-Moreno, L., Morency, P., and El-Geneidy, A. (2011) The link between built environment, pedestrian activity and pedestrian-vehicle collision occurrence at signalized intersections. *Accident Analysis and Prevention* 43(5): 1624–1634.

Morency, P., and Cloutier, M.S. (2006) From targeted "black spots" to area-wide pedestrian safety. *Injury Prevention* 12: 360–364.

Naderan, A., and Shahi, J. (2009) Aggregate crash prediction models: Introducing crash generation concept. *Accident Analysis and Prevention* 42 (1): 339-346.

Neider, M. B., McCarley, J.S., Crowell, J.A., Kaczmarski, H., and Kramer, A.F. (2010) Pedestrians, vehicles, and cell phones. *Accident Analysis & Prevention* 42 (1): 589-594

National Population Projections 2009 (2008) U.S. Census Bureau. Washington, D.C. URL: http://www.census.gov/population/projections/data/national/2009.html.

NHSTA (2011) Traffic Safety Facts 2009 Data, National Highway Traffic Safety Administration, US DOT. URL: http://www-nrd.nhtsa.dot.gov/Pubs/811394.pdf.

Park, E.S., and Lord, D. (2007) Multivariate Poisson-lognormal models for jointly modeling crsh frequency by severity. *Transportation Research Record* 2019: 1-6.

Pettitt, A., Weir, I.S., and Hart, A. G. (2002) A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modeling large sets of binary data. *Statistics and Computing* 12(4): 353-367.

Quintero, J. (2007) Regional Economic Development: An Economic Base Study and Shift and Shares Analysis of Hays County, Texas. Applied Research Project. Texas State University. URL: http://ecommons.txstate.edu/arp/259/

Reynolds, P. D., Miller, B., and Maki, W. R. (1995) Explaining Regional Variation in Business Births and Deaths: U.S. 1976-88. *Small Business Economics* 7: 389-407.

Song, J., Ghosh, M., Miaou, S., and Mallick, B. (2006) Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis* 97: 246-273.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003) WinBUGS User Manual Version 1.4. URL: http://voteview.org/manual14.pdf.

Tobler W. (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46 (2): 234-240.

United States Department of Agriculture (USDA) (2003) Rural-Urban Continuun Codes, Economic Research Service, URL: http://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx

Wall, M. (2004) A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning & Inference* 121: 311-324.

Wang, C., Quddus, M.A., and Ison, S.G. (2009) Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention* 41 (4): 798-808.

Wang, C., Quddus, M. A., and 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: 1979-1990.

Wang, X., and K. M. Kockelman. (2009) Application of the Dynamic Spatial Ordered Probit Model: Patterns of Land Development Change in Austin, Texas. *Papers in Regional Science* 88 (2): 345-366.

Wang, X., and Kockelman, K. (2009) Forecasting Network Data: Spatial Interpolation of Traffic Counts Using Texas Data. *Transportation Research Record* 2105: 100-108.

Wang, X., Kockelman, K., Lemp, J. (2012) The Dynamic Spatial Multinomial Probit Model: Analysis of Land Use Change Using Parcel-Level Data. Forthcoming in the *Journal of Transport Geography*.

Wang, Y., Kockelman, K., and Damien, P. (2012) A Spatial Autoregressive Multinomial Probit Model for Anticipating Land Use Change in Austin, Texas. Proceedings of IATBR's 13th International Conference on Travel Behavior Research Board, in Toronto.

Washington, S. P., Karlaftis, M. G., and Mannering, F. L. (2011) *Statistical and Econometric Methods for Transportation Data Analysis*. CRC Press, Chapman & Hall, Boca Raton, FL.

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

Zegeer, C., and Bushell, M. (2011) Pedestrian crash trends and potential countermeasures from around the world. *Accident Analysis & Prevention* 44 (1): 3-11.