COMPUTATIONALLY EFFICIENT HIERARCHICAL SPATIAL MODELS FOR LARGE DATASETS: A CASE STUDY FOR THE ASSESSMENT OF FOREST CHARACTERISTICS ACROSS THE LAKE STATES

By

Huirong Zhu

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Forestry

2011

ABSTRACT

COMPUTATIONALLY EFFICIENT HIERARCHICAL SPATIAL MODELS FOR LARGE DATASETS: A CASE STUDY FOR THE ASSESSMENT OF FOREST CHARACTERISTICS ACROSS THE LAKE STATES

By

Huirong Zhu

The scientific community is moving into an era where data rich environments provide extraordinary opportunities to understand the spatial complexity of ecological processes. Across scientific fields, researchers face the challenge of coupling these data with imperfect models to better understand variability in their system of interest. In the environmental sciences there is recognized urgent need to develop and disseminate methodology capable of accurately accounting for multiple sources of uncertainty. Accordingly, the goal of this thesis was to explore and illustrate the properties of promising new modeling tools that will enable researchers to extract more information from large spatial datasets. In particular, this thesis was motivated by a larger project's need to analyze a large forest inventory dataset with the intent to better understand the potential of managing forests for increased complexity as a climate change mitigation and adaptation strategy. The thesis yields results from the analysis of synthetic and forestry datasets that clearly demonstrate how model misspecification, specifically ignoring spatial dependence among model residuals, can result in incorrect inference about regression parameters of interest. These results have important implications for hypothesis testing and ultimately forest management and policy decisions. The thesis details some modeling tools and useful guidelines that allow practitioners to more fully accommodate model assumptions and draw correct inference for large spatial datasets.

To my parents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The scientific community is moving into an era where data rich environments provide extraordinary opportunities to understand the spatial complexity of ecological processes. Increased availability of environmental datasets is a result of investments to collect data for regulatory, monitoring, and resource management objectives, and technological advances in spatially-enabled monitor and sensor networks along with geospatial information storage, analysis, and distribution systems. Across scientific fields, researchers face the challenge of coupling these data with imperfect models to better understand variability in their system of interest. The inference gained through these analyses often supports decisions with important economic and ecological implications; therefore, it is critical to correctly estimate inferential uncertainty. In the environmental sciences there is recognized urgent need to develop and disseminate methodology capable of accurately accounting for multiple sources of uncertainty (Cressie et al. 2009). However, developing modeling frameworks capable of accounting for various sources of uncertainty is not a trivial task, especially when datasets are large and data come from multiple sources.

The statistical literature acknowledges that spatial and temporal associations are captured most effectively using hierarchical models that build dependencies in different stages. Hierarchical models are especially advantageous with datasets having several lurking sources of uncertainty and dependence, where they can estimate much richer models with less stringent assumptions than traditional modeling paradigms (Banerjee et al. 2004). These hierarchical models, implemented through Markov chain Monte Carlo (MCMC) methods, offer increased flexibility to fit models that would be infeasible with classical methods (see e.g., Carlin and Louis 2000; Gelman et al. 2004). Despite their power and flexibility, practitioners have been slow to adopt these methods, due in part to lack of statistical/mathematical training, computationally efficient software, and illustration of these methods within their area of interest.

Accordingly, the goal of this thesis was to explore and illustrate the properties of promising new modeling tools that will enable researchers to extract more information from large spatial datasets. In particular, this thesis was motivated by a larger project's need to analyze a large forest inventory dataset with the intent to better understand the potential of managing forests for increased complexity as a climate change mitigation and adaptation strategy. Therefore, a central objective of this thesis is illustrate how the proposed methods can be applied to generate valid inference about patterns in forest ownership, forest complexity, and biomass that ultimately help support the goals of the larger project. To this end, and in an effort to make these illustrated methods more accessible to practitioners, this work also produced software and associated documentation for the highlighted geostatistical models.

The thesis yields results from the analysis of synthetic and forestry datasets that clearly demonstrate how model misspecification, specifically ignoring spatial dependence among

model residuals, can result in incorrect inference about regression parameters of interest. These results have important implications for hypothesis testing and ultimately forest management and policy decisions. The thesis details some modeling tools and useful guidelines that allow practitioners to more fully accommodate model assumptions and draw correct inference for large spatial datasets.

