2012-06-26

An Applied Investigation of Gaussian Markov Random Fields

Jessica Lyn Olsen
Brigham Young University - Provo

An Applied Investigation of Gaussian Markov Random Fields

Jessica Lyn Olsen

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

Master of Science

Shannon Neeley, Chair
Candace Berrett
Bruce Collings

Department of Statistics

Brigham Young University

August 2012

ABSTRACT

An Applied Investigation of Gaussian
Markov Random Fields

Jessica Lyn Olsen
Department of Statistics, BYU
Master of Science

Recently, Bayesian methods have become the essence of modern statistics, specifically, the ability to incorporate hierarchical models. In particular, correlated data, such as the data found in spatial and temporal applications, have benefited greatly from the development and application of Bayesian statistics. One particular application of Bayesian modeling is Gaussian Markov Random Fields. These methods have proven to be very useful in providing a framework for correlated data. I will demonstrate the power of GMRFs by applying this method to two sets of data; a set of temporal data involving car accidents in the UK and a set of spatial data involving Provo area apartment complexes. For the first set of data, I will examine how including a seatbelt covariate effects our estimates for the number of car accidents. In the second set of data, we will scrutinize the effect of BYU approval on apartment complexes. In both applications we will investigate Laplacian approximations when normal distribution assumptions do not hold.

Keywords: Gaussian Markov Random Fields, Spatial, Correlated Data

ACKNOWLEDGMENTS

First and foremost, I need to thank my advisor, Dr. Shannon Neeley. She has been beyond understanding and patient, and has gone above and beyond the call of Professor. This project would not have been possible without her, and I imagine that the project would not have been as successful under the tutelage of any other professor. This specific project is the culmination of my entire educational experience at Brigham Young University. With that, I must also thank my professors, the Statistics Department staff, my classmates, and my family. Successful people are often the product of supportive environments, and I have been fortunate enough to be surrounded by a fantastic team.

CONTENTS

CHAPTER 1

INTRODUCTION

Although the discipline of statistics has long been employed in aiding and supporting researchers and scientists solve complex problems, there seems to be no foreseeable end to the areas in which statistics can be beneficial and valuable. Karl Pearson, esteemed mathematician and founder of the worlds first department of statistics, once noted that, "Statistics is the grammar of science" (Pearson, 1911). With that, we can argue that statistics is the means by which we are able to transform the study of natural phenomena into not only the understanding of natural phenomena, but also the communication of the why, where, how, and when of scientific events of interest. As a result of the universal need and ubiquitous demand for explanation and exploration of the world around us, the field of statistics will continue to grow and expand, in areas that have long been established as well as disciplines that have yet to be discovered.

Appealing to a very economic sense of creation driven by demand, statistics is an ever evolving and developing field of study. As new problems arise, new methodology and techniques for dealing with those problems are developed, tested, and scrutinized. Spatial statistics is an exemplary case of this principle. Although evidence of statistics for spatial data can be traced back to very simple weather data maps from the 17th century, true statistical spatial models did not appear until much later. More recently, hierarchical modeling via the Bayesian paradigm has permitted scientists and researchers to analyze increasingly complex systems of correlated spatial data. One such method is Gaussian Markov Random Fields, a method that focuses on defining covariance structures via precision matrices.

The ultimate aim of this investigation is to understand Gaussian Markov Random Fields; how they can be measured, quantified and utilized. After an in depth investigation into the theory surrounding classical prediction methods, and hierarchical Bayesian methods, we will demonstrate the powerful nature of Gaussian Markov Random fields via temporal and spatial data applications. Through this inquiry we will gain a better idea of how Gaussian Markov Random Fields behave and determine efficient techniques for estimation of specific data types.

CHAPTER 2

CLASSICAL APPROACH TO SPATIAL STATISTICS

2.1 Brief Introduction to the General Spatial Model and Kriging

Generally speaking, kriging is a member of the least-squares class of statistical methods. Ultimately, kriging attempts to find optimal interpolation of unobserved quantities of a function, $Z(\cdot)$ using regression on some collection of the weighted observed $Z(\cdot)$ values. The method is named after D.G. Krige, a mining engineer from South Africa, who first used a crude method to determine actual ore-grade behavior based on sample ore grades (Oliver, 2010). French mathematician, George Matheron, formally developed the theoretical foundation for interpolation and extrapolation based on Krige's work (Waller and Gotway, 2004). Methods have since been developed to relax model assumptions and improve model fit.

2.2 Defining the Spatial Model

We begin this section by making appropriate definitions and explanations of the general spatial model. Doing so will allow us to establish the language for kriging. The notation used here is borrowed from Cressie (Cressie, 1993). This conventional presentation is flexible enough to be applied to a wide set of problems and purposes. Let $\mathbf{s} \in \mathbb{R}^d$ be a generic data location in d-dimensional Euclidean space. We will define $\mathbf{Z}(\mathbf{s})$ to be a random data point measured at spatial location $\mathbf{s}$. We can then allow $\mathbf{s}$ to vary over the indexed set $D \subset \mathbb{R}^d$. Doing so results in a generation of the multivariate random process

$$\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D\}.$$

Realizations from this random process are represented by $\{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$. While randomness is not strictly assumed when defining the indices in D or the process Z, in practical applications, these quantities are usually fixed. This notation can be used to describe a number of spatial areas of interest including geostatistical data, lattice data, and point patterns. While much of the statistical approach focuses on developing models and methodologies, it is equally important to allot time and attention to the outcomes that a particular model produces. In spatial statistics, this concept translates into building models where prediction is the desired outcome. It is imperative to create methodological frameworks that accurately portray the dependence structures that inherently exist in spatial data. To continue, we must further define and expand the spatial model.

2.3 Variogram Structures

Any discussion of spatial models is not complete without a complete understanding of the correlation structure. As such, we will investigate the variogram, the covariogram, and the correlelogram; the three standard measures used in setting up spatial dependencies. These quantities become critically important in laying the foundation for the assumptions surrounding the random process, $Z(\cdot)$. Thus, understanding these terms is paramount to the framework for kriging.

The variogram is the function that describes the strength and form of the relationship between points. In essence, the variogram defines the amount of spatial dependency in a model. I will further define

$$Var(Z(\mathbf{s_1}) - Z(\mathbf{s_2})) = 2\gamma(\mathbf{s_1} - \mathbf{s_2}),$$

where $2\gamma(\cdot)$ is called the variogram and $\gamma(\cdot)$ is called the semivariogram. The valuation of this function is directly related to assumptions concerning the stationarity of the function

$Z(\cdot)$. Stationarity refers to how the $Z(\cdot)$ process changes as we move from $\mathbf{s_i}$ to $\mathbf{s_j}$ in our defined space. Strict stationarity requires that the joint distribution of the $Z(\cdot)$ process remain invariable from one location to another. This is a very strong, and sometimes unrealistic assumption. Even the most basic methods of kriging do require strict stationarity. Rather, kriging generally requires one of the following forms:

Intrinsic Stationarity: The weakest form of stationary, Intrinsic Stationarity requires two main conditions:

$$E[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})] = 0,$$
$$Var[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})] = 2\gamma(\mathbf{h}).$$

where the variogram, $2\gamma(\mathbf{h})$, is known.

Second-Order Stationarity: Also known as wide-sense stationarity, requires the following conditions:

$$E[Z(\mathbf{s})] = \mu,$$
$$Cov[Z(\mathbf{s_1}), Z(\mathbf{s_2})] = C(s_1 - s_2).$$

The quantity $C(\cdot)$, called the covariogram, is a function of direction and distance between points. Unlike intrinsic stationarity, which is heavily dependent on $\mathbf{s}$ and $\mathbf{h}$ for defining the first moment of the process, second-order stationarity restricts the first moment of the process to be a finite constant. In both instances, the covariance structure (expressed through the variogram or the covariogram) depend only on $\mathbf{h}$, and not $\mathbf{s}$. Under the assumption of second-order stationarity we can show that

$$2\gamma(\mathbf{h}) = 2[C(\mathbf{0}) - C(\mathbf{h})],$$

which is a convenient relationship between the variogram and the covariogram.

Finally, it is convenient to define dependencies in terms of correlations. Under the provision that $C(\mathbf{0}) > 0$, we can define the correlelogram to be

$$\rho(\mathbf{h}) \equiv C(\mathbf{h})/C(\mathbf{0}).$$

The manner in which the process behaves with respect to spatial dependency is crucial to the model. As such, estimation of dependency is a subject that has received a lot of attention in spatial statistics. Different assumptions of stationarity and sample sizes suggest different optimal methods for estimating this quantity. Cressie (Cressie, 1993) makes the argument that it is nearly always preferable to estimate the variogram rather than the covariogram since the covariogram is not always an identifiable parameter in a model. French geologist Georges Matheron developed the first procedure for estimating variograms by using the Method of Moments techniques (Piegorsch and Bailer, 2005). When the random process $Z(\cdot)$ is intrinsically stationary, this quantity is unbiased for the variogram. However, if the process is truly second-order stationary, this measure has bias of $O(1/n)$, which may present issues with small sample sizes.

