Spatial Modelling of Highway Crash Risk in Minnesota

Cameron Faerber

A selected project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Dr. Matthew Heaton, Chair
Dr. Candace Barrett
Dr. Robert Richardson

Department of Statistics

Brigham Young University

April 2016

ABSTRACT

Spatial Modelling of Highway Crash Risk in Minnesota

Cameron Faerber
Department of Statistics, BYU
Master of Science

Systemic safety improvements to roadways, when selected and targeted appropriately, provide a tremendous opportunity to proactively reduce automobile crashes. While state and local agencies are starting to embrace the benefit of systemic safety improvements, additional guidance is needed to help these agencies select and target their improvements. As such, detailed information is needed on the crashes and contributing factors that are best targeted by systemic improvements and the types of locations and situations where these crashes occur. In this regard, the Highway Safety Information System (HSIS) collects data such as traffic volume, road characteristics, and number and type of crashes that, through statistical analysis, allow us to pinpoint locations with high crash risk and identify roadway features that lead to a higher crash risk (so called risk factors). In this project, we model the number of car crashes along road segments in Minnesota using Poisson regression with spatially correlated random effects. These spatial random effects serve to account for important covariates left out of the model or potential residual spatial correlation in the crash counts. Estimating both the coefficients associated with covariates and estimated spatial components, we are be able to determine (1) which road characteristics lead to higher accident rates and (2) which road segments have higher than expected crash rates for further investigation.

Keywords: Poisson regression, CAR model, spatial, LASSO, adaptive MCMC

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Overview

The Federal Highway Administration (FHWA) conducts research in an effort to improve road safety. In these research efforts, the FHWA collects data on the location of crashes along with characteristics of a given road. Knowing the location of accidents on a road can help the FHWA determine which segments of a road are dangerous (i.e. so called "hot spots"). Additionally, characteristics of the road give insight into why the crash happened. For example, if a section of the road containing strong curvature has a high rate of crashes, the FHWA can suggest safety improvements to that particular section of road in an effort to reduce the number of crashes.

The purpose of this project is to regress the number of car crashes on segments of a road onto variables such as road characteristics (type of pavement, shoulder width, number of lanes, etc.) and annual average daily traffic. The ultimate goal of this modeling is to be able to identify segments of road that have more accidents than would be expected given the traffic levels, and road characteristics. Using the results of this model, the FHWA can identify, and make changes to, dangerous road segments and, subsequently, reduce the number of crashes.

1.2 Highway Safety Information System Data

The Highway Safety Information System (HSIS) is a roadway-based system that provides data about crash, roadway, and traffic variables. HSIS is currently managed by the University of North Carolina Highway Safety Research Center (HSRC) under contract with FHWA. Importantly, each state involved in the HSIS research project collects data independently

Table 1.1: Name, and description, of each variable that is considered in this analysis.

Variable Name	Description
aadt	Average annual daily traffic
medwid	Median width (feet)
med_type	Median type
lshldwid	Left shoulder width (feet)
lshl_type	Left shoulder type
rshldwid	Right shoulder width (feet)
rshl_typ	Right shoulder type
surf_wid	Surface width (feet)
no_lanes	Total number of lanes
lanewid	Lane width

of HSIS, and HSIS is merely charged with collecting and processing the data on an annual basis. There are currently nine states in the HSIS: California, Illinois, Maine, Minnesota, Michigan, Ohio, North Carolina, Utah, and Washington. However, Utah and Michigan were removed from HSIS in 2000 and 1997, respectively, because of changes in inventory data collection. These nine states were selected to participate in HSIS primarily because of the availability of data and the high quantity and quality of information gathered.

Databases for each state include up to seven different data files: crash, roadway, traffic volume, curve/grade, intersection, and interchange. Given the goals of this analysis, only the crash and roadlog files were used. The crash data files for each state are split up into two separate files: a crash file and a vehicle file. The crash file contains information on each crash while the vehicle file contains information on each vehicle and each driver. The roadlog file contains information on homogeneous sections of roads such as number of lanes, lane width, annual average daily traffic, etc. that will be used to assess road characteristics that lead to a higher than expected volume of crashes. Due to the large volume of data in HSIS, this project will focus on modeling the data from Minnesota (MN). Table 1.1 displays all covariates, and their description, that will be used as a part of the analysis.

Figure 1.1 displays the observed number of crashes minus expected number of crashes (see Chapter 3 for a discussion of how to calculate the expected number of crashes) along

Interstate 35 (I-35) and Interstate 35 East (I-35E) in Minnesota. There is clearly a difference in the observed and the expected number of counts. We wish to model and explain the deviations from the expected counts. Notice the spatial "stickiness" in Figure 1.1 from road segment 180 to 200 on the I35. Figure 1.2 shows the major interstate and US highways. The noticeably spatially "sticky" piece in I-35, is the portion of freeway that goes through the Duluth-Superior metropolitan area in northern Minnesota.

1.3 Modeling Challenges

In modeling the HSIS data from MN, several challenges need to be addressed. First, as mentioned above, the data are subject to spatial correlation. Spatial correlation is the idea that locations that are close (in time or space) tend to have similar characteristics (e.g. similar high rates of crashes). In the case of this project, segments of road that are connected will most likely have a high correlation for various reasons including similar traffic patterns, similar road characteristics as well as closeness in space.

A second challenge in regards to this analysis is the non-Gaussian structure of the data. Namely, since HSIS data is in terms of the number of crashes on a given road segment, standard linear regression techniques are not appropriate. For example, linear regression would produce a model that could potentially predict negative and/or non-integer number of crashes.

Inherent in the data is some measure of collinearity. For instance, average annual daily traffic and the number of lanes are highly correlated as one would expect (a higher number of lanes typically indicates a greater volume of traffic flowing through a street). Additionally, as discussed in Hughes and Haran (2012), adding spatial random effects to a statistical model to account for potential spatial correlation adds collinearity to the main effects. Collinearity presents a problem because if present, the uncertainty estimates of the collinear effects will likely be inflated.