The thesis is organized as follows: Chapter 2 details the univariate spatial regression model, its properties, theoretical assumptions, and how it can be implemented within a Bayesian framework. Chapter 3 reviews covariance tapering and its ingredients including compactly supported correlation functions, sparse matrix structures, and software routines for sparse matrix linear algebra computations. Chapter 4 offers two numerical experiments used to compare computation efficiency and predictive information loss incurred when using compactly supported correlation functions in the spatial regression model. These models are then used to explore the relationship between indices of forest complexity, biomass, and private forest ownership across Lake States' forest in Chapter 5. A brief summary and discussion of future work is given in Chapter 6.

Chapter 2

Models for spatial data

2.1 A review of hierarchical Bayesian modeling

While the traditional methods, such as maximum likelihood, are reasonable in simpler contexts with assumptions of limited sources of variability, they are likely to be inadequate in capturing the total uncertainty in settings with multiple sources of variation and uncertainty. In such cases, one needs to model the data *hierarchically* by introducing parameters in different levels (or hierarchies) that will account for the different sources of measurement error and other extraneous variation. Over the past two decades hierarchical Bayesian modeling has seen enormous developments in the statistical literature. I provide a brief review of the underlying principles of hierarchical Bayesian modeling below.

By modeling both the observed data and any unknown parameters or other unobserved effects as random variables, the hierarchical Bayesian approach to statistical analysis provides a cohesive framework for combining complex data models and external knowledge or expert opinion (e.g., Berger, 1985; Carlin and Louis, 2000; Robert, 2001; Gelman et al., 2003; Lee,

2004). In this approach, in addition to specifying the distributional model $P(\mathbf{y}|\boldsymbol{\theta})$ for the observed data $\mathbf{y} = [y_i]_{i=1}^n$ given a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, we suppose that $\boldsymbol{\theta}$ is a random quantity sampled from a *prior* distribution $P(\boldsymbol{\theta}|\boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a vector of hyperparameters. Inference concerning $\boldsymbol{\theta}$ is then based on its *posterior* distribution,

$$P(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\lambda}) = \frac{P(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{P(\mathbf{y}|\boldsymbol{\lambda})} = \frac{P(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{\int P(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta}} = \frac{P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\int P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta}}\,. \tag{2.1}$$

Notice the contribution of both the data (in the form of the likelihood $P(\mathbf{y}|\boldsymbol{\theta})$) and the external knowledge or opinion (in the form of the prior $P(\boldsymbol{\theta}|\boldsymbol{\lambda})$) to the posterior. If $\boldsymbol{\lambda}$ is known, this posterior distribution is fully specified; if not, a second-stage prior distribution (called a *hyper-prior*) may be specified for it, leading to a *fully Bayesian* analysis. Alternatively, we might simply replace $\boldsymbol{\lambda}$ by an estimate $\hat{\boldsymbol{\lambda}}$ obtained as the value which maximizes the marginal distribution $P(\mathbf{y}|\boldsymbol{\lambda})$ viewed as a function of $\boldsymbol{\lambda}$. Inference proceeds based on the estimated posterior distribution $P(\boldsymbol{\theta}|\mathbf{y}, \hat{\boldsymbol{\lambda}})$, obtained by plugging $\hat{\boldsymbol{\lambda}}$ into equation (2.1). This is called an *empirical Bayes* analysis and is closer to maximum likelihood estimation techniques.

The Bayesian decision-making paradigm improves on the classical approaches to statistical analysis in its more philosophically sound foundation, its unified approach to data analysis, and its ability to formally incorporate prior opinion or external empirical evidence into the results via the prior distribution. Statisticians, formerly reluctant to adopt the Bayesian approach due to general skepticism concerning its philosophy and a lack of necessary computational tools, are now turning to it with increasing regularity as classical methods emerge as both theoretically and practically inadequate. Modeling the θ_i's as random (instead of fixed) effects allows us to induce specific (e.g. spatial, temporal or more general) correlation

structures among them, hence among the observed data y_i as well. Hierarchical Bayesian methods now enjoy broad application in the analysis of complex systems, where it is natural to pool information across from different sources (e.g. Gelman et al., 2003).