In addition to Matheron's original approach to the problem, other statisticians have suggested more robust measures of the variograms. Cressie and Hawkins (1980) use fourth roots of squared differences and incorporate corrections for biased behavior to provide robust estimators of the variogram. Building upon the work of Cressie and Hawkins, Armstrong and Delfiner (1980) have proposed scale and quantile estimators. More recently, Dan Nord-

man has proposed an empirical likelihood method for estimation.

In many cases, we are more concerned with the "noiseless" form of the random process, $S(\mathbf{s})$, where

$$Z(\mathbf{s}) = S(\mathbf{s}) + \epsilon(\mathbf{s}) \quad \mathbf{s} \in D,$$

and $\epsilon(\cdot)$ is a white-noise measurement-error process. The field of spatial prediction is primarily interested in predicting a function of either $Z(\cdot)$ or $S(\cdot)$ using the information that we actually observe. We can further decompose $S(\cdot)$ into the sum of a deterministic mean structure and two zero-mean intrinsically stationary processes. This gives

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}) + \eta(\mathbf{s}) + \epsilon(\mathbf{s}), \ \ \mathbf{s} \in D,$$

where:

$\mu(\cdot) \equiv E[Z(\cdot)]$ denotes the deterministic mean structure (also called large-scale variation),
$W(\cdot)$ is a zero mean intrinsically stationary process representing the variogram of the smooth process, $S(\cdot)$, and
$\eta(\cdot)$ is an additional zero mean intrinsically stationary process representing the variogram capturing microscale variation.

2.4 Different Kriging Methods

Much of the differentiation between kriging methods relies on how we define $\mu(\mathbf{s})$, the deterministic mean structure. Keeping in mind that each of the following methods assumes

wide sense stationarity (i.e., strict or intrinsic stationarity) of the field, several methods are summarized below.

Simple

- The mean structure, $\mu(\mathbf{s})$, is assumed to be 0 everywhere
- Known variogram

Ordinary

- Assumes unknown, but constant mean structure, $\mu(\mathbf{s})$
- Unknown variogram, but must have sufficient data to estimate the function

Universal

- Assumes unknown and inconstant mean structure, $\mu(\mathbf{s})$ (assumes a genearl linear trend model)
- Unknown variogram, but must have sufficient data to estimate the function

2.5 Types of Spatial Data

Spatial data is generally classified into several broad categories. I now provide a short summary of two of these categories, point-reference data and areal data, where kriging methods are most typically employed.

Point-Reference Data

Often referred to as geostatistical data or point-level data, point-referenced data allows the location of index $\mathbf{s}$ to vary randomly and continuously over fixed subset D. In this type

of data structure, we observe r points within our subset D, and allow correlation between points to vary by some measure of distance between the points (Banjerjree, et. al, 2004).

Areal Data

Areal Data, also called lattice data, also assumes that D is some fixed subset, but introduces spatial correlation through the notion of blocks; the r points (as with geostatistical data) are grouped together into blocks. These blocks define neighborhood structures that are used to characterize semivariogram functions (Banjerjree, et. al, 2004). Although blocks can be defined a number of different ways, blocks often have real world interpretations such as zip codes, counties, etc.

My project proposal concerns the second data type, areal data. Although classical statistical methods are satisfactory for many applications, I will investigate whether or not introducing the Bayesian paradigm through Gaussian Markov Random Fields can improve model fit and spatial prediction. In order to properly execute this task, we must first understand the theory that supports Gaussian Markov Random Fields, a brief introduction to which is found in the next chapter.

CHAPTER 3

HIERARCHICAL SPATIAL PREDICTION METHODS: GAUSSIAN MARKOV RANDOM FIELDS

Although Bayes rule, which governs the Bayesian paradigm, has been established for centuries, use of a fully Bayesian model has only become possible more recently with the advent of faster and more efficient computing power. Bayesian models allow us to incorporate antecedent beliefs about parameter behavior, as well as incorporate and account for uncertainty in our beliefs. Frequentist perspectives offer very little leeway in this respect, and force a single layered approach to a problem. Complex hierarchical modeling, which is arguably the core of modern statistics, can be achieved through adoption of the Bayesian paradigm.

The aim of the following section is to explore the properties of a specific type of hierarchical modeling, Gaussian Markov Random Fields (GMRFs). I will first discuss some of the theory and basic properties of GMRFs and then discuss a particular application to temporal modeling. Our goal is to understand what GMRF's are, why they work, and how they contribute to the expanding world of statistical techniques and methodology.

3.1 Brief Lesson on Gaussian Markov Random Fields

Theory Behind Gaussian Markov Random Fields

Although specific implementation of Gaussian Markov Random Fields can prove to be quite complex, the basic concept of a GMRF is quite simple. A GMRF is simply a random vector (Random Field) of finite-dimension that follows a multivariate normal distribution (Gaussian). In the most relaxed and basic settings, we assume independence of elements in a random vector, however, practical application restricts the GMRF to satisfy additional conditional independence assumptions (Markov). A more formal mathematical definition is

given by

GMRF:*A random vector* $\mathbf{x} = (x_i, ..., x_n)^T \in \mathbb{R}^n$ *is called a GMRF with respect to a graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *with mean* μ *and precision matrix* $\mathbf{Q} > 0$*, iff its density has the form*

$$\pi(\mathbf{x}) = (2\pi)^{-n/2}|Q|^{1/2} exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{Q}(\mathbf{x}-\mu)\right)$$

where $\mathcal{V}$ *and* $\mathcal{E}$ *are the set of nodes in the graph, and the set of edges in the graph, respectively.*

In laymen's terms, a GMRF governs connected points (nodes) are within a well-defined (by edges) set are connected. These nodes and areas typically represent movements in space, time, or space-time. We will assume that these relationships can be modeled by a multivariate normal distribution. Under this construct, we must further define parameterization for the multivariate normal distribution, namely a mean and covariance. For this, we turn to the properties of a sparse precision matrix (instead of explicitly defining a covariance matrix), and theorems regarding conditional expectations. Namely,

$$x_i \perp x_j | \mathbf{x_{-ij}} \Leftrightarrow Q_{ij} = 0.$$

A sparse precision matrix greatly increases computational efficiencies. Conditional Markov results prove to be important and useful. If we partition a set of data into A and B, and in turn, condition A upon B the following result holds true.

Let $\mathbf{x} \sim N_C(\mathbf{b}, \mathbf{Q})$*, then*

$$\mathbf{x_A}|\mathbf{x_B} \sim N_C(\mathbf{b_A} - \mathbf{Q_{AB}}x_B, \mathbf{Q_{AA}})$$

Later on we will define $\mathbf{x}$ and $\mathbf{x_A}|\mathbf{x_B}$ to be Gaussian Markov Random Fields. These are presented in canonical form, where $N(\mu, \mathbf{Q^{-1}}) = N_C(\mathbf{Q}\mu, \mathbf{Q})$. Most notable from this parameterization is the fact that we can successively update without explicitly computing the mean until we actually need it! While this is certainly a very short course on what GMRFs, we now have enough of the general theory behind GMRFs to implement these techniques in data analysis.

3.2 General Areas of Applications of GMRF's

Gaussian Markov Random Fields offer a very general and flexible framework correlated data, making them easily applicable to a number of different hierarchical modeling structures. Autoregressive models, which are used extensively in structural time-series analysis, are GMRFs on a linear space. Gaussian Markov Random Fields have also proven to be very useful in longitudinal and survival data (Rue and Held, 2005). Not surprisingly, GMRFs are used extensively in spatial statistics. More recently, researchers have been exploring the properties of using GMRFs on spatial binary data. Graphical models and image analysis have also made wide use of GMRF techniques.

3.3 Estimation of GMRFs with Laplacian Approximations

Although specifying correlation through precision matrices within the construct of Gaussian Mark Random Fields can be computationally convenient, high-dimension problems can present issues. As such, a method which approximates posterior distributions has been proposed. Rather than directly sampling from posterior distributions, we can estimate these distributions. Consider the marginal posterior density of parameter θ_i with hyperpriors $\pi(\lambda)$ as

$$\pi(\theta_i|y) = \int_{\lambda} \pi(\theta_i|\lambda, y)\pi(\lambda|y)d\theta.$$

Integrated Nested Laplacian Approximation (INLA) uses an iteratively weighted average of the i^{th} component of the posterior parameters. Namely

$$\tilde{\pi}(\theta_i|y) = \sum_k \tilde{\pi}(\theta_i|\Lambda_k, y)\tilde{\pi}(\theta_k|y)\Delta_k.$$

Estimating $\tilde{\pi}(\lambda|y)$ also requires some theoretical background. We begin with the strict definition of $\pi(\theta, \lambda, y)$ as

$$\pi(\theta, \lambda, y) = \pi(\theta|\lambda, y)\pi(\lambda|y)\pi(y),$$

which implies

$$\pi(\lambda|y) \propto \frac{\pi(\theta, \lambda, y)}{\pi(\theta|\lambda, y)}.$$

Since we already have explicitly information about the numerator, our task is to estimate the denominator. Tierney and Kadane (1986) proposed the Laplacian approximation, for which INLA get its name. Further work by Havard and Rue (2009) suggested a Gaussian approximation. Finally, Azzalini and Capitano (1999) suggested addition of a cubic spline for improved fit.