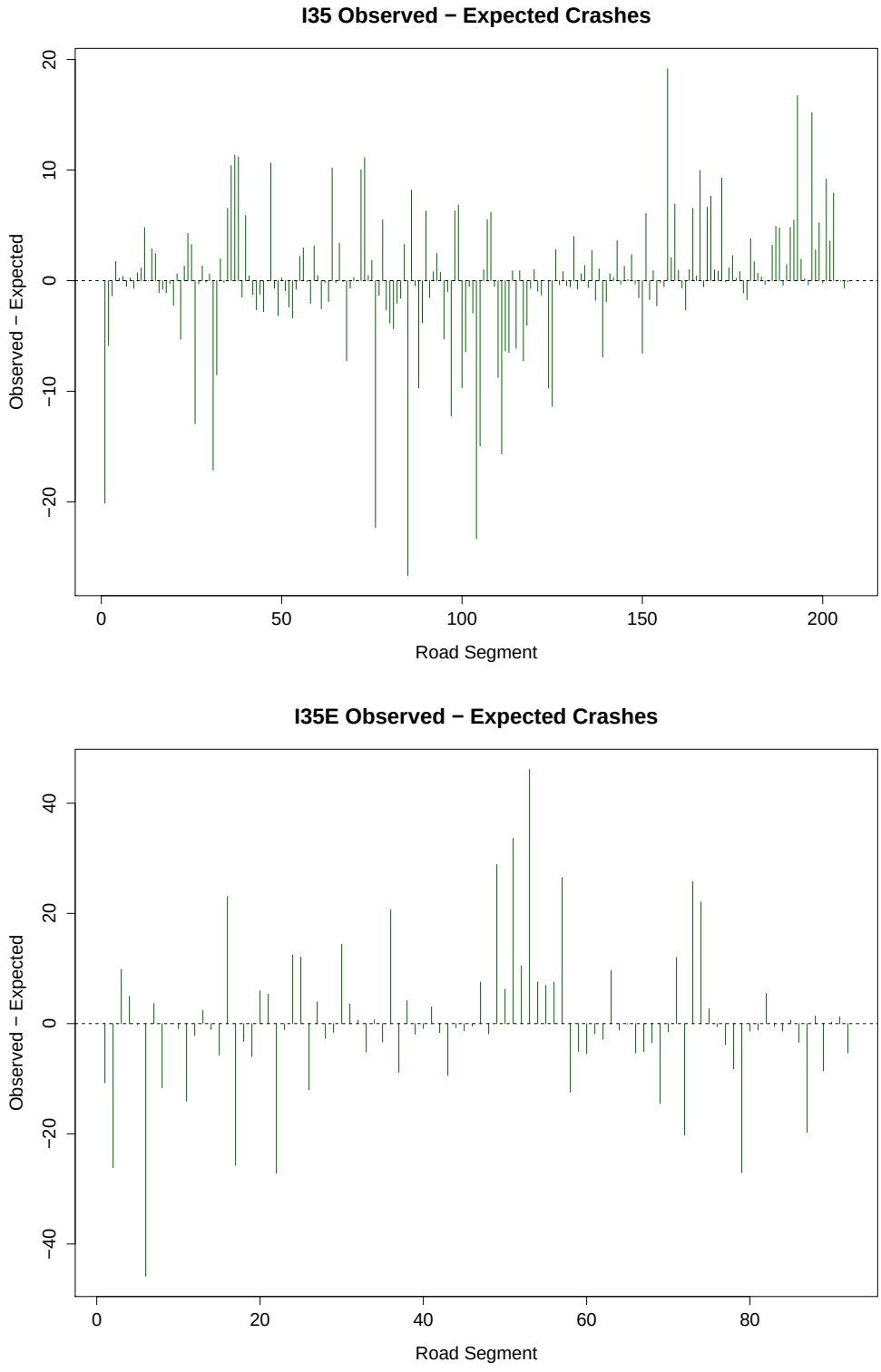

Figure 1.1: Observed - expected number of crashes along I-35 and I-35E in MN. The calculation of expected values is discussed in Chapter 3.

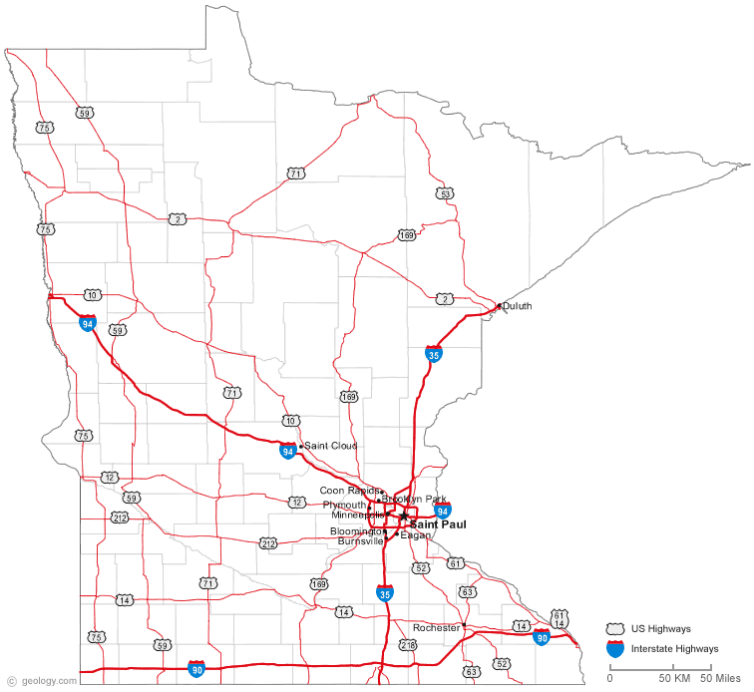

Figure 1.2: Major roadways in MN. I35 runs north to south (leading up to the twin cities metropolitan area but noth through), and I35E goes through the twin cities metropolitan area .

1.4 Methods Overview

In this project, we will address the above challenges by modeling the data using a Poisson regression with spatially varying effects. A Poisson regression ensures that the data have the proper support (integers greater than or equal to zero). Additionally, by including spatially varying effects we will capture the spatial dependency inherent in the data. To include spatial effects, we will use a conditional autoregressive (CAR) model. A CAR model is appropriate to fit the data because the data is areal (region specific as opposed to point specific locations). To correct for spatial collinearity, we will use the parameterization of the CAR model advocated by Hughes and Haran (2012) such that the spatial random effects are orthogonal to the main effects. Finally, a Bayesian LASSO will be used to select the most important main effects (road characteristics) that lead to increased or decreased crash rates.

CHAPTER 2

METHODS REVIEW

The purpose of the chapter is to provide a review of the main statistical methods used in this project. See the cited references for a more thorough discussion on each topic.

2.1 Poisson Regression

Poisson regression is used to model count data and/or contigency tables. For example, Poisson regression can be used to model the number of incoming calls in a call center for different time segments throughout the day or, as in the case in this project, the number of car crashes on a segment of road. Let Y_i for $i = 1, \ldots, n$ denote a count response variable for the i^{th} individual (e.g. road segment) and n is the total number of individuals in the dataset (the sample size). Furthermore, define $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{iP})'$ to be a vector of P covariates (and an intercept term) associated with Y_i. Poisson regression is a tool to estimate the relationship between Y_i and $\boldsymbol{x}_i$.

As described in Weisberg (2005), the three main components of Poisson regression are the following.

1. The conditional distribution of $Y_i|\boldsymbol{x}_i$ is referred to as the likelihood. As the name implies, Poisson regression assumes $Y_i \mid \boldsymbol{x}_i \sim \mathcal{P}(\mu(\boldsymbol{x}_i))$ where $\mathcal{P}(\cdot)$ denotes the Poisson distribution and $\mu(\boldsymbol{x}_i) > 0$ is the mean of the Poisson distribution and is a function of the covariate vector $\boldsymbol{x}_i$.

2. The link function, denoted by $\ell(\cdot)$, is responsible for mapping the support of $\mu(\boldsymbol{x}_i)$ to the real line. In Poisson regression, $\ell(\cdot)$ is typically taken to be the log-link function because $\ell(\mu(\boldsymbol{x}_i)) = \log(\mu(\boldsymbol{x}_i)) \in (-\infty, \infty)$.

3. Finally, $\ell(\mu(\boldsymbol{x}_i))$ is related to $\boldsymbol{x}_i$ through the linear predictor $\ell(\mu(\boldsymbol{x}_i)) = \log(\mu(\boldsymbol{x}_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)'$ is a vector of unknown coefficients to be estimated from the data.

Importantly, estimates of $\boldsymbol{\beta}$ cannot be solved for explicitly (as in normal linear regression) and, therefore, numerical methods are used to obtain the maximum likelihood estimates (denoted by $\hat{\boldsymbol{\beta}}$).

Because we are using the log-link function to ensure the mean parameter of the Poisson distribution is strictly greater than zero, interpretation of the $\boldsymbol{\beta}$ coefficients is not as simple as in traditional linear regression. For example, consider β_1, the coefficient associated with the first covariate x_1. The interpretation of β_1 is as follows: for each one unit increase in x_1, holding all else constant, the log expected count increases by β_1. Because the log-scale is not always intuitive, the coefficients are often interpreted on the exponential scale. For example, each one unit increase in x_1, holding all else constant, increases the expected count, multiplicatively, by $\exp\{\beta_1\}$ (i.e. as x_1 increases by one, holding all else constant, the mean $\mu(\boldsymbol{x}_i)$ is $\exp\{\beta_1\}$ times larger). Oftentimes, an even simpler (and less informative) interpretation is used based on the sign of the coefficient. If the $\beta_1 > 0$, increases in x_1 are associated with increases in the expected count while if $\beta_1 < 0$, increases in x_1 are associated with decreases in the expected count.