Modern Bayesian methods seek complete evaluation of the posterior distributions using simulation methods that draw samples from the posterior distribution. This sampling-based paradigm enables *exact* inference free of unverifiable asymptotic assumptions on sample sizes and other regularity conditions. A computational challenge in applying Bayesian methods is that for many complex systems, the simulations required to do inference under (2.1) generally involve distributions that are intractable in closed form, and thus one needs more sophisticated algorithms to sample from the posterior. Forms for the prior distributions (called *conjugate* forms) may often be found which enable at least partial analytic evaluation of these distributions, but in the presence of nuisance parameters (typically unknown variances), some intractable distributions remain. Here the emergence of inexpensive, high-speed computing equipment and software comes to the rescue, enabling the application of recently developed Markov Chain Monte Carlo (MCMC) integration methods, such as the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990). Univariate MCMC algorithms are particularly attractive for general purpose implementation, since all that is required is the ability to sample easily from each parameter's complete conditional distribution, namely $P(\theta_i|\mathbf{y}, \theta_{j\neq i}),\ i = 1, \ldots, k$. The recently developed `WinBUGS` language (`www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`) and the `R` statistical platform (`www.r-project.org`) with its Bayesian packages are promising steps towards a general purpose software package for hierarchical modeling, though it may be insufficiently general

in some advanced analysis explored in this thesis.

Statistical prediction in a Bayesian settings is particularly elegant and intuitive. Let $\mathbf{y}_{pred}$ denote the variables (they can be a collection) we seek to predict. Then, we simply treat $\mathbf{y}_{pred}$ as a random variable whose *prior*, conditional upon the parameters, is the data distribution $P(\mathbf{y}_{rep}|\boldsymbol{\theta})$. Then, all predictions will be summarized in the *posterior predictive* distribution:

$$P(\mathbf{y}_{pred}|\mathbf{y}) = \int P(\mathbf{y}_{pred}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.$$

Once the posterior samples are available from $P(\boldsymbol{\theta}|\mathbf{y})$, it is routine to draw samples from $P(\mathbf{y}_{pred}|\mathbf{y})$ using the principle of *composition*(Banerjee et al. 2004): for each posterior draw of $\boldsymbol{\theta}$, say draw a $\mathbf{y}_{pred}$ from $P(\mathbf{y}_{pred}|\boldsymbol{\theta})$. Details of such methods are particularly well explained in the texts by Carlin and Louis (2000) and Gelman et al. (2004).

2.2 Univariate spatial regression

Geostatistical settings typically assume, at locations $\mathbf{s} \in D \subseteq \Re^2$, where $\mathbf{s}$ indexes spatial coordinates (e.g., easting and northing, longitude and latitude), D is region of interest, and $\Re^2$ represents a 2-dimentional space, a Gaussian outcome variable $y(\mathbf{s})$ along with a $p \times 1$ vector of spatially referenced predictors $\mathbf{x}(\mathbf{s})$ are associated through a spatial regression model such as,

$$y\left(\mathbf{s}\right) = \mathbf{x}\left(\mathbf{s}\right)'\boldsymbol{\beta} + w\left(\mathbf{s}\right) + \epsilon\left(\mathbf{s}\right). \tag{2.2}$$

In this model the residual comprises a spatial process, $w(\mathbf{s})$, and an independent white-noise process, $\epsilon(\mathbf{s})$, that captures measurement error or micro-scale variation. With any

collection of n locations, say $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, we assume that the $\epsilon(\mathbf{s}_i)$'s are identically and independently distributed as $N(0, \tau^2)$, where τ^2 is often called the *nugget*.

The $w(\mathbf{s})$ are spatial random effects, providing local adjustment (with structured dependence) to the mean, interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern. A popular modeling choice for the spatial random effects is the *Gaussian process*, $w(\mathbf{s}) \sim GP(0, K(\cdot, \cdot; \boldsymbol{\theta}))$, specified by a valid covariance function $K(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) = \text{Cov}(w(\mathbf{s}_1), w(\mathbf{s}_2))$ that models the covariance corresponding to a pair of locations $\mathbf{s}_1$ and $\mathbf{s}_2$. The process realizations are collected into a $n \times 1$ vector, say $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))'$, which follows a multivariate normal distribution $MVN(\mathbf{0}, \Sigma_{\mathbf{w}})$, where $\Sigma_{\mathbf{w}}$ is the $n \times n$ covariance matrix of $\mathbf{w}$ with (i, j)-th element given by $K(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})$.