Held, Rue, and Schrodle (2009) explored comparisons of MCMC and INLA Bayesian inference using leave-one-out cross-validatory predictive checks. They found that this measure established INLA as comparable to actual MCMC draws.

CHAPTER 4

DATA APPLICATION: UK DRIVER DATA

4.1 Data Description

Our data represent the monthly totals of car drivers in Great Britain killed or seriously injured from January 1969 to December 1984. Count data traditionally do not exhibit normal properties, however, a square root transformation offers acceptable approximations to a Gaussian distribution. We assume that the data are conditionally independent normal variables, or more formally,

$$y_i \sim N(s_i + t_i, 1/\kappa_y).$$

The mean is represented as the combination of a smoothing trend and a seasonal effect, while the variance is defined by precision parameter κ. The smoothing trend, t_i, follows a second-order Random Walk model, and the season effect, s_i, follows a seasonal model. Methods, to be discussed later, will allow us to side step specific forms of the distributions of the seasonal and smoothing trends.

An initial plot of the data, seen in Figure 4.1, shows a seasonal pattern. Although not as obvious as the seasonal variation, visual inspection also identifies a slight downward trend. The black horizontal line represents the point in time at which compulsory seatbelt laws were enforced. This significantly reduced the number of deaths and serious accidents. Hopefully, we will capture these separate effects via our specification of $\mathbf{s}$ and $\mathbf{t}$. However, we must make two important observations. First, we assume that $\mathbf{s}$ and $\mathbf{t}$ are *a priori* independent. Secondly, we assume that these models are both governed by respective precision parameters,

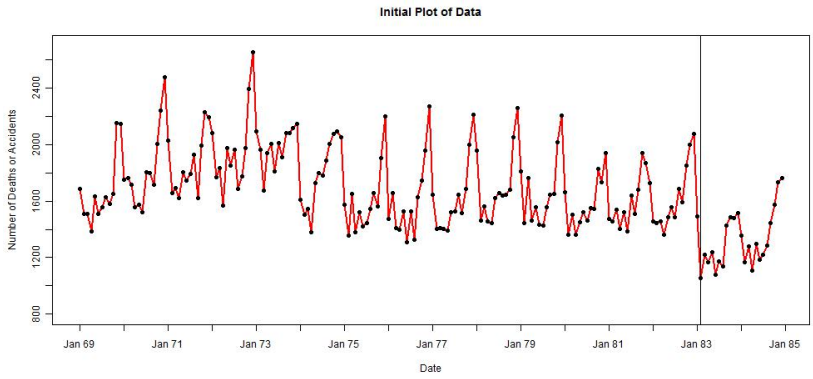

Figure 4.1: Initial Data Plot

denoted as κ. Therefore, our parameter vector has five elements: $\kappa_y, \kappa_s, \kappa_t, \mathbf{s}, \mathbf{t}$. Ultimately, this parameter vector is our primary object of inference.

Although we will eventually do inference on $(\mathbf{s}, \mathbf{t}, \kappa)$, our analysis begins with deriving the joint density

$$\pi(\mathbf{s}, \mathbf{t}, \mathbf{y}|\kappa) = \pi(\mathbf{y}|\mathbf{s}, \mathbf{t}, \kappa)\pi(\mathbf{s}|\kappa_s)\pi(\mathbf{t}|\kappa_t),$$

which is a GMRF with precision matrix, $\mathbf{Q}$. If we condition this distribution on $\mathbf{y}$, we will arive at one density of interest, $\pi(\mathbf{s}, \mathbf{t}|\mathbf{y}, \kappa)$. In order to do this we write the likelihood as

$$\begin{aligned} \pi(\mathbf{s}, \mathbf{t}|\mathbf{y}, \kappa) &= \prod_{i=1}^{n} \left(\frac{\kappa_y}{2\pi}\right)^{-.5} exp\left[-\frac{\kappa_y}{2}(y_i - s_i - t_i)^2\right] \\ &\propto exp \sum_{i=1}^{n} \left[-\frac{\kappa_y}{2}(y_i - s_i - t_i)^2\right]. \end{aligned}$$

If we multiply out the squared term in the exponent, we see that the likelihood induces pairwise independence between y_i, s_i, and t_i. To account for this, we partition the precision matrix $\mathbf{Q}$ into

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q_{ss}} & \mathbf{Q_{st}} & \mathbf{Q_{sy}} \\ \mathbf{Q_{ts}} & \mathbf{Q_{tt}} & \mathbf{Q_{ty}} \\ \mathbf{Q_{ys}} & \mathbf{Q_{yt}} & \mathbf{Q_{yy}} \end{pmatrix}.$$

We can now derive $\pi(\mathbf{s}, \mathbf{t}|\mathbf{y}, \kappa)$ by using theorems regarding conditional expectations, which were discussed earlier. The distribution that results is bivariate normal distribution with canonical parameterization

$$\begin{pmatrix} \mathbf{s} \\ \mathbf{t} \end{pmatrix} = N \left[- \begin{pmatrix} \mathbf{Q_{sy}} \\ \mathbf{Q_{sy}} \end{pmatrix} \mathbf{y}, \begin{pmatrix} \mathbf{Q_{sy}} & \mathbf{Q_{st}} \\ \mathbf{Q_{ty}} & \mathbf{Q_{tt}} \end{pmatrix}. \right]$$

Our acceptance probability for use during MCMC successive sampling then becomes

$$\begin{aligned} \alpha &= min \left\{ 1, \frac{\pi(\mathbf{s}^*, \mathbf{t}^*, \kappa^*|\mathbf{y})}{\pi(\mathbf{s}, \mathbf{t}, \kappa|\mathbf{y})} \times \frac{\pi(\mathbf{s}, \mathbf{t}|\mathbf{y})}{\pi(\mathbf{s}^*, \mathbf{t}^*|\mathbf{y})} \right\} \\ &= min \left\{ 1, \frac{\pi(\kappa^*|\mathbf{y})}{\pi(\kappa|\mathbf{y})} \right\}. \end{aligned}$$

Since we are integrating out $(\mathbf{s}, \mathbf{t})$, only our posterior marginals and the normalizing constants from $\pi(s)$ and $\pi(t)$ are included in our acceptance probability. This greatly reduces computation. We use Gamma priors on all precision parameters, so the full conditionals become

$$[\kappa] \propto \kappa^{\alpha_\kappa - 1} exp(-b_\kappa \kappa) exp \left[-\frac{\kappa_y}{2} \sum_{i=1}^{n} (y_i - s_i - t_i)^2 \right] \left(\frac{\kappa_s}{2\pi} \right)^{1/2} \left(\frac{\kappa_s}{2\pi} \right)^{1/2}.$$

The distribution of κ can then be used to find the center of the bivariate normal distribution of $\pi(\mathbf{s}, \mathbf{t}|\mathbf{y}, \kappa)$.

4.2 Prior Specification

Now that the problem has been discussed, we can show how Integrated Nested Laplacian Approximation works in practice. First we will explicitly define our prior distributions for the κ vector. Following recommendations from Rue and Held (2005), each κ distribtuion will have a *Gamma* form where $E(\kappa) = b/a$. Prior parameter values are found in Table 4.1

Parameter	a	b
κ_y	4	4
κ_s	1	.1
κ_t	1	.005

Table 4.1: Prior Parameter Values

These values resulted in a practical prior predictive distribution throughout the time series, as shown in Figure 4.2.
The black line represents the mean value over a thousand iterations at each point in time. The dotted bands are 95% probability intervals.

Even with all of the computational niceties that result from the theorems regarding Gaussian Markov Random Fields, this problem still presents issues in actual evaluation; convergence was problematic and actual computation time was interminable. This is likely due to the high dimensionality of the problem (we are attempting to calculate several parameters per data point). Harvard and Rue suggest using a Laplacian Approximation as suggested in Schrodle and Held (2005) and described in Chapter 2.

Using Integrated Nested Laplacian Approximation (INLA) significantly reduced computation time, and made model comparison possible. We fit three models to the data. These models are

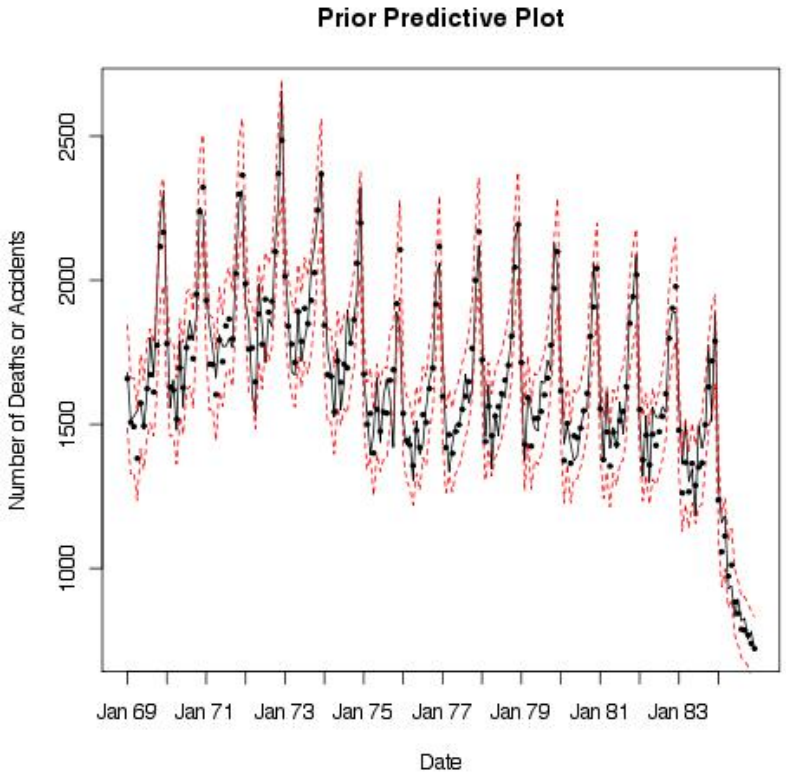

Figure 4.2: Initial Data Plot

Model 1: $y \sim Normal(s_i + t_i, 1/\kappa_y)$,

Model 2: $y \sim Normal(s_i + t_i + \beta_1 I(Belt = 1), 1/\kappa_y)$, and

Model 3: $y \sim Normal(s_i + t_i + \beta_1 I(Belt = 1) + g_i, 1/\kappa_y)$.