The three main assumptions of Poisson regression are:

1. observations are independent conditional on the covariates $\boldsymbol{x}_i$,
2. $Y_i|\boldsymbol{x}_i \sim \mathcal{P}(\mu(\boldsymbol{x}_i))$, (i.e. the Poisson assumption of a squared relationship between mean and variance), and,
3. the log-transformed expected counts, $\log(\mu(\boldsymbol{x}_i))$, have a *linear* association with the covariates.

The distributional assumption that $Y_i|\boldsymbol{x}_i \sim \mathcal{P}(\mu(\boldsymbol{x}_i))$ is difficult to test. Faraway (2006) suggests a crude test, by plotting $\hat{\mu}(\boldsymbol{x}_i) = \exp\{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\}$ and $(y_i - \hat{\mu}(\boldsymbol{x}_i))^2$. This can give a general

idea if the variance is somewhat close to the mean. An alternative to this method would be to group similar values of covariates together and compare the mean of their expected values with the variance within each group. To check condition 3 (linear relationship between log expected counts and the covariates), plots of log counts vs. each covariate can be used to assess linearity. Independence of the observations conditional on the covariates is difficult to test for and can be assumed unless there is a good reason to suppose they are not independent (as is the case in this project due to adjacent spatial locations).

To test whether or not the Poisson regression model is correctly specified, residual deviance (G^2) is often used. If the Poisson mean function is correctly specified, the residual deviance $G^2 \sim \chi^2(n-(P+1))$ where n is the number of individuals and $P+1$ is the rank of the design matrix $\boldsymbol{X}$. G^2 is calculated as $G^2 = 2\sum_{i=1}^{n}(y_i \log(\frac{y_i}{\hat{\mu}(\boldsymbol{x}_i)})) - (y_i - \hat{\mu}(\boldsymbol{x}_i)))$. An alternative to the G^2 statistic is Pearson's χ^2 which is given by $\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}(\boldsymbol{x}_i))^2}{\hat{\mu}(\boldsymbol{x}_i)}$. Similar to the G^2 statistic, χ^2 follows a $\chi^2(n-(P+1))$ distribution if the model fits well. At large sample sizes, the G^2 and χ^2 give the same inference, but at smaller sample sizes, Pearson's χ^2 is more powerful.

2.2 Conditional Autoregressive Models

Conditionally autoregressive models (CARs; Banerjee et al. (2015); Cressie and Wikle (2011)) were originally developed by Besag (1974) and are used to capture spatial correlation in areal data (areal data is data aggregated over regions such as states, countries, road segments, etc.). For this subsection, let Y_i denote a random response variable associated with spatial areal unit i where $i = 1, \ldots, n$. CAR models assume that,

$$Y_i \mid \{y_j\}_{j \neq i} \sim \mathcal{N}\left(\sum_j b_{ij} y_j, \tau_i^2\right), \tag{2.1}$$

for all i. The $\{b_{ij}\}$ in (2.1) are spatial weights where larger b_{ij} denote a stronger weight, or connection, between areal unit i and j. For reasons as will be seen below, $\{b_{ij}\}$ are typically specified by the researcher.

The set of equations in (2.1) define the full set of complete conditional distributions rather than the joint distribution of $\boldsymbol{y} = (Y_1, \ldots, Y_N)'$. However, through Brook's Lemma, the joint distribution of $\boldsymbol{y}$, up to a proportionality constant, is given by

$$p(\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{y}'\boldsymbol{D}^{-1}(\boldsymbol{I}-\boldsymbol{B})\boldsymbol{y}\right\} \tag{2.2}$$

where $\boldsymbol{B} = \{b_{ij}\}$, $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{D}$ is diagonal with $D_{ii} = \tau_i^2$. Upon inspection of (2.2), it appears that the set of complete conditionals in (2.1) implies a multivariate normal distribution for $\boldsymbol{y}$. However, for this to be the case, it must be ensured that $\boldsymbol{D}^{-1}(\boldsymbol{I}-\boldsymbol{B})$ (the implied precision matrix) is symmetric. The symmetry condition is ensured by enforcing the constraint, $b_{ij}/\tau_i^2 = b_{ji}/\tau_j^2$ for all i, j.

Note from (2.2), that the matrix $\boldsymbol{B}$ is associated with the precision matrix of the vector $\boldsymbol{y}$. However, in the majority of settings, only a single vector $\boldsymbol{y}$ will be observed. Hence, $\boldsymbol{B}$ cannot be estimated from the data and needs to be user specified. One common specification, referred to as the intrinsic CAR (ICAR) model, is to define a weight or "proximity" matrix $\boldsymbol{W}$ where the ij^{th} element of $\boldsymbol{W}$, denoted by w_{ij}, gives a weight to the association between areal units i and j. As the CAR model is typically specified for spatial data, the weights w_{ij} are larger for areal units that are closer to one another. The ICAR model sets $b_{ij} = w_{ij}/w_{i+}$ where $w_{i+} = \sum_{j \neq i} w_{ij}$ and $\tau_i^2 = \tau^2/w_{i+}$. This specification leads to the set of complete conditionals,

$$Y_i \mid \{Y_j\}_{j \neq i} \sim \mathcal{N}\left(\frac{1}{w_{i+}}\sum_{j \neq i} w_{ij} y_j, \tau^2/w_{i+}\right). \tag{2.3}$$

In (2.3), note that the distribution of Y_i given $\{Y_j\}_{j \neq i}$ is a weighted average of the other areal units. This specification, when w_{ij} is larger for areal units closer in space, allows for a spatial smoothing of the areal random variables. Additionally, the variance of Y_i conditional on $\{Y_j\}_{j \neq i}$ is smaller for those locations where w_{i+} is bigger (i.e. those locations with many areal units in close proximity).

Under (2.3), the joint distribution of $\boldsymbol{y}$ now becomes

$$p(\boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{y}'(\boldsymbol{D}_w - \boldsymbol{W})\boldsymbol{y}\right\}, \tag{2.4}$$

where $\boldsymbol{D}_w$ is a diagonal matrix with ii^{th} entry w_{i+}. Inspection of (2.4) reveals one problem with this joint density, namely, the precision matrix $(\boldsymbol{D}_w - \boldsymbol{W})$ is singular (therefore the implied covariance matrix of the multivariate normal distribution does not exist). The solution to this problem is to enforce contraints on $\boldsymbol{y}$ (the solution used in this project follows Hughes and Haran (2012)). As these constraints are an important part of the model used in this project, discussion is deferred to Chapter 3.

The ICAR model is widely used in spatial statistics literature. Example applications include ecology (He and Sun 2000), disease mapping (Waller et al. 1997; Lawson 2013), mortality estimation (MacNab and Dean 2000), real estate prices (Pace et al. 2000) and climate (Zhou et al. 2012).

2.3 The LASSO

In linear models where the number of observations is significantly greater than the number ($n \gg p$), the least squares estimates tend to have low variance. When the difference between n and p is not quite so large, there can be a lot of variability in the model fit. In order to compensate for this, James et al. (2013) suggest using shrinkage and variable selection techniques. Shrinkage techniques constrain (or "shrink") coefficient values closer to zero, and in the case of LASSO models, shrink some of the coefficients all the way to zero (a method of variable selection). While there is some bias introduced in the model, the decrease in the variance of the model is substantial leading to improved mean square error.