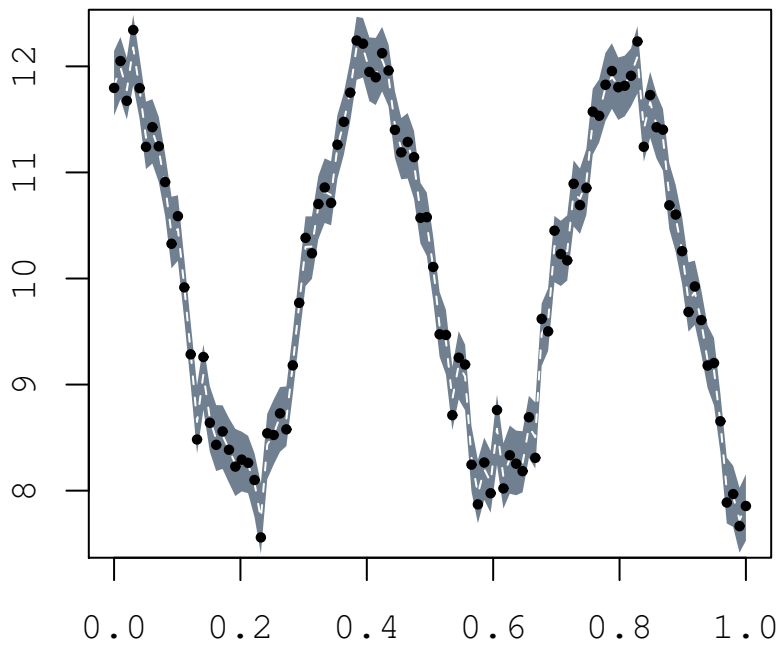

Figure 2.1: Illustrative example of model (2.2) fit to $\Re^1$ data generated from an *unknown* function indicated with point symbols. White dashed lines and gray regions indicate the median and 95% credible intervals of the fitted $\mathbf{y}$. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this thesis.

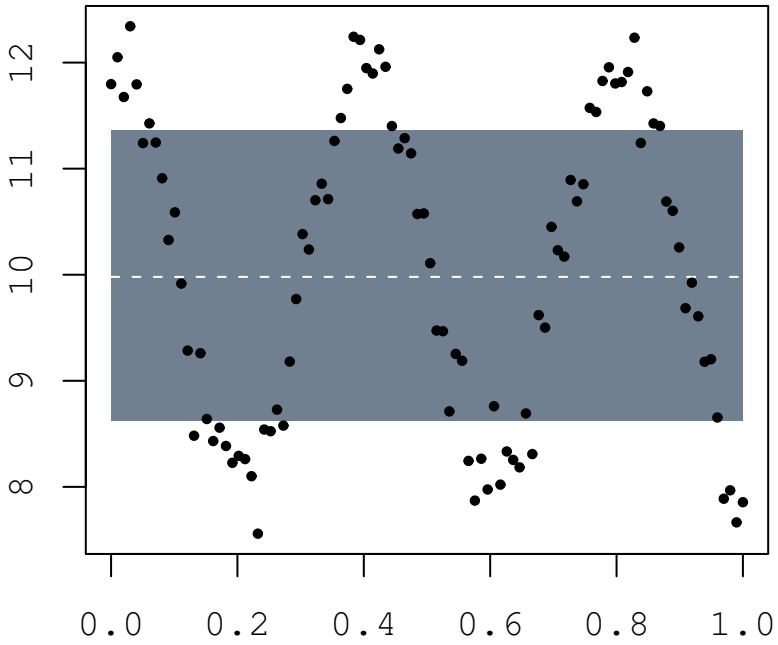

Figure 2.2: Illustrative example of model (2.2) fit to $\Re^1$ data generated from an *unknown* function indicated with point symbols. White dashed lines and gray regions indicate the median and 95% credible intervals of the intercept β_0.

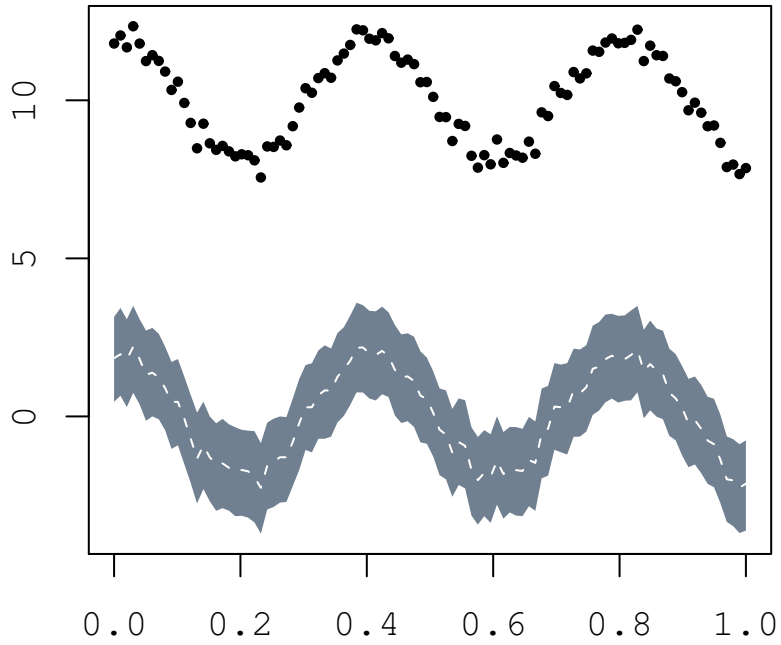

Figure 2.3: Illustrative example of model (2.2) fit to $\Re^1$ data generated from an *unknown* function indicated with point symbols. White dashed lines and gray regions indicate the median and 95% credible intervals of the random effects $\mathbf{w}$.