The models increase in complexity. The first model is the base model described earlier in this chatper. The second model adds a covariate for a seatbelt indicator fixed effect. The third model adds both a covariate for a seatbelt indicator and a gas price random effect. We use a flat, non informative prior for β_1 and allowed gas price to follow an RW(1) model. This model is written as

$$\begin{aligned} \pi(\mathbf{g}|\kappa_g) &\propto \kappa_g^{(n-1)/2} exp\left(-\frac{\kappa_g}{2}\sum_{i=1}^{n-1}(g_i - g_{i-1})^2\right) \\ \pi(\kappa_g) &\sim Gamma(1, .05). \end{aligned}$$

4.3 Model Assessment

I examined four different model assessment techniques. The first three, AIC, BIC, and DIC, are information criterion, which have different penalites to account for additional model complexity. Specifically,

Akaike Information Criterion = 2k - 2L,
Bayesian Information Criterion = k ln(n) - 2L, and
Deviance Information Criterion = $p_D + \bar{D}$

Table 4.2: Model Evaluation Criterion

where k is the number of parameters in the model, L is the log-likelihood, p_D is the number of effective parameters and $\bar{D}$ is the expected model deviance. The values for each measure applied to each model are given in Table 4.3.

	Model 1	Model 2	Model 3
AIC	771.92	753.13	752.88
BIC	879.46	861.83	860.93
DIC	682.81	670.30	670.16

Table 4.3: Summary of the Information Criterion Model Assessment

All of the criterion tell the same story: Models 2 and 3 are significantly better than Model 1, however, there is little difference between Models 2 and 3. This is confirmed in the fourth model assessment technique, Bayes Factors.

Bayes Factors are the Bayesian equivalent of the likelihood ratio test. However, instead of computing a test statistic and comparing to some asymptotic distribution, we compute values of K, where

$$K_{i,j} = \frac{P(Data|Model_i)}{P(Data|Model_j)}$$

and inherit K using the rules found in Table 4.4. Actual computed values of K can be found in Table 4.5.

Value of K	Decision
Less than 1	Support Model j
1-3	Barely worth mentioning
3-10	Substantial
10 - 30	Strong
30 - 100	Very strong
Greater than 100	Decisive

Table 4.4: Bayes Factors strength of evidence

$K_{2,1} = 17{,}188$
$K_{3,1} = 15{,}892$
$K_{2,3} = .92453$

Table 4.5: Bayes Factors for all possible model comparison

As we expected, Models 2 and 3 are clearly superior to Model 1, but the difference between Model 2 and 3 is marginal at best. This suggests that gas price is insignificant in our model.

4.4 Posterior Results: Model 2

We now present some of the interesting features of our INLA model choice. First we examine the two fixed effects, the overall intercept and the seatbelt indicator variable. The posterior distributions are given in Figure 4.3.

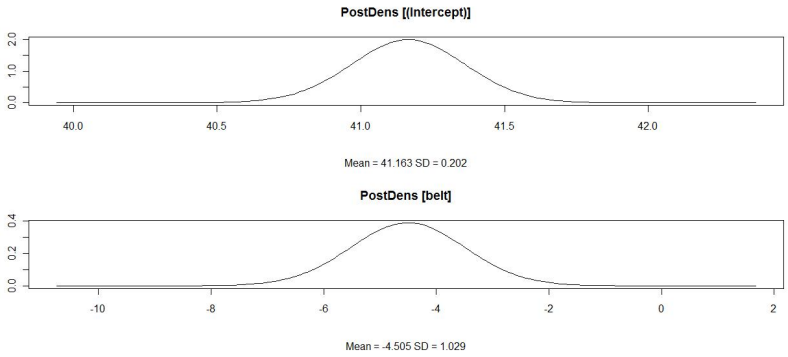

Figure 4.3: Posterior Distributions of Fixed Effects

We can examine the random effects, the seasonal trend and the smoothing trend. The posterior distributions are given in Figure 5.6. These distributions are of particular interest since they are specified through precision matrices.

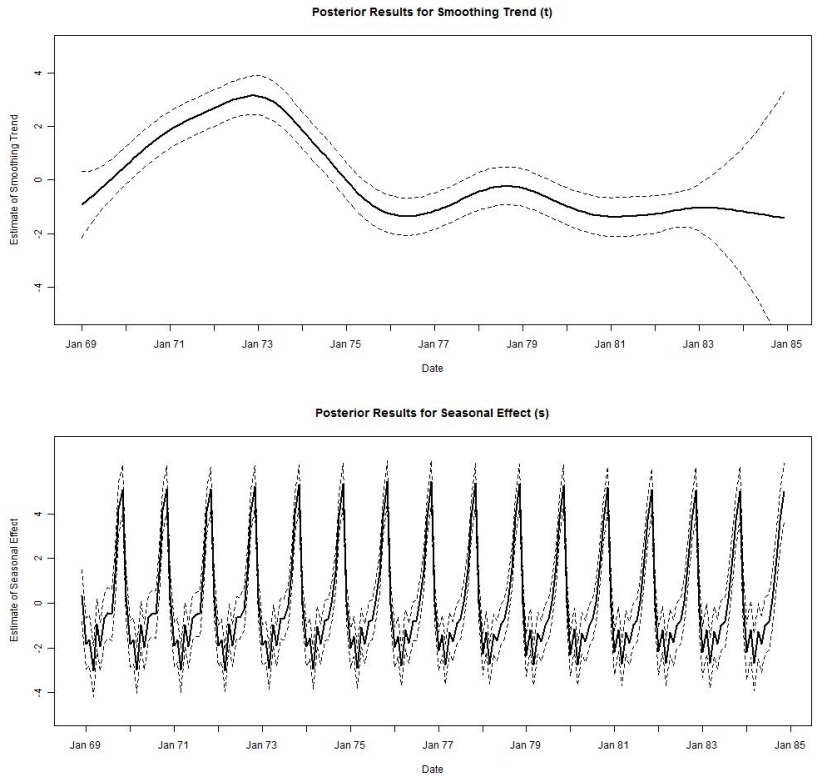

Figure 4.4: Posterior Distributions of Random Effects

The posterior distribution for the smoothing trend is not as smooth as might be expected. In fact, for the first several years of observations there seems to be a bit of a cycle in the smoothing trend itself. However, the downward trend is clearly visible. Also worth noting, is the fact that the trend effect seems to diminish as the time increases. The posterior distribution for the season effect appears as if there are peaks in winter and summer, with lulls in the spring and fall. Unlike the downward smoothing trend, the season variation remains relatively constant over time.

CHAPTER 5

DATA APPLICATION: PROVO APARTMENT DATA

Now that we have examined a temporal application of Integrated Nested Laplacian Approximation techniques, we will analyze a set of spatially correlated data. As a member of the Provo, UT community and a former student of Brigham Young University (BYU), I have always been strangely fascinated by the variables that determine rental prices. This investigation will probe at the heart of this very question.

5.1 Data Description

Our data set represent a number of characteristics and qualities of 118 apartment buildings from Utah County. This is an original data set, meaning it was collected and verified by myself. This involved a great deal of effort, including, but not limited to, emails, phone calls, searching county records, and collecting information from apartment directories maintained by various institutions. A description of the data collected pertaining to each apartment building can be found in the Table 5.1 through Table 5.4.

Informational Variables	
Variable Name	Description
Address	Physical Address of Building (Lat/Long)
Apartment Name	Name of the Apartment
Contact	Contact information (phone number, email, etc.)