The typical linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ traditionally, estimates $\boldsymbol{\beta}$ using maximum likelihood yielding the closed form solution $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Lasso estimates, in contrast, find values, say $\hat{\boldsymbol{\beta}}_\lambda$, that minimize the quantity $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$. Here λ is a penalty parameter that penalizes the model for having large coefficient estimates and

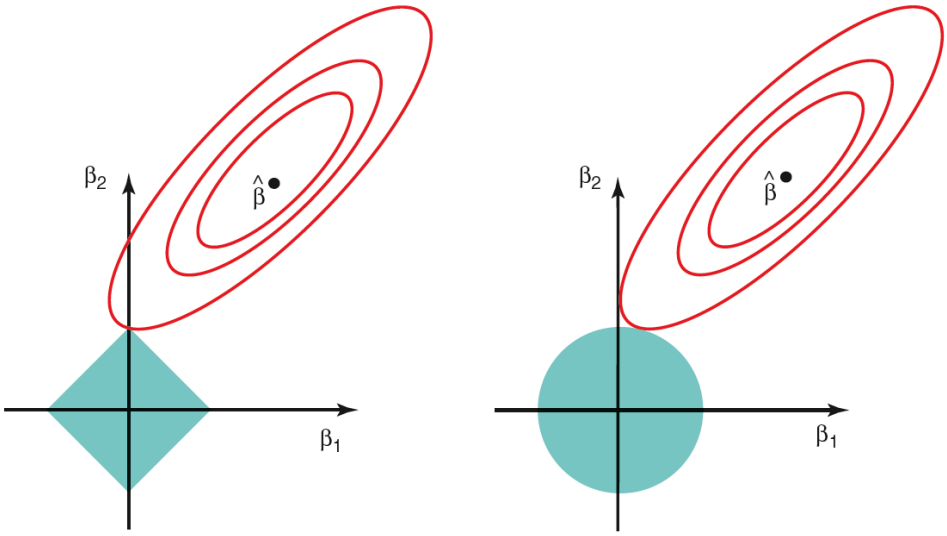

Figure 2.1: Two types of regression regularization. On the left shows is LASSO regression, on the right is Ridge regression. Figure from James et al. (2013).

effectively, acts as a constraint on the β coefficient that may pull some of the coefficients to zero. Figure 2.1, from James et al. (2013) shows how the LASSO is able to zero out coefficients.

Casella and Park (2008) suggested the Lasso estimates are comparable to the mode of the posterior distributions for the regression parameters when using identical Laplace priors. The prior on $\boldsymbol{\beta}$ is (setting the location parameter μ to zero):

$$p(\boldsymbol{\beta}|s) = \left(\frac{1}{2s}\right)^p \exp\left\{-\frac{\sum_{i=1}^{p}|\beta_i|}{s}\right\} \tag{2.5}$$

When we combine the prior on $\boldsymbol{\beta}$ with a normal likelihood it becomes proportional to $\exp\left\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{s}\sum_{i=1}^{p}|\beta_i|\right\}$. Now s is similar to the λ penalty of the traditional LASSO approach.

2.4 Markov chain Monte Carlo

Bayesian inference is a method of statistical inference which uses Bayes' theorem to update probabilities of a hypothesis using data. Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.6}$$

While Equation 2.6 refers to simple, discrete events, it can be used to determine conditional densities as well. The conditional densities of interest are known as posterior densities which are of the following form:

$$\pi(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \tag{2.7}$$

where $f(x|\theta)$ is often referred to as the likelihood, $f(\theta)$ is known as the prior, and $f(x) = \int_\theta f(x|\theta)f(\theta)d\theta$.

In the majority of statistical models, the posterior distribution in Equation (2.7) is only known up to the proportionality constant. This unknown proportionality prohibits calculation of posterior summaries such as the expectation, variance, and quantiles of interest. To overcome this obstacle, posterior summary statistics are calculated via Monte Carlo integration using a large sample from the posterior distribution. However, because the posterior distribution is often of unknown form, Markov chain Monte Carlo (MCMC) are a class of algorithms that can be used to sample from the posterior distribution (see Robert and Casella 2004; Gamerman and Lopes 2006). While there are many MCMC algorithms, this section will focus on only one such algorithm: the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970; Chib and Greenburg 1995).

Let $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ represent the posterior distribution of interest that can be evaluated for any value $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ up to a proportionality constant. The Metropolis-Hastings algorithm is as follows:

1. Choose a starting value: $\boldsymbol{\theta}_0$

2. For iteration $t = 1, \ldots, T$

a) Propose a new value by drawing $\boldsymbol{\theta}^\star \sim q_\sigma(\cdot \mid \boldsymbol{\theta}_{t-1})$ where $q_\sigma(\cdot \mid \boldsymbol{\theta}_{t-1})$ is a common density function (e.g. the normal) with scale parameter σ. Note that the proposal distribution q can depend on $\boldsymbol{\theta}_{t-1}$ but this does not necessarily have to be so.

b) Calculate the following ratio:

$$R = \frac{q_\sigma(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}^\star)\pi(\boldsymbol{\theta}^\star \mid \boldsymbol{y})}{q_\sigma(\boldsymbol{\theta}^\star \mid \boldsymbol{\theta}_{t-1})\pi(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{y})} \tag{2.8}$$

c) Set $\boldsymbol{\theta}_t = \boldsymbol{\theta}^\star$ with probability $\min(1, R)$.

Using properties of Markov chains, Metropolis et al. (1953) show that as $T \to \infty$ then draws $\boldsymbol{\theta}_t$ will resemble random vectors drawn from $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$.

In practice, having a proper proposal distribution q_σ is key to obtaining draws from $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ in a finite amount of time. Typically, a symmetric proposal density is used because R in (2.8) reduces $R = \pi(\boldsymbol{\theta}^\star)/\pi(\boldsymbol{\theta}_{t-1})$ since $q_\sigma(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}^\star) = q_\sigma(\boldsymbol{\theta}^\star \mid \boldsymbol{\theta}_{t-1})$ Even when q_σ is symmetric, the choice of the scale parameter σ is important: if σ is too large then the Markov chain will rarely accept proposed values but a σ too small will produce a chain that does not efficiently explore the space of the posterior distribution. For this reason, Haario et al. (2001) proposed using an adaptive metropolis algorithm.

Suppose at iteration t, the chain has sampled points $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{t-1}$. The adaptive algorithm of Haario et al. (2001) lets q_σ be a Gaussian distribution centered at the current point, with covariance matrix equal to $\boldsymbol{C}_t = (2.4^2/P) \times \text{Cov}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{t-1}) + (2.4^2/P)\epsilon I_P$ where P is the dimension of $\boldsymbol{\theta}$. The $\epsilon > 0$ is a chosen small constant and I_P is a P-dimensional identity matrix. In practice, a starting covariance matrix C_0 is chosen and then choose an index $t_0 \in \{1, \ldots, T\}$ such that:

$$C_t = \begin{cases} C_0, & \text{if } t \leq t_0 \\ \frac{2.4^2}{P} \times \text{Cov}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{t-1}) + \frac{2.4^2}{P}\epsilon I_P, & t > t_0 \end{cases}$$

The parameter ϵ must be greater than zero to ensure that the covariance matrix remains positive definite but ϵ is preferrably very small.

CHAPTER 3

A SPATIAL MODEL FOR CAR CRASH RATES

This chapter provides rigorous detail on the statistical model used to identify risk factors and trouble spots for highway safety.

3.1 Model

Let Y_s denote the number of crashes that occur on road segment s for $s = 1, \ldots, S$. Because Y_s is a count, we assume

$$Y_s \overset{ind}{\sim} \mathcal{P}\left(E_s \mu_s\right) \tag{3.1}$$

where $\mathcal{P}(\mu)$ represents the Poisson distribution with mean μ. In (3.1), E_s is the *known* expected number of crashes that occur in segment s and is calculated as

$$E_s = \frac{\sum_j Y_j}{\sum_j (\text{AADT}_j \times \text{length}_j)} \times \text{AADT}_s \times \text{length}_s \tag{3.2}$$

where AADT_s is the average annual daily traffic and length_s is the length (in miles) of segment s. Note that, from (3.2), as the traffic in segment s increases, so does E_s. Likewise, E_s increases with length_s suggesting that traffic and a long segment are associated with a higher number of crashes.