To illustrate the flexibility of a GP, 100 observations were generated from a normal distribution $N(10 + 2\sin(5\pi x + 1.5), 0.1)$, where x ranges from 0 to 1, indicated with point symbols in Figure 2.1, 2.2 and 2.3. These data were then fit using model (2.2). Figure 2.1, 2.2 and 2.3 shows the estimated fitted values, and mean components β_0 and $\mathbf{w}$. This figure clearly shows that even if the underlying function that generates the data is not known, it can be closely approximated using a GP.

Clearly $K(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta})$ cannot be just any function; it must ensure that the resulting $\Sigma_{\mathbf{w}}$

matrix is symmetric and positive definite. Such functions are known as positive definite functions and are characterized as the characteristic function of a symmetric random variable (due to a famous theorem by Bochner (see, e.g., Gikhman and Skorokhod, 1974, p.208)). Further technical details about positive definite functions can be found in Cressie (1993), Chilés and Delfiner (1999), and Banerjee et al. (2004).

We customarily specify $K(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) = \sigma^2 \rho(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\phi})$ with $\boldsymbol{\theta} = \{\sigma^2, \boldsymbol{\phi}\}$ where $\rho(\cdot; \boldsymbol{\phi})$ is a *correlation function* while $\boldsymbol{\phi}$ includes parameters quantifying rate of correlation decay and smoothness of the surface $w(\mathbf{s})$. Then $\text{Var}(w(\mathbf{s})) = \sigma^2$ represents the spatial variance component in the model (2.2). This specification assumes a *stationary* and *isotropic* processes. Stationarity, in spatial modeling contexts, refers to the setting when $K(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) = K(\mathbf{s}_1 - \mathbf{s}_2; \boldsymbol{\theta})$; that is, the covariance function depends upon the separation of the locations. Isotropy goes further and specifies $K(\mathbf{s}_1, \mathbf{s}_2) = K(\|\mathbf{s}_1 - \mathbf{s}_2\|)$, where $\|\mathbf{s}_1 - \mathbf{s}_2\|$ is the Euclidean distance between the locations $\mathbf{s}_1$ and $\mathbf{s}_2$. A very versatile class of correlation functions is the Matérn correlation function given by

$$\rho(\|\mathbf{s}_1 - \mathbf{s}_2\|; \boldsymbol{\phi}) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(\|\mathbf{s}_1 - \mathbf{s}_2\|\phi)^{\nu}\mathcal{K}_{\nu}(\|\mathbf{s}_1 - \mathbf{s}_2\|; \phi); \quad \phi > 0, \nu > 0. \tag{2.3}$$

Here $\boldsymbol{\phi} = \{\phi, \nu\}$ with ϕ controlling the decay in spatial correlation and ν controlling process smoothness. Specifically, if ν lies between positive integers m and $(m+1)$, then the spatial process $w(\mathbf{s})$ is mean-square differentiable m times, but not $m+1$ times. This fact has been used to study the derivatives of spatial surfaces (e.g., Banerjee, Gelfand and Sirmans, 2003). Also, Γ is the usual Gamma function while $\mathcal{K}_{\nu}$ is a modified Bessel function of the third kind with order ν. Note, when $\nu = 1/2$ (2.3) reduces to the familiar exponential correlation function.

For multilevel models, one assigns prior (hyperprior) distributions to the model parameters (hyperparameters) and inference proceeds by sampling from the posterior distribution of the parameters (see, e.g., Banerjee et al., 2004). Note that the spatial process induces a $MVN(\mathbf{0}, \sigma^2 R(\boldsymbol{\phi}))$ distribution on $\mathbf{w}$. For the remaining parameters, suppose that $\boldsymbol{\beta}$ is assigned a multivariate Gaussian prior, say $MVN(\boldsymbol{\mu}_\beta, \Sigma_\beta)$, while the spatial variance component σ^2 and the measurement error variance τ^2 are assigned inverse-Gamma (IG) priors. The process correlation parameter(s), $\boldsymbol{\phi}$, are usually assigned informative priors (e.g., uniform over a finite range and to perhaps ensure a desired degree of freedom for the spatial surface) based on the underlying spatial domain and the physical processes under study.