Table 5.1: Informational Variables Collected

Spatial Variables		
Variable Name	Description	Details
Ward Membership	Used Address field to determine which LDS ward the apartment building belongs to	Within the data set, 23 unique wards are represented
Distance from BYU Campus	Used Address field to determine how far away the apartment is from BYU campus	Used the Spencer W. Kimball Tower as a point of reference

Table 5.2: Spatial Variables Collected

Other Variables		
Square Footage	Total Square footage for the apartment	Many apartment buildings had a varied square footage. This measure is an average of the high and low square footage given
Average Monthly Rent	Independent variable of interest	Many apartment buildings had a varied monthly rent. This measure is an average of the high and low monthly rent given
Year	Year the apartment building was originally constructed	
Students in Unit	The average number of students occupying a unit	

Table 5.3: Other Variables Collected

Indicator Variables	
Variable Name	Description
BYU Approved	Does the apartment building meet BYU approval standards
Power	Indicator variable for whether or not electricity is included in the rent
Trash	Indicator variable for whether or not trash collection is included in the rent
Gas	Indicator variable for whether or not gas is included in the rent
Internet	Indicator variable for whether or not internet is included in the rent
Sewer	Indicator variable for whether or not sewer is included in the rent
Water	Indicator variable for whether or not water is included in the rent
Furnished	Indicator variable for whether or not the apartment is furnished
Air Conditioning	Indicator variable for whether or not the apartment is air conditioned
Dishwasher	Indicator variable for whether or not the apartment has a dishwasher
Disposal	Indicator variable for whether or not the apartment has a garbage disposal

Indicator Variables (continued)	
Microwave	Indicator variable for whether or not the apartment has a microwave
Washer and Dyer	Indicator variable for whether or not the apartment has a washer and dryer
Laundry Hookups	Indicator variable for whether or not the apartment has laundry hookups
Cable	Indicator variable for whether or cable is included in the rent
Cable Hookups	Indicator variable for whether or not the apartment has a cable hookups
Internet Hookups	Indicator variable for whether or not the apartment has a internet hookups
On-sight Coin Laundry	Indicator variable for whether or not the apartment has an on-sight coin laundry
Recreation Area	Indicator variable for whether or not the apartment has a recreation area
Pool	Indicator variable for whether or not the apartment has a pool
Spa	Indicator variable for whether or not the apartment has a spa
Storage	Indicator variable for whether or not the apartment has a storage facility

Table 5.4: Indicator Variables Collected

This data represents a set of spatially correlated data which will be fit with a geoadditive, Bayesian semiparametric regression model. Using this data set, we will investigate the

properties of INLA in estimating the fixed and random effects represented by the variables detailed above. A regression analysis could provide potential rental guidelines for new and existing apartment buildings in the area, as well as offer insight into the effect of requiring students to live in BYU approved housing units.

5.2 Model Setup

Response Variable and Nonparametric Effects

We must first establish that we are working within the framework of a normal response model, where the conditional mean is a linear function of underlying unknown parameters. To help meet the conditions for normality, we investigated several transformations of our response variable, Average Monthly Rent. In accordance with the literature, we satisfy conditions for normality by modeling Rent per Square Foot (Rue and Held, 2005); thus,

$$y_i = \text{Rent per Square Foot.}$$

Let us define

$$\mathbf{z^S} = \text{Size of the apartment in Square Footage}$$

and

$$\mathbf{z^C} = \text{Year of Construction.}$$

We will incorporate a nonparametric modeling approach for the effect of these two covariates. Consequently, define vectors $\mathbf{s_S}$ and $\mathbf{s_C}$ as the ordered distinct covariate values of $\mathbf{z^S}$ and $\mathbf{z^C}$. Thus

$$\mathbf{s_S} = (s_1^S, ..., s_n^S),$$
$$\mathbf{s_C} = (s_1^C, ..., s_n^C),$$

where $\mathbf{s_S}$ values range from 500 to 3500 and $\mathbf{s_C}$ values range from 1962 to 2007. In accordance with a nonparametric approach, we will estimate a parameter at each value of $\mathbf{s_S}$ and

$\mathbf{s_C}$. These parameter values are denoted by vectors $\mathbf{x_S}$ and $\mathbf{x_C}$. For all locations of both $\mathbf{s_S}$ and $\mathbf{s_C}$, we assume a Random Walk of Order 1. Due to this fact, our parameters can be modeled as previously defined.

The precision parameter, κ, which defines the behavior of the GMRF for both square footage and year of construction parameters, is fit with a *LogGamma* distribution. We chose fairly diffuse priors, specifically

$$\kappa_S \sim LogGamma(1, .01), \text{ and}$$
$$\kappa_C \sim LogGamma(1, .01).$$

Geoadditive Effect

Spatial models require that we define a neighborhood structure. For purposes of this study, we define observations to be part of the same neighborhood if they belong to the same LDS church ward. Determining which wards were neighbors was determined more objectively, but is visually represented in Figure 5.1, showing a node graph and spatial matrix (where neighbors are represented by dark entries).

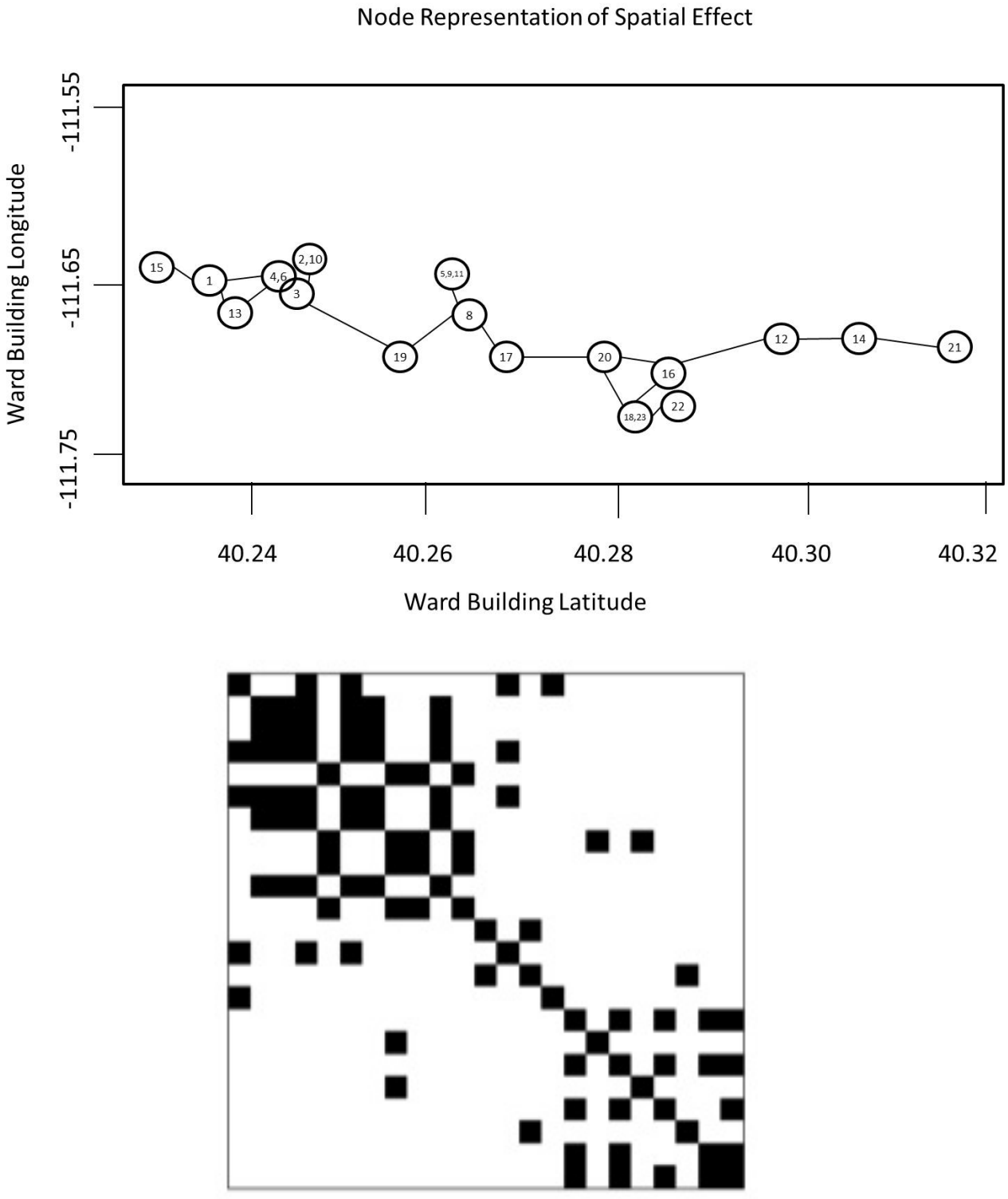

Figure 5.1: Node and Matrix Representation of Neighborhood Structure

Similarly to how we defined a corresponding vector of parameters for the nonparametric effects, we now define a vector of covariates for the location effect of each ward, $\mathbf{x_L}$. These effects are estimated with a spatial model where

$$x_i|x_j, i \neq j, \kappa \sim N\left(\frac{1}{n}\sum_{i\sim j} x_j, \frac{1}{n_j\kappa}\right)$$

where n_i is the number of neighbors of node i, and $i \sim j$ indicates that the two nodes i and j are neighbors. This model fits node i with a Normal distribution with mean equal to the average of all neighbors of node i, and variance equal to precision κ weighted by the number of neighbors of node i. The relationships defined in the node graph in Figure 5.1 are followed in the expression of the spatial model in practice. Again, we specify the parameter κ to be LogGamma with hyperparameters chosen to produce a fairly diffuse distribution. More precisely,

$$\kappa_L \sim LogGamma(1, .01)$$

Fixed Effects

Lastly, various indicator variables have been constructed to examine possible fixed effects. These variables are variables seven through thirty described in above table, and will be denoted as vector $\mathbf{z_i}$. Additionally, the continuous variables Distance from BYU Campus and Students in Each Unit, are assumed to behave linearly. These covariates will be denoted by vector β. Each element of β is assumed to Normally distributed with a prior mean of 0 and a prior precision of 1.