The μ_s in (3.1) can be thought of as a relative risk. If $\mu_s > 1$, then segment s has a higher than expected number of crashes, while if $0 < \mu_s < 1$ then segment s has a lower than expected number of crashes. Estimating μ_s for all road segments can identify which segments of road have a higher than expected number of crashes, i.e. dangerous segments of road.

To identify risk factors while accounting for spatial correlation, we set

$$\log(\mu_s) = \boldsymbol{x}_s' \boldsymbol{\beta} + \phi_s \tag{3.3}$$

where $\boldsymbol{x}_s = (1, x_{s1}, \ldots, x_{sP})'$ is a vector of covariates describing the roadway (see Table 1.1) and ϕ_s is a spatial random effect. The vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)'$ quantifies the effect of each of the road factors on the log-relative risk ($\log(\mu_s)$) for road segment s. Specifically, if $\beta_p > 0$ ($\beta_p < 0$) then increases in x_{sp} are associated with increases (decreases) in the relative risk of car accidents.

The ϕ_s captures the spatial dependence among crash rates as seen in the data in Figure 1.1 and is traditionally modeled using a CAR model. However, as shown in Hughes and Haran (2012), the ϕ_s can potentially confound the main effects ($\boldsymbol{\beta}$) leading to an increase in the posterior variance of the $\boldsymbol{\beta}$. This is highly undesirable for this project because $\boldsymbol{\beta}$ will identify roadway characteristics that are potentially harmful. To correct for this, we follow Hughes and Haran (2012) and constrain the spatial effects to be orthogonal to the main effects. To do this, let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_s) = \boldsymbol{M}\boldsymbol{\phi}^\star$ where $\boldsymbol{M}$ is a known matrix (see below) and $\boldsymbol{\phi}^\star$ is a vector of unknown coefficients. Under this model we can write (3.3) jointly as:

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{\phi}^\star \tag{3.4}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_S)'$ is the vector of relative risks. In the framework of Equation (3.4), constraining the spatial effects $\boldsymbol{\phi}$ to be orthogonal to the main effects $\boldsymbol{\beta}$ amount to ensuring that $\boldsymbol{M}$ is orthogonal to the design matrix $\boldsymbol{X}$. In this project, we take $\boldsymbol{M}$ to be the Moran Operator from Hughes and Haran (2012) and is obtained using the following steps:

1. Calculate $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

2. Calculate $\boldsymbol{M}^\star = \boldsymbol{P}^\perp \boldsymbol{W} \boldsymbol{P}^\perp$ where $\boldsymbol{W}$ is the weight (adjacency) matrix for the CAR model described in Chapter 2.

3. Calculate the eigendecomposition of $\boldsymbol{M}^\star = \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}'$ where $\boldsymbol{C}$ is the matrix of (column)-eigenvectors and $\boldsymbol{D}$ is the diagonal matrix of eigenvalues.

4. Set $\boldsymbol{M}$ equal to the Q columns of $\boldsymbol{C}$ associated with the Q largest eigenvalues in $\boldsymbol{D}$ where $Q \ll S$.

To define $\boldsymbol{M}$ as above, we must choose an appropriate value for Q (the number of eigenvectors taken from $\boldsymbol{C}$ to create the Moran operator). The parameter Q can be thought of as a smoothing parameter: the smaller the value of Q, the smoother crash rates. Instead of arbitrarily choosing a value for Q, we will use deviance information criterion (DIC; Spiegelhalter et al. 2002) to determine the most appropriate value for Q. Specifically, we will compare DIC values for Q ranging from 10 to 100, and choose the value of Q that is associated with the lowest DIC value. Because selecting Q is computationally challenging, we will assume that the most appropriate value of Q is constant across all highways in Minnesota.

As prior distributions, the prior on the spatial random effects $\boldsymbol{\phi}^\star$ is

$$p(\boldsymbol{\phi}^\star|\tau) \propto \tau^{Q/2} \exp\left\{-\frac{\tau}{2}\boldsymbol{\phi}^* \boldsymbol{M}' \left[\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}\right] \boldsymbol{M}\boldsymbol{\phi}^\star\right\} \tag{3.5}$$

where, again, Q is the number of eigenvectors chosen from $\boldsymbol{C}$ to form the Moran operator $\boldsymbol{M}$ and $\boldsymbol{W}$ is the weight (adjacency) matrix from the CAR model. Finally, we use a conjugate gamma prior on τ such that $\tau \sim \mathcal{G}(2.01, 1)$ where $\mathcal{G}(a, b)$ denotes the Gamma distribution with shape parameter a and rate parameter b. Under this prior specification, the complete conditional posterior distribution of τ is also a gamma distribution with shape parameter $Q/2 + 2.01$ and rate parameter $1/2\boldsymbol{\phi}^\star \boldsymbol{M}' \left[\text{diag}(\boldsymbol{W}\mathbf{1}) - \boldsymbol{W}\right] \boldsymbol{M}\boldsymbol{\phi}^\star + 1$.

For the coefficients to roadway factors, we assume that $\beta_p \overset{iid}{\sim} \mathcal{L}(0, t)$ for $p = 1, .., P$, where $\mathcal{L}(m, t)$ denotes the Laplace distribution with location parameter μ and scale parameter t. As described in Chapter 2, the Laplace distribution corresponds to the LASSO penalty. Hence, our choice of prior here shrinks coefficients to zero and allows for variable selection. The scale parameter t determines the strength of the LASSO penalty term, or in other words, determines how fast the β_p coefficients are pulled towards zero. Since we don't have information on what value of t to use, we place a hyper-prior on t such that $t \sim \mathcal{IG}(2.01, 1)$ where $\mathcal{IG}(a, b)$ denotes the inverse-gamma distribution with shape a and rate b. An inverse gamma hyperprior is a conjugate prior distribution for t. In other words, the complete con-

ditional distribution of t will also be an inverse-gamma with shape parameter $2.01 + P$ and rate parameter $\sum_{p=1}^{P} |\beta_p| + 1$.

3.2 Model Fitting

In the model described above, the unknown parameters are $\boldsymbol{\beta}$, $\boldsymbol{\phi}^\star$, t and τ. We use the adaptive Markov chain Monte Carlo described in Section 2.4 to obtain posterior summaries of these parameters. Specifically, we "update" the unknown parameters in three blocks. First, we update $(\boldsymbol{\beta}, \boldsymbol{\phi}^\star)$ jointly. For the first 250 draws, we generate a proposal from a multivariate Gaussian distribution centered at the current value and covariance matrix $(0.01^2)\boldsymbol{I}$. The covariance matrix $(.01^2)\boldsymbol{I}$ is used because it allows enough proposals to be accepted such that the MCMC can begin to adapt. After the 250th draw, we generate proposals from the adaptive covariance matrix described in Section 2.4. The value for ϵ was set to $.01^2$.

After updating $(\boldsymbol{\beta}, \boldsymbol{\phi}^\star)$ jointly, t and τ are updated sequentially. Because conjugate priors were used for both of these parameters, they are easily updated by obtaining a random draw from their respective complete conditional distributions.

We ran the MCMC for 120,000 draws, after a burn-in period of 15,000 leaving a total of 105,000 draws. Appropriate convergence of the Markov chain was assessed using Monte Carlo standard errors (Jones et al. 2006). The starting values for the vector $(\boldsymbol{\beta}, \boldsymbol{\phi}^\star)$ were chosen to be the maximum likelihood estimates. The starting values for t and τ were set to 1.