Using notations similar to Gelman et al. (2004), we can write the posterior distribution by $p(\boldsymbol{\beta}, \mathbf{w}, \sigma^2, \tau^2, \boldsymbol{\phi} \,|\, \mathbf{y})$ as proportional to

$$p(\boldsymbol{\phi}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \times N(\mathbf{w} \,|\, \mathbf{0}, \sigma^2 R(\boldsymbol{\phi})) \times \prod_{i=1}^{n} N(y(\mathbf{s}_i) \,|\, \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + w(\mathbf{s}_i), \tau^2). \tag{2.4}$$

Estimation of (2.4) customarily proceeds using an MCMC algorithm. Often a marginalized likelihood is used that is obtained by integrating out the spatial effects $\mathbf{w}$. This yields the posterior distribution $p(\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\phi} \,|\, \mathbf{y})$ that is proportional to

$$p(\boldsymbol{\phi}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \times N\left(\mathbf{y} \,|\, X\boldsymbol{\beta}, \sigma^2 R(\boldsymbol{\phi}) + \tau^2 I_n\right), \tag{2.5}$$

where X is the matrix of regressors whose i-th row is $\mathbf{x}(\mathbf{s}_i)'$ and $\mathbf{I}_n$ is the $n \times n$ identity

matrix. The log likelihood of $N\left(\mathbf{y} \mid X\boldsymbol{\beta}, \sigma^2 R(\boldsymbol{\phi}) + \tau^2 I_n\right)$ is proportional to

$$-\frac{1}{2}\log|\Sigma_y| - \frac{1}{2}\mathbf{z}'\Sigma_y^{-1}\mathbf{z}, \tag{2.6}$$

where $\Sigma_y = \sigma^2 R(\boldsymbol{\phi}) + \tau^2 I_n$ and $\mathbf{z} = \mathbf{y} - X'\boldsymbol{\beta}$.

An efficient MCMC algorithm for updating $\boldsymbol{\beta}$ from its full conditional $MVN(\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot}, \Sigma_{\boldsymbol{\beta}|\cdot})$, where

$$\Sigma_{\boldsymbol{\beta}|\cdot} = [\Sigma_{\boldsymbol{\beta}}^{-1} + X'\Sigma_y^{-1}X]^{-1} \text{ and } \boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot} = \Sigma_{\boldsymbol{\beta}|\cdot}X'\Sigma_y^{-1}\mathbf{y}\ . \tag{2.7}$$

The remaining parameters can be updated using Metropolis steps, where the target distribution for any (set of) parameter(s) is proportional to the product of the terms in (2.5) that involve that (those) parameter(s). These will yield posterior samples of $\Omega = \{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}, \tau^2\}$.

In this sampling scheme, the spatial random effects $\mathbf{w}$ are not sampled directly. This reduces the parameter space, which results in a more efficient MCMC algorithm. A key advantage of the first stage Gaussian model (as in 2.2) is that samples from the posterior distribution of $\mathbf{w}$ can be recovered in a posterior predictive fashion. More precisely, we seek to evaluate

$$p(\mathbf{w} \mid \mathbf{y}) \propto \int p(\mathbf{w} \mid \Omega, \mathbf{y}) p(\Omega \mid \mathbf{y})\ d\Omega\ . \tag{2.8}$$

Because the full conditional distribution of $\mathbf{w}$ in (2.8) is again multivariate normal, (2.8) is easily evaluated using *composition sampling*. For each sample $\{\boldsymbol{\beta}^{(l)}, \sigma^{2(l)}, \tau^{2(l)}, \boldsymbol{\phi}^{(l)}\}$, we draw $\mathbf{w}^{(l)}$ from $N(\boldsymbol{\mu}_{\mathbf{w}|\cdot}, \Sigma_{\mathbf{w}|\cdot})$, where

$$\Sigma_{\mathbf{w}|\cdot} = \left[\frac{R\left(\boldsymbol{\phi}^{(l)}\right)^{-1}}{\sigma^{2(l)}} + \frac{I_n}{\tau^{2(l)}}\right]^{-1} \text{ and } \boldsymbol{\mu}_{\mathbf{w}|\cdot} = \Sigma_{\mathbf{w}|\cdot}\frac{\left(\mathbf{y} - X\boldsymbol{\beta}^{(l)}\right)}{\tau^{2(l)}}. \tag{2.9}$$