Full Model

Combining the random, nonparametric, and fixed effects results in a model specification that follows

$$y_i \sim N\left(\mu + x^S(i) + x^C(i) + x^L(i) + x^T(i)\beta, 1/\kappa_y\right)$$

Finally, the distribution for κ_y and will follow the same distribution as all other precisions. The full posterior distribution is

$$\begin{aligned}\pi(\mathbf{x}^S, \mathbf{x}^C, \mathbf{x}^L, \mu, \beta, \kappa_y|\mathbf{y}) \propto \pi(\mathbf{x}^S|\kappa_S)\pi(\mathbf{x}^L|\kappa_L)\pi(\mathbf{x}^C|\kappa_C)\pi(\mu) \\ \times\pi(\beta)\pi(\kappa_y)\pi(\mathbf{y}|\mathbf{x}^S, \mathbf{x}^C, \mathbf{x}^L, \mu, \beta, \kappa_y)\end{aligned}$$

We assume *a priori* conditional independence among our parameter vectors, however, once we condition upon the data, we introduce dependence, which is clearly seen by examining the exponential term of the posterior distribution,

$$exp\left(\sum_i (y_i - (\mu + x^S(i) + x^C(i) + x^L(i) + \mathbf{z}_i^T\beta))^2\right).$$

Each combination of parameter values introduces dependence, and therefore, the precision matrix will be nonsparse. We can make use of a subblock algorithm, introduced in Hue and Reld (2005) by grouping the precision matrix into four subblocks, namely

$$(\mathbf{x}^S, \kappa_S), \quad (\mathbf{x}^C, \kappa_C), \quad (\mathbf{x}^L, \kappa_L), \quad (\beta, \kappa_y, \mu).$$

Consequently, each block of the precision matrix updates one at a time and accepts or rejects κ and $\mathbf{x}$ parameter vectors jointly. Deriving the full precision matrix for $\mathbf{x}$ can be illustrated

by first introducing a set of data for prediction, $\widetilde{\mathbf{y}}$. Let us first examine $\pi(\mathbf{x}^{L,*}|all\ else)$ by defining

$$\widetilde{y}_i = y_i - (\mu + x^S(i) + x^C(i) + x^C(i) + \mathbf{z}_i^T\beta).$$

The full conditional of $\mathbf{x}^L$ can then be defined as

$$\begin{aligned}\pi(\mathbf{x}^L|all\ else) \propto \exp(-\frac{\kappa_L}{2}\sum_{i\sim j}(x_i^L - x_j^L)^2) \\ \times\ \exp(-\frac{\kappa_y}{2}\sum_{i\sim j}(\widetilde{y}_k - x^L(k))^2).\end{aligned}$$

Furthermore, let n_i represent the number of neighbors to location i and define

$$L(i) = \{k : x^L(k) = x_i^L\}$$

such that its size is $|L(I)|$. We can then specify the full conditional of $\mathbf{x}^L$ as a GMRF with canonical parameters $(\mathbf{b}, \mathbf{Q})$, where

$$b_i = \kappa_y \sum_{k\in L_i} \widetilde{y}_k$$

$$Q_{ij} = \begin{cases} \kappa_L n_i + \kappa_y|L(i)| & \text{if } i = j \\ -\kappa_L & \text{if } i \sim j \\ 0 & otherwise \end{cases}$$

We can go through a similar process to define all other necessary precision matrices.

Even with all of the computational niceties that result from the theorems regarding Gaussian Markov Random Fields, this problem still presents issues in actual evaluation;

convergence was problematic and actual computation time was interminable. This is most likely due to the high dimensionality of the problem (we are attempting to calculate several parameters per data point). Harvard and Rue suggest using a Laplacian Approximation as suggested in Schrodle and Held (2005) and described in Chapter 2 of this document. Using an Integrated Nested Laplacian Approximation (INLA) significantly reduced computation time, and made model comparison possible. As such, we went through an extensive process for fitting the data and selecting a model, which is described in the following section.

5.3 Model Selection

We built a script that employed a modified forward regression algorithm. My base model included only the nonparametric and geoadditive effects, Year of Construction, Square Footage, and Location. In addition to the Mean Absolute Percentage Error (MAPE) of the fitted values, for model comparison we used the same measurement criterion from the UK Driver data analysis, AIC, BIC, and DIC, which are defined in Table 4.2.

Starting with the base model, the algorithm then examined the possible addition of another variable out of all other remaining variables and then added the variable that resulted in the best fit according to the different criteria. While the order in which variables were added to the model was the same across all different measures of model fit, the optimal number of parameters was not. Figure 5.2 shows the relationship between model fit and number of variables for the different criterion.

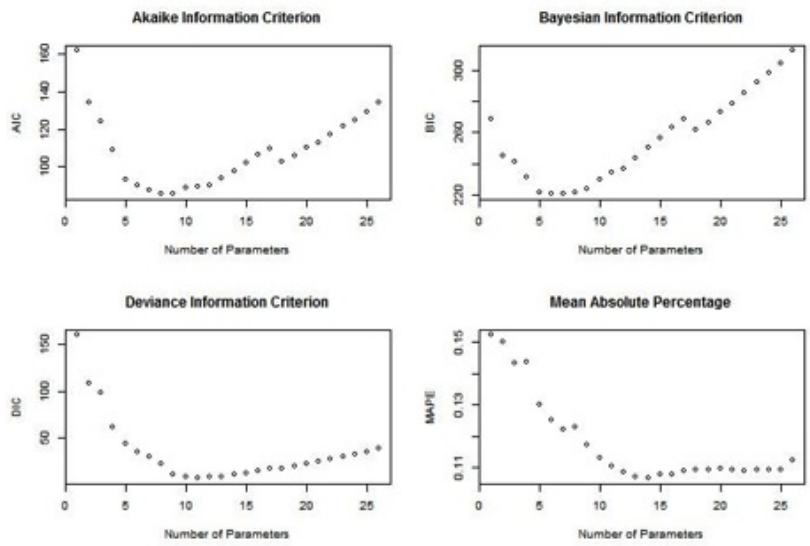

Figure 5.2: Information Criterion for Varying Levels of Covariates

The optimal number of parameters chosen by each criterion is shown in Table 5.5 and the variables chosen in each of the models can be found in Table 5.6.

Measurement Type	Number of Parameters
AIC	8
BIC	6
DIC	11
MAPE	14

Table 5.5: Number of Covariates to Include for each Information Criterion

1. Students in Each Unit	8. Laundry Hookups
2. Distance from BYU Campus	9. Internet
3. Furnished	10. Air conditioning
4. Washer & Dryer	11. On-sight Coin Laundry
5. Disposal	12. Wheelchair Accessible
6. Storage	13. Gas
7. Cable Hookups	14. Internet Access

Table 5.6: Covariates Included for each Model

To make a final decision between the most efficacious models chosen by the different techniques, I turned to Bayes Factor as defined in the previous chapter. As a reminder, Bayes Factor is the ratio of the likelihoods of two different models (called K) and uses Table 4.4 as guidelines in determining which model is best. Actual computed values of K can be found in Table 5.7, where the model labeled in the first row represents $Model_i$ and subsequent rows represent $Model_j$.

$Model_i$ = AIC	
BIC	3.48
DIC	0.52
MAPE	0.86

$Model_i$ = BIC	
AIC	0.29
DIC	0.15
MAPE	0.25

$Model_i$ = DIC	
AIC	1.93
BIC	6.73
MAPE	1.67

$Model_i$ = MAPE	
AIC	1.16
BIC	4.03
DIC	0.60

Table 5.7: All Bayes Factors Values

With a Bayes Factor of less than 1 for each model comparison, the best model chosen using BIC is clearly inferior to all other models. The comparisons of AIC, DIC, and MAPE models prove less straightfoward. The Bayes Factor for each pair of these models ranges in the 1

to 3, which according to Table 4.4 is barely worth mentioning. However, since DIC was the only model that was, at a minimum, supported in all comparisons, we will conclude that the model computed using DIC is the best.

5.4 Posterior Results: DIC Model

The final covariates included in the DIC model are given in Table 5.8. We now present some of the interesting findings of our INLA model of choice, beginning with the posterior results. First we will examine the fixed effects. The posterior distributions are given in Figure 5.3.