CHAPTER 4

RESULTS

4.1 Model Diagnostics

To perform inference, 120,000 (15,000 burn-in) draws of the unknown parameters from the posterior distribution were obtained using the adaptive Markov Chain Monte Carlo algorithm described in Chapter 3. Figure 4.1 shows trace and autocorrelation plots for randomly selected marginal posteriors of β and ϕ parameters, as well as the marginal posteriors for both s and τ.

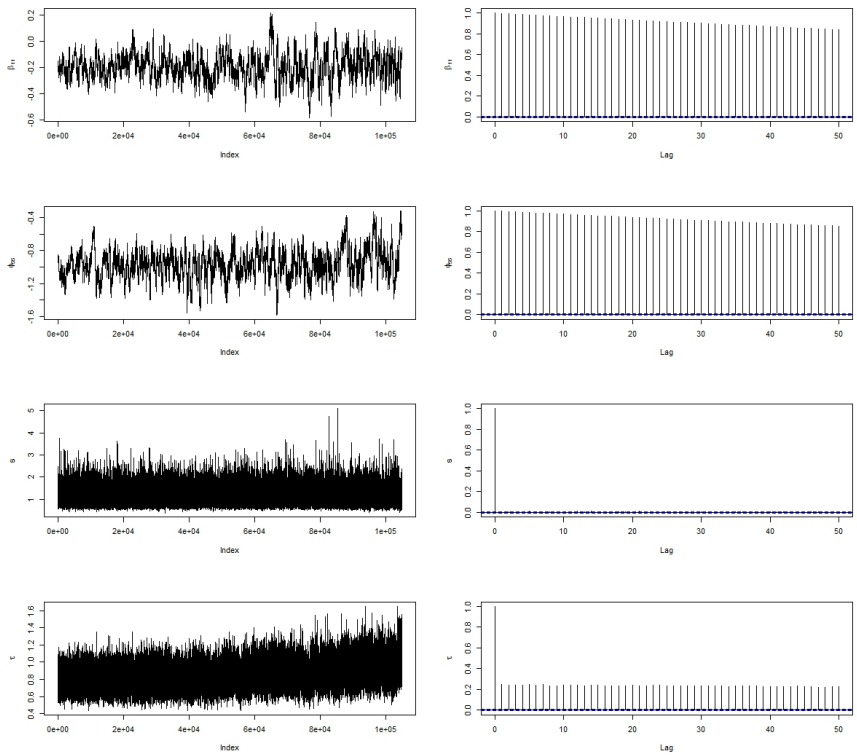

Figure 4.1: Trace and autocorrelation plots to assess convergence

Marginal distributions for nearly all β and ϕ parameters have very high autocorrelation that can be seen in Figure 4.1. Despite the high autocorrelation, the trace plots show that the MCMC has sufficiently explored the parameter space. The trace and autocorrelation plots for s and τ indicate adequate convergence. We further verify convergence of the MCMC draws using Monte Carlo (MC) standard errors (Jones et al. 2006). All of the MC standard errors for the β's were less than .01 except for the intercept (β_0) which has a standard error of .014. Likewise all of the MC standard errors for the ϕ parameters, except thirteen, were smaller than .01. However, the largest MC standard was .011 suggesting sufficiently high precision in the associated posterior summaries.

To verify that the fitted model explains the data well, Figure 4.2 shows 95% predictive intervals for each of the 207 road segments on I-35. Of the 207 intervals, 197 contain the observed value ($\approx$ 95% coverage) suggesting that the model accurately captures the variation seen in the data. Likewise, for I-35E, the predictive intervals have 97% coverage also suggesting adequate model fit for I-35E.

4.2 Results

The quantity $\mu_s = \exp\{\boldsymbol{x}_s'\boldsymbol{\beta} + \phi_s\}$ denotes the relative risk (relative to the expected number of crashes E_s from (3.2)) for each road segment where $\mu_s > 1$ indicate regions of elevated risk. Figure 4.3 displays 95% central credible intervals for each μ_s. Segments in red correspond to locations where the posterior probability that $\mu_s > 1$ is greater than 0.95 whereas segments in blue correspond to those regions where the posterior probability of $\mu_s < 1$ is greater than 0.95. In this context, segments in red can be viewed as road segments that pose a high risk to travelers while blue segments are those that are safer than expected. On the I35, three percent of segments (7) have relative risks of over two. I35E has, in general, a much higher relative risk rate than I35. Ten percent (10) road segments have relative risks of higher than 2 and over one-third of segments on the I35E have relative risks over one. The road segments with a high relative risk should have high priority for systemic improvement.

Table 4.1: 95% central credible intervals for the coefficient of each road characteristic included in this analysis for I35 and I35E.

	I35			I35E		
	2.5%	97.5%	Mean	2.5%	97.5%	Mean
Intercept	-1.690	-0.460	-1.038	2.602	4.081	3.380
Median width < 30ft	0.557	1.053	0.828	-0.869	0.095	-0.398
Median width varies	0.222	0.492	0.366	-0.940	-0.201	-0.567
No median barrier	-0.241	0.255	0.007	-1.126	-0.294	-0.753
Left shoulder width (ft)	-0.120	-0.005	-0.064	-0.175	-0.029	-0.105
Left shoulder type other	-0.182	0.051	-0.069	-0.115	0.362	0.114
Road surface width (ft)	-0.012	0.050	0.021	-0.045	0.054	0.010
Right shoulder width	0.078	0.219	0.150	-0.139	0.023	-0.052
Number of lanes	-0.466	0.015	-0.235	-0.357	0.229	-0.089
Lane width (ft)	-0.006	0.010	0.002	-0.199	-0.050	-0.129

As described in Chapter 3, the β coefficients give the increase (or decrease) in the relative risks associated with an increase in each road characteristic. For example, if $\beta_p > 0$ then increases in road characteristic x_p are associated with increased relative risk of accidents. Table 4.1 shows the central 95% credible intervals for the intercept and each of the 9 road characteristics. Interestingly, road characteristics have different effects on two highways. On the I35, four covariates are above zero: poor pavement conditions, small and/or variable median width, road width, and right shoulder width. Small median width in particular stands out as having a large, positive association with increased risk of accidents. Hence, safer roads typically have large, constant, median widths, wide left shoulders and fewer lanes. On the I35E, the majority of road characteristics are below zero. Only left shoulder type and road width have a $P(\beta_i > 0)$ of over .5 with probabilities of .85 and .66 respectively.

The covariates in the model do not a constitute a complete description of each road segment. For this reason the model in Chapter 3 included spatial random effects. Figure 4.4 includes 95% credible intervals for the spatial effects on each road segment. Intuitively, segments with a high (low) spatial random effect correspond to areas where unobserved covariates may be increasing (decreasing) the relative risk.

From Figure 4.4, notice that segments 180-200 on the I35 (between mile markers 250 and 256) correspond to large spatial random effects. Hence, these road segments have a higher relative crash rate than we would expect given the observed covariates. Segments 180-200 are associated with freeway segments in the Duluth-Superior metropolitan area. We hypothesize that potential reasons why crash rates are higher in Duluth include not only unobserved characteristics of the road such as high road curvature and steep grades but also socio-economic variables such as a higher than average number of bars per capita. However, additional socioeconomic data would be required to verify such claims.