These posterior samples from $\mathbf{w}$ can be mapped using contours or interpolation routines to produce plots of the spatial process. For predictions, if $\mathcal{S}_0 = \{\mathbf{s}_{0,1}, \mathbf{s}_{0,2}, \ldots, \mathbf{s}_{0,n_0}\}$ is a collection of n_0 *new* locations, the posterior predictive distribution of $p(\mathbf{w}_0 \,|\, \mathbf{y})$, where $\mathbf{w}_0 = (w(\mathbf{s}_{0,1}), w(\mathbf{s}_{0,2}), \ldots, w(\mathbf{s}_{0,n_0}))'$, is

$$p(\mathbf{w}_0 \,|\, \mathbf{y}) \propto \int p(\mathbf{w}_0 \,|\, \mathbf{w}, \Omega, \mathbf{y}) p(\mathbf{w} \,|\, \Omega, \mathbf{y}) p(\Omega \,|\, \mathbf{y}) d\Omega \mathbf{w}. \tag{2.10}$$

Given posterior samples, $\{\Omega^{(l)}\}_{l=1}^{L}$, this distribution can again be obtained via composition sampling: we first draw $\mathbf{w}^{(l)}$ for each l as described in (2.8) and then draw $\mathbf{w}_0^{(l)}$ from $p(\mathbf{w}_0 \,|\, \mathbf{w}^{(l)}, \Omega^{(l)}, \mathbf{y})$, where this last distribution is derived as a conditional distribution from a multivariate normal and, hence, is again multivariate normal. More precisely, the process realizations over the new locations are conditionally independent of the observed outcomes given the realizations over the observed locations and the process parameters. In other words, $p(\mathbf{w}_0 \,|\, \mathbf{w}, \Omega, \mathbf{y}) = p(\mathbf{w}_0 \,|\, \mathbf{w}, \Omega)$, which is a multivariate normal distribution with mean and variance given by

$$\begin{aligned} \mathrm{E}[\mathbf{w}_0 \,|\, \mathbf{w}, \Omega] &= \mathrm{Cov}(\mathbf{w}_0, \mathbf{w}) \mathrm{Var}^{-1}(\mathbf{w}) \mathbf{w} = r_0(\boldsymbol{\phi})' R(\boldsymbol{\phi})^{-1} \mathbf{w} \\ \text{and } \mathrm{Var}[\mathbf{w}_0 \,|\, \mathbf{w}, \Omega] &= \sigma^2 \left\{ R(\boldsymbol{\phi}) - r_0(\boldsymbol{\phi})' R(\boldsymbol{\phi})^{-1} r_0(\boldsymbol{\phi}) \right\} , \end{aligned}$$

where $r_0(\boldsymbol{\phi})'$ is the $n_0 \times n$ matrix with (i, j)-th element given by $\rho(\mathbf{s}_{0,i}, \mathbf{s}_j; \boldsymbol{\phi})$ and $R_0(\boldsymbol{\phi})'$ is the $n_0 \times n_0$ matrix define analogous to $R(\boldsymbol{\phi})$. Finally, given a set of covariates at a new location $\mathbf{s}_0$, samples from the posterior predictive distribution of the outcome variable, $y(\mathbf{s}_0)^{(l)}$, are drawn from $N(\mathbf{x}(\mathbf{s}_0)' \boldsymbol{\beta}^{(l)} + \mathbf{w}(\mathbf{s}_0)^{(l)}, \tau^{2(l)})$ for $l = 1, 2, \ldots, L$.

2.2.1 Non-spatial and spatial model numerical experiment

In this section a synthetic dataset analysis is used to illustrate how model misspecification can negatively effect inference. In particular, the analysis shows how regression slope parameter estimates and model fit changes when one fails to accommodate spatial dependence among the regression model's residuals. The synthetic data were generated at 500 locations within a $[0, 100] \times [0, 100]$ unit domain. Observations arose from the likelihood portion of model (2.2). The design matrix X included an intercept and two covariates each with values drawn from a $N(0, 1)$ distribution. Data were generated using the slope parameters $\boldsymbol{\beta} = (1, 0.01, 0.5)$ and Σ_y was formed using $\sigma^2 = 20$, $\tau^2 = 0.1$, and spatial decay parameter $\phi = 0.015$. The exponential spatial correlation function was used to construct the correlation matrix $R(\phi)$. Obviously, with such a long *effective spatial range* of 200 map units and large σ^2 these data exhibit strong residual spatial dependence. Here, the *effective spatial range* is defined as the distance at which the correlation drops below some value, e.g., 0.05 which is $-\log(0.05)/\phi \approx 3/\phi$.