Fixed Effects Included	Random Effects Included	Spatial Effects Included
1. Students in Each Unit	Year of Construction	Ward Membership
2. Distance from BYU Campus	Square Footage	
3. Furnished		
4. Washer & Dryer		
5. Disposal		
6. Storage		
7. Cable Hookups		
8. Laundry Hookups		
9. Internet		
10. Air Conditioning		
11. On-sight Coin Laundry		

Table 5.8: Final Covariates Included

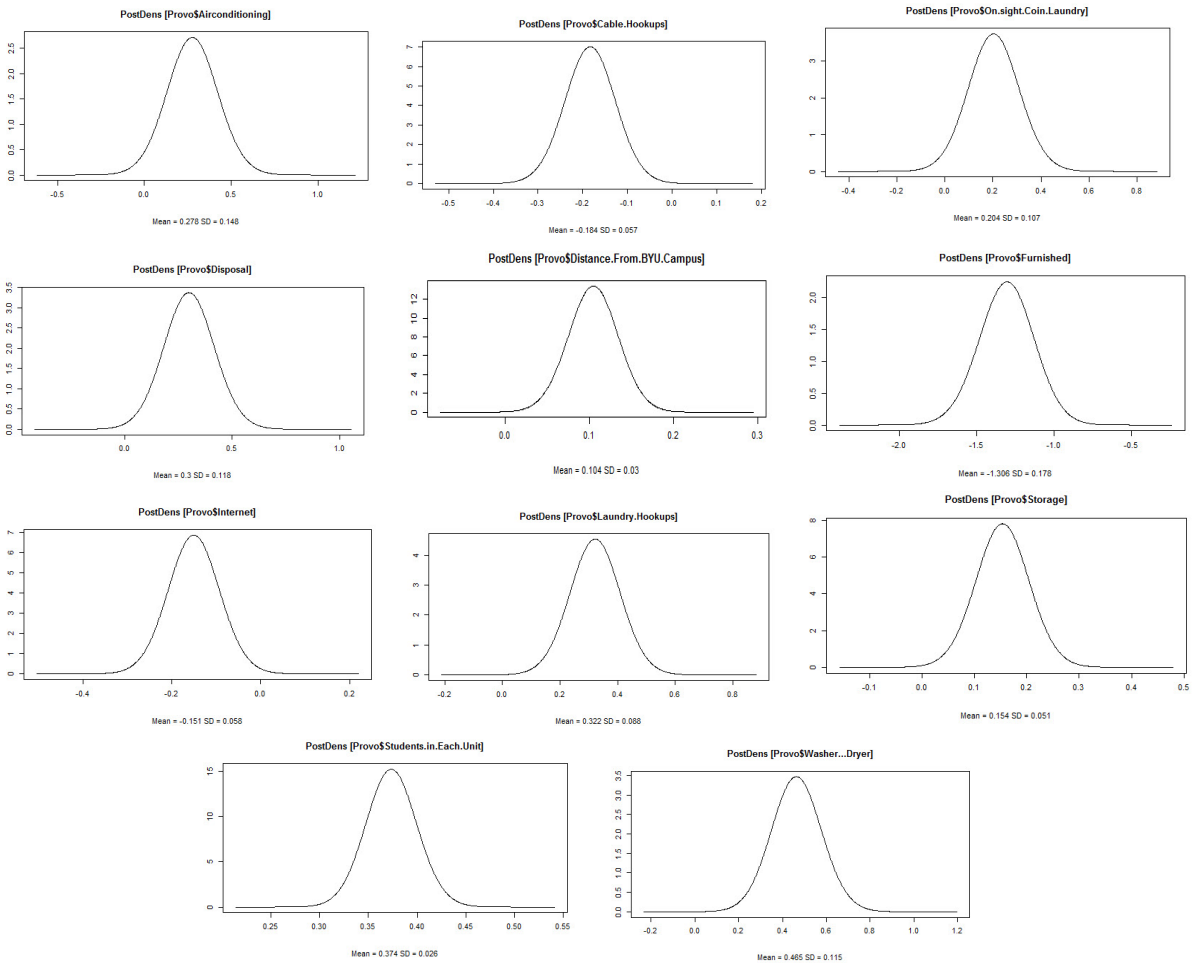

Figure 5.3: Posterior Distributions of Fixed Effects

Many of the posterior distributions contain 0 within their 95% Credibility Intervals, however there are several which do not, which we shall discuss further. The first interesting thing to note is the exclusion of the BYU Approved indicator variable. In the forward regression, this variable was not selected until step 19. This may indicate that it is insignificant in predicting or determining apartment rental prices in Utah County. I believe that this effect is being hidden by an interaction of the Distance from BYU variable (see in the second posterior distribution) and another variable which is not represented: apartment condition. I think a variable for the general condition of the apartment could provide some evidence that conditions in BYU approved apartments close to campus tend to be poorer than their

non-BYU approved counterparts.

One particularly interesting facet to note is the sign of the Furnished effect, see in Figure 5.4,which is not only negative, but also the largest effect in either direction.

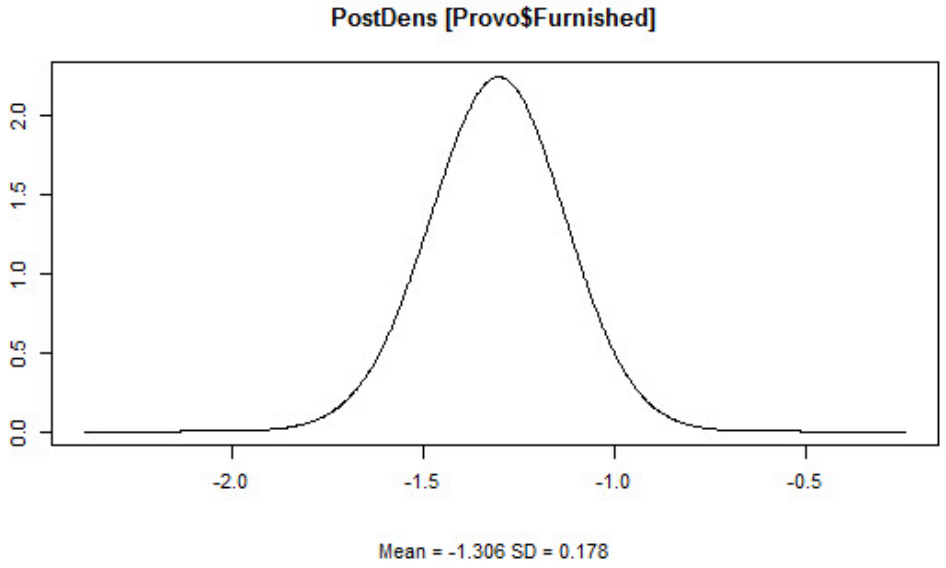

Figure 5.4: Posterior Distribution of Furnished Effect

With a mean of -1.306, this suggests that unfurnished apartments are significantly more expensive than furnished apartments. I would expect that the cost to the owner of furnishing an apartment would increase the cost of the rent charged on the apartment. This could also be related to the BYU approved variable: in our sample, 1.09% of the BYU approved apartments were unfurnished, as opposed to 15.38% of non-BYU approved apartments. There appears to be a relationship between furnished and non-furnished apartments and their BYU approved status that has not been reflected in our model.

The direct of the effect of the Internet indicator variable is also counterintuitive, with a posterior mean of -0.151.

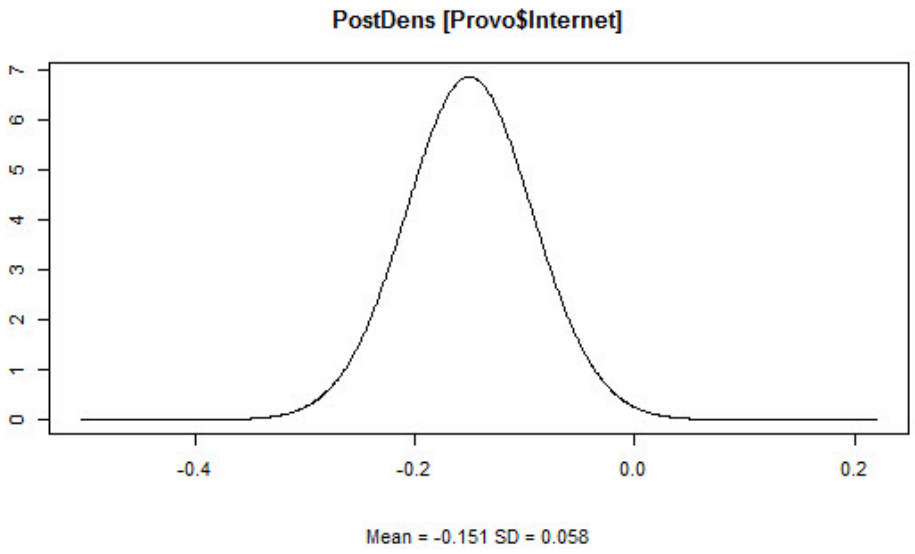

Figure 5.5: Posterior Distribution of Internet Effect

This implies the average monthly rent per square foot for apartments where internet access is included in the price of rent is $0.15 less than an apartment of the same condition that does not include internet access.

Now we can examine the random effects, the Year of Construction and Square footage. The posterior distributions are given in Figure 5.6. These distributions are of particular interest since they were specified through precision matrices.