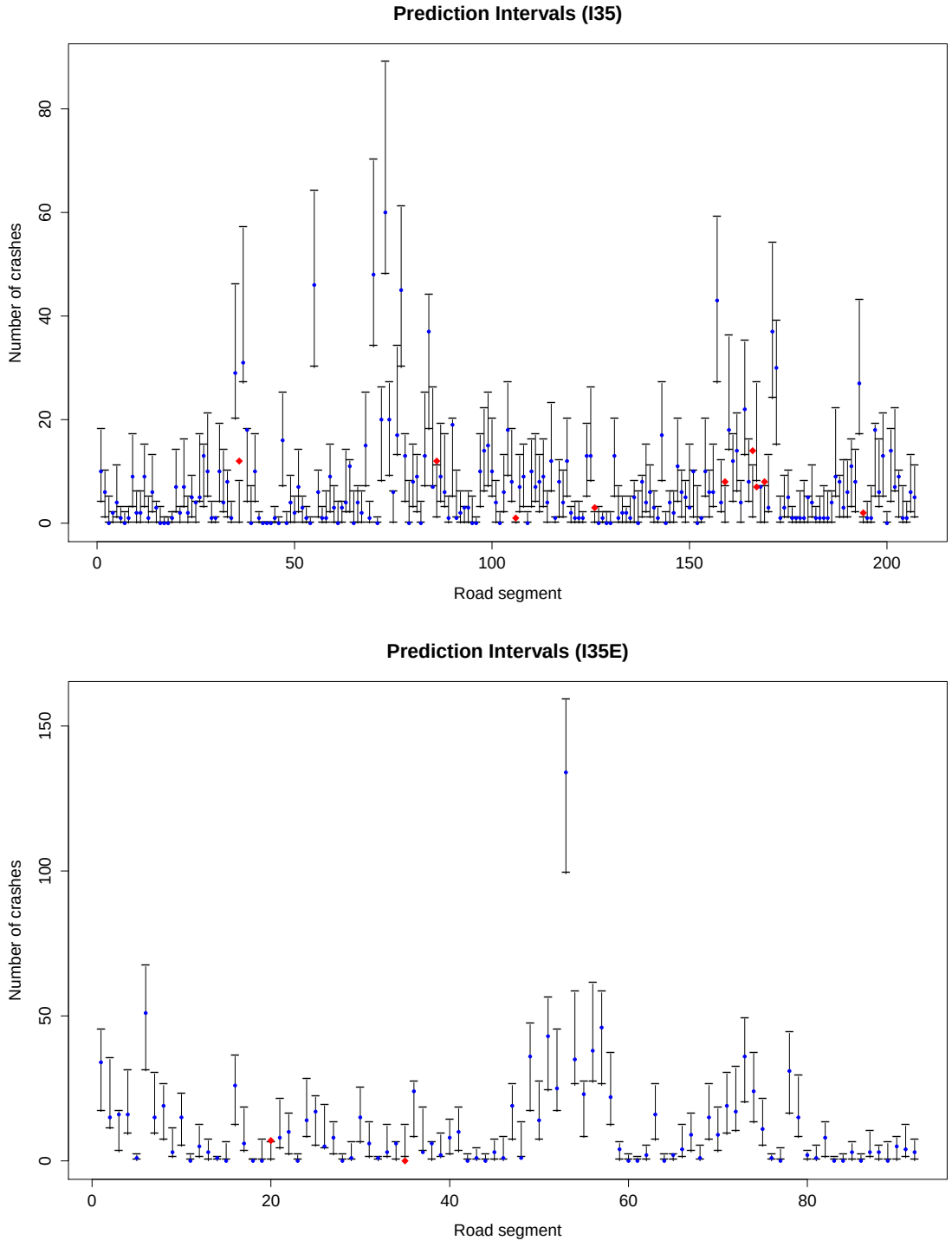

Figure 4.2: Posterior predictive intervals: blue dots indicate the observed number of crashes fell within a 95% interval, while red dots indicate the observed crashes did not fall within the interval.

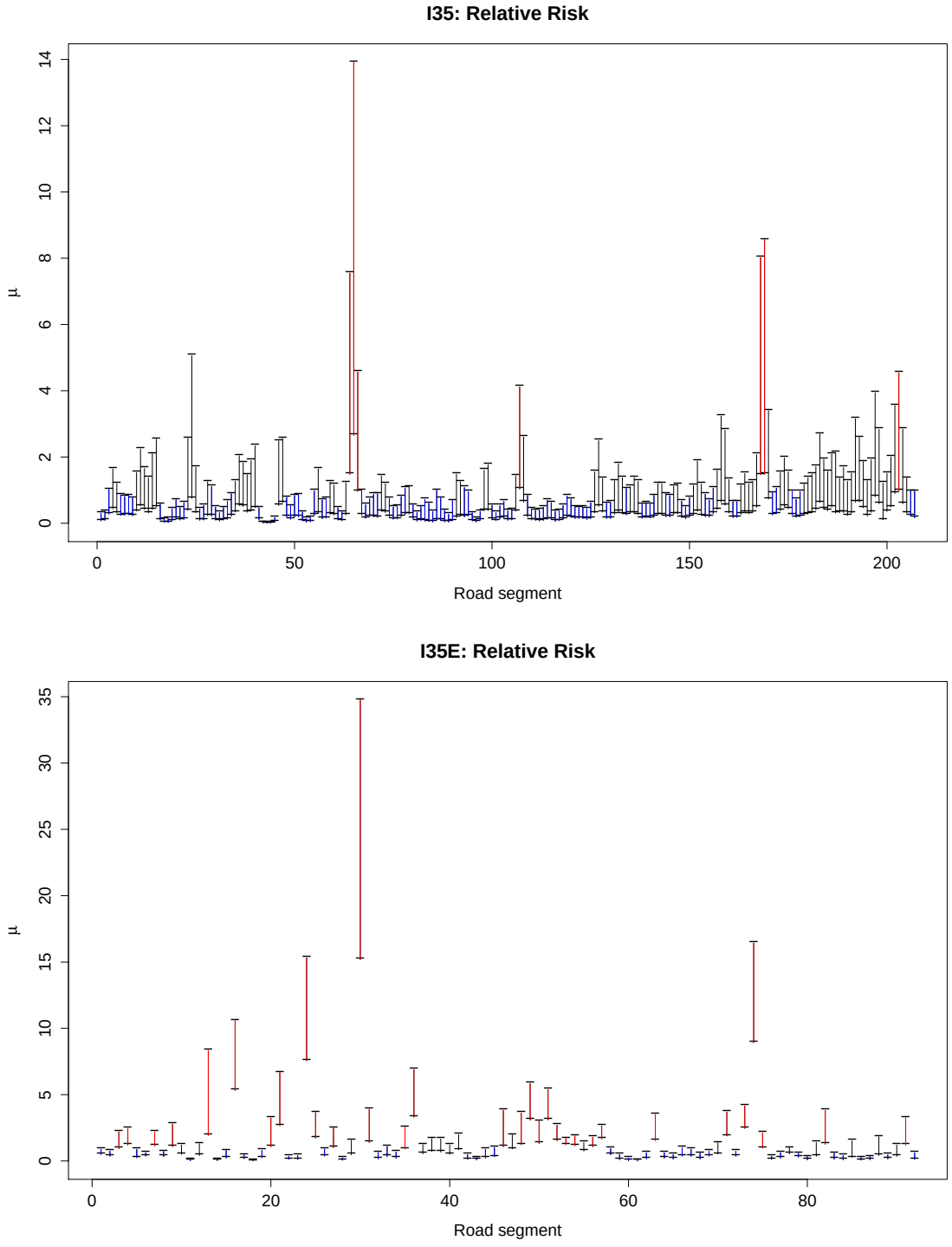

Figure 4.3: 95% credible intervals for the relative risks. Red lines indicate that 95% of draws are above 1 (indicating an increased risk of an accident) and blue lines indicate that 95% of draws are below 1 (indicating a decreased risk of an accident).

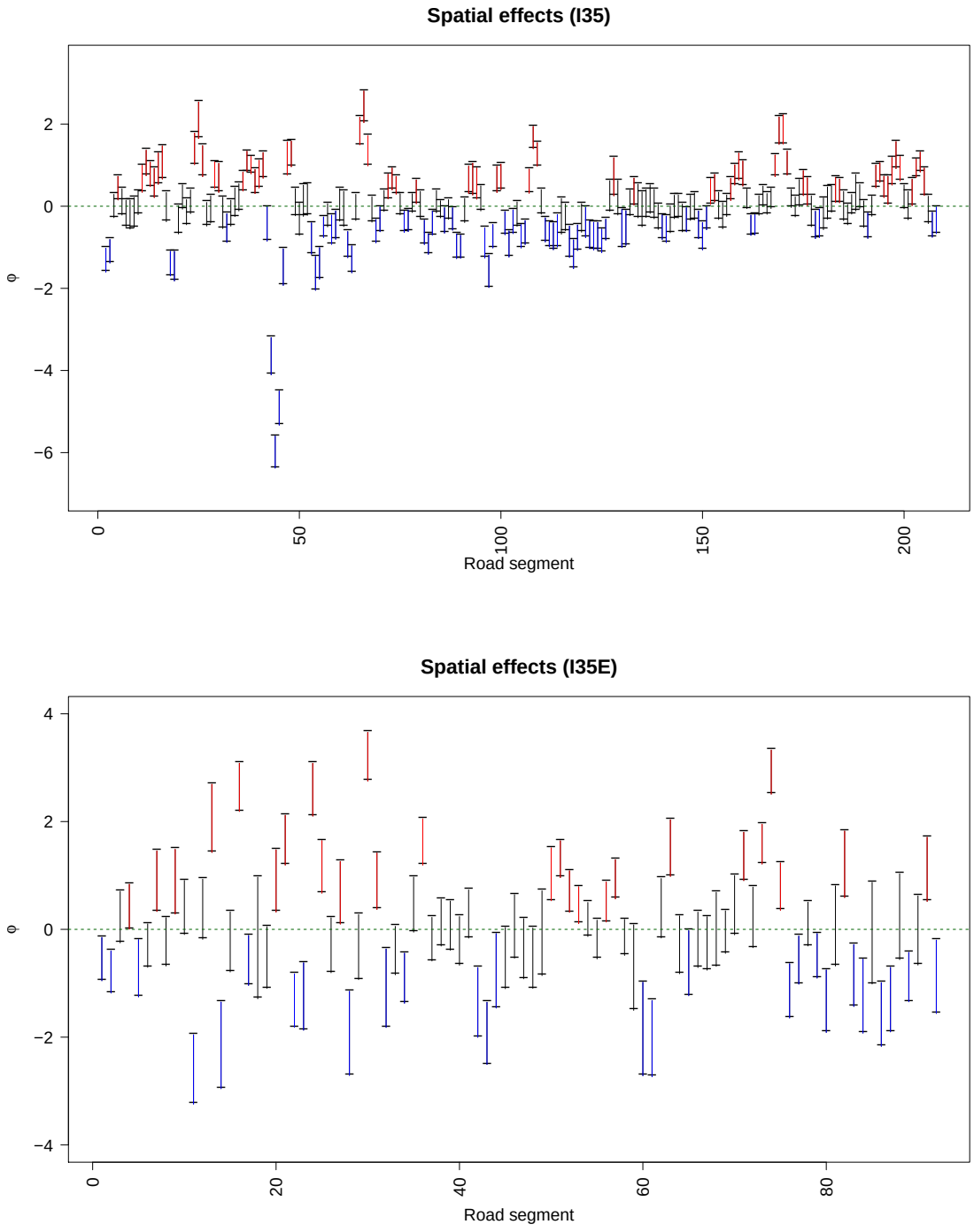

Figure 4.4: 95% credible intervals for the spatial effects. Blue lines indicate the intervals fall below zero while red lines indicate intervals are above zero.

CHAPTER 5

CONCLUSION

The goal of this analysis was to (1) identify road characteristics that lead in an increased risk of accident and ultimately to (2) identify road segments that have more accidents than is expected given the traffic levels and road characteristics. Through fitting a Poisson regression, we were able to identify which road characteristics are most strongly associated with increased accident risk. By including a spatial component, we were able to account for both spatial correlations and unmeasured effects on road segments. These unmeasured effects allowed us to identify road segments with higher accidents than expected given the traffic levels and road characteristics.

The spatial effects make it clear that the road characteristics alone do not contain all the information on number of accidents in a given road segment. There may be one or several road characteristics that haven't been included in the model that should be. One puzzling result seen above is that the effect of road characteristics differs drastically from road to road. We hypothesize that this may be due to the fact that I-35E is a purely metropolitan interstate whereas I-35 spans urban and rural areas. This suggests that, when considering systemic improvements, the type of improvement needed to prevent crashes depends on the type of interstate.

While this model presents a good start to identify road characteristics, the modeling can be improved. Specifically, suggested next steps in risk factor analysis include, first, collect and use more covariates (road characteristics) in the model. From the results seen in Chapter 4, the spatial random effects are clearly capturing unobserved covariate data. Particularly, including variables such as road curvature, grade (steepness), and whether or not an entrance or exit is present are important variables not included in this analysis (or the

HSIS database) that affect crash risk. And, second, include cross-road spatial correlations. This analysis considered one road segment at a time. However, I-35 and I-35E connect near the Twin Cities area. Instead of making inference on one road at a time, including other roads which connect to or are close to the road of interest may give more information on potential risk and preventative factors.

BIBLIOGRAPHY

Banerjee, S., Carlin, B., and Gelfand, A. (2015), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, Florida: CRC Press.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Roal Statistical Society*, 36, 192–236.

Casella, G., and Park, T. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103.

Chib, S., and Greenburg, E. (1995), "Understanding the Metropolis Hastings Algorithm," *The American Statistician.*

Cressie, N., and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Hoboken, New Jersey: John Wiley & Sons, Inc.

Faraway, J. (2006), *Extending the Linear Model with R* (1st ed.), Boca Rotan, Florida: Chapman & Hall CRC.

Gamerman, D., and Lopes, H. F. (2006), *Markov Chain Monte Carlo* (2nd ed.), Boca Raton, Florida: Chapman & Hall/CRC.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An Adaptive Metropolis Algorithm," *Bernoulli*, 7, 223–242.

Hastings, W. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

He, Z., and Sun, D. (2000), "Hierarchical Bayes Estimation of Hunting Success Rates with Spatial Correlations," *Biometrics*, 360–367.

Hughes, J., and Haran, M. (2012), "Dimension Reduction and Alleviation of Confounding for Spatial Generalized Linear Mixed Models," *Journal of the Royal Statistical Society*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006), "Fixed-Width Output Analysis for Markov Chain Monte Carlo," *Jornal of the American Statistical Association*.

Lawson, A. B. (2013), *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology* (2nd ed.), CRC Press.

MacNab, Y. C., and Dean, C. B. (2000), "Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models," *Statistics in Medicine*, 19.

Metropolis, N., Rosenbbluth, A., Rosenbluth, M., Teller, A. H., and Teller, E. (1953), "Equation of state calculations by fast computing machine," *Journal of Chemical Physics*, 21, 1087–91.

Pace, R. K., Barry, R., Gilley, O. W., and Sirmans, C. (2000), "A method for spatial-temporal forecasting with an application to real estate prices," *International Journal of Forecasting*, 229–246.

Robert, C. P., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed.), New York, New York: Springer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Line, A. (2002), "Bayesian measures of model complexity and fit," *J.R. Statistics*.

Waller, L., Carlin, B., Xia, H., and Gelfand, A. (1997), "Hierarchical Spatio-Temporal Mapping of Disease Rates," *Journal of the American Statistical Association*, 92, 607–617.

Weisberg, S. (2005), *Applied Linear Regression* (3rd ed.), Hoboken, New Jersey: John Wiley & Sons, Inc.

Zhou, J., Change, H. H., and Fuentes, M. (2012), "Estimating the Health Impact of Climate Change With Calibrated Climate Model Output," *Journal of Agricultural, Biological, and Environmental Statistics*.