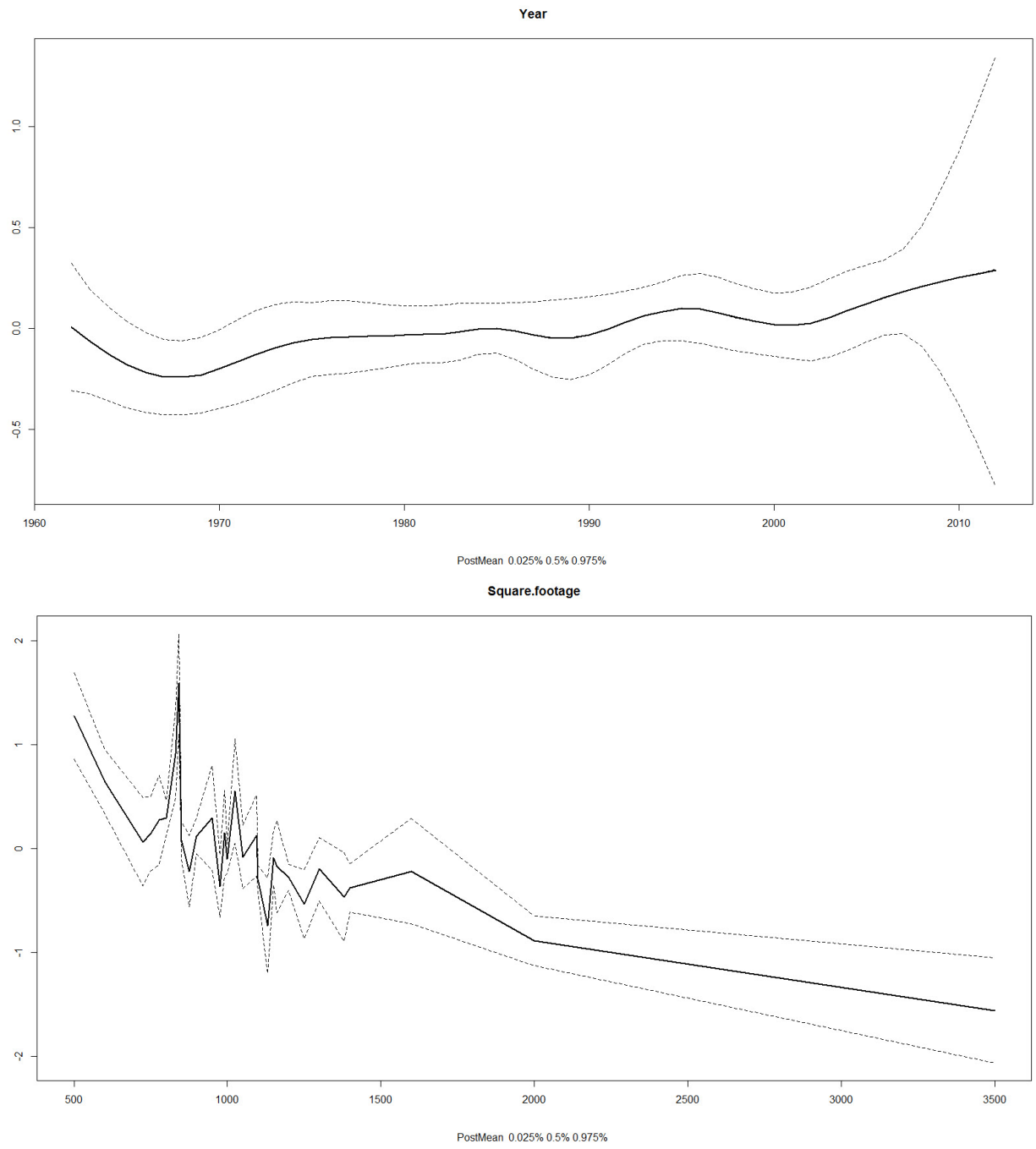

Figure 5.6: Posterior Distributions of Random Effects

We notice that there is certainly a nonlinear effect present in the square footage, as the effect is positive for smaller apartment units. The effect of Year of Construction is slightly linear at best, although it appears negligible at all points. Later years appear to have a slight

positive effect.

Finally, we examine the location effect in Figure 5.7, which we specified with a spatial model.

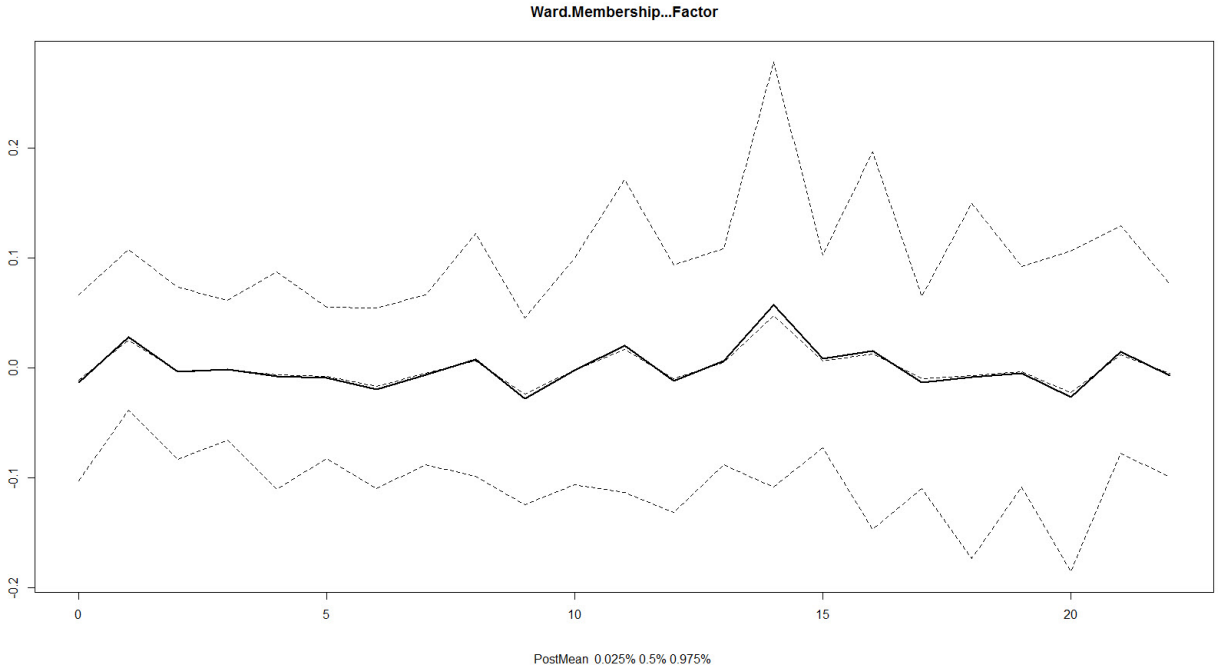

Figure 5.7: Posterior Distributions of Geoadditive Effect

I believe that this variable suffered the most from such a small sample size. Many of the districts contained only one data point. Having said that, the model suggests that a spatial effect is trivial at best.

5.5 Conclusion

Although the BYU approved effect was not incorporated into our final model, we were able to build a solid model for estimating the rental prices of apartments in the Provo, UT community and surrounding areas. Some of our results seem counterintuitive, but could possibly be explained by interactions with the BYU approved variable, but these interactions

were not tested in our model. We also observed a nonlinear effect for the square footage and a relatively linear, but small effect of the Year of Construction.

5.6 Future Work

While I feel that this analysis was thorough, I do believe that this exploration has opened up several questions that could be answered with more research. First and foremost, I think that some of the effects, particularly the two random effects and the spatial effect may have suffered from a small sample size. Any further study would certain improve with an increase in sample size, particularly with non-BYU approved housing available in Utah county. Second, as mentioned in some of the model discussion, I think any future work should include interaction effects between the BYU approved variable and several of the fixed effects. I think this would offer greater insight into how much the BYU approved variable really effects apartment rental prices. Last, it would be interesting to compare these results to a similar analysis done for apartments surrounding the University of Utah or Utah State University where no such living location approval by the university is required.

BIBLIOGRAPHY

Azzalini, A., and Capitano, A. (1999), "Statistical Applications of the Multivariate Skew Normal Distribution," *Journal of the Royal Statistical Society: Series B*, 61, 579–602.

Banjerjee, S., Carlin, B. P., and Gelfand, A. E. (2003), *Hierarchical Modeling and Analysis for Spatial Data*, London, UK: Taylor & Francis, Inc.

Cressie, N. A. (1993), *Statistics for Spatial Data*, New York, New York: Wiley Series in Probability and Mathematical Statistics.

Oliver, M. A. (2010), *Geostatistical Applications for Precision Agriculture*, New York, New York: Springer.

Pearson, K. (1911), *The Grammar of Science*, New York, New York: Cosimo.

Piegorsch, W. W., and Bailer, A. J. (2005), *Analyzing Environmental Data*, West Sussex, England: John Wiley & Sons.

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, London, UK: Chapman & Hall/CRC Press.

Schrodle, B., and Held, L. (2009), "Evaluation of Case Reporting Data From Switzerland: Spatio-temporal Disease Mapping Using INLA," *University of Zurich*, Tehnical Report, Biostatistics.

Tierney, L., and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of American Statistics Association*, 81, 82–86.

Waller, L. A., and Gotway, C. A. (2004), *Applied Spatial Statistics for Public Health Data*, Hoboken, New Jersey: John Wiley & Sons.