The candidate models included the spatial model (2.2) and its non-spatial equivalent (i.e., $\mathbf{w} = \mathbf{0}$). Posterior distributions of the parameters are used to compare between the non-spatial and spatial models. In addition to comparing the parameter estimates, non-spatial and spatial models are assessed based on fit using the popular *Deviance Information Criterion*, or DIC (Spiegelhalter et al., 2002). This criterion is the sum of the Bayesian deviance (a measure of model fit) and the (effective) number of parameters p_D (a penalty for model complexity).lower DIC indicates *better* fit.

Table 2.1 summarizes parameter estimates from non-spatial model. Here, β_0, β_1, and β_2 are statistically different from zero at the 0.05 level. This level of significance would suggest

Table 2.1: Non-spatial model's parameter credible intervals for the synthetic data. Parameters that are statistically significant at the 0.05 level are indicated with an "*".

	2.5%	25%	50%	75%	97.5%
β_0^*	2.95	4.48	5.42	6.34	7.96
β_1^*	0.09	0.26	0.34	0.42	0.60
β_2^*	0.08	0.23	0.32	0.42	0.57
τ^2	7.66	8.27	8.62	8.97	9.79

that the covariates are *useful* for explaining some of the variability in the outcome variable and, if this was a *real* dataset, one would make inferential statements about the associated covariates, e.g., regarding some posited hypotheses.

A simple exploration of the non-spatial model's residuals would reveal the strong spatial dependence, which would encourage the subsequent specification of model (2.2). Table 2.2 summarizes parameter estimates from spatial model. Here, intervals for β_0 and β_1 now included zero, which suggests they are not contributing significantly to explaining the variability in the outcome variable. As a result, one would come to very different conclusions about the impact of these variables. Further, the median parameter estimates are much closer to the *true* values used to generate the data.

Table 2.2: Spatial model's parameter credible intervals for the synthetic data. Parameters that are statistically significant at the 0.05 level are indicated with an "*".

	2.5%	25%	50%	75%	97.5%
β_0	-2.40	1.45	2.90	4.47	8.25
β_1	-0.01	0.04	0.06	0.09	0.14
β_2^*	0.48	0.54	0.56	0.59	0.65
σ^2	7.40	13.73	19.86	25.02	33.89
τ^2	0.02	0.04	0.06	0.10	0.15
ϕ	0.01	0.01	0.02	0.03	0.05

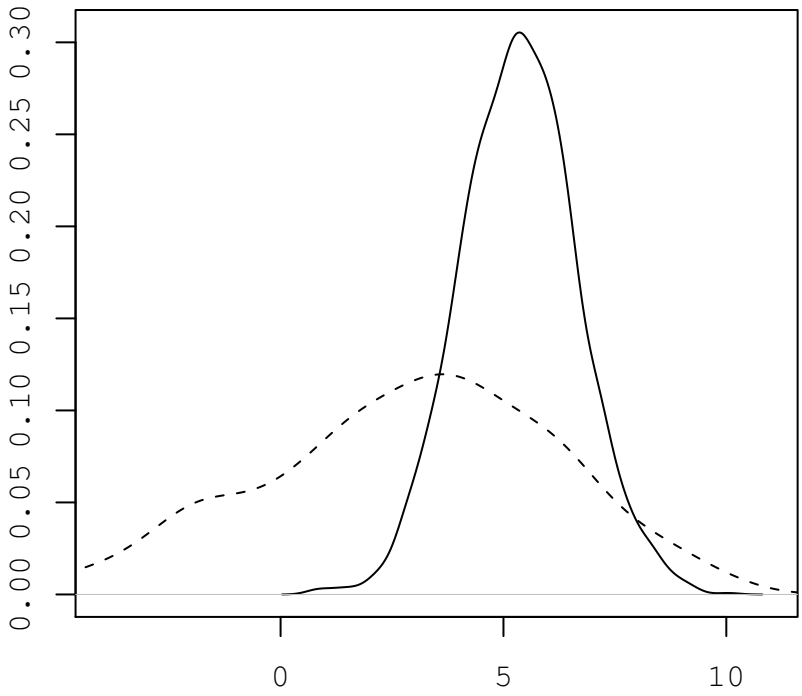

Figure 2.4: Illustration of intercept' posterior distribution in non-spatial model and spatial model: solid line represents non-spatial model; dashed lines represents spatial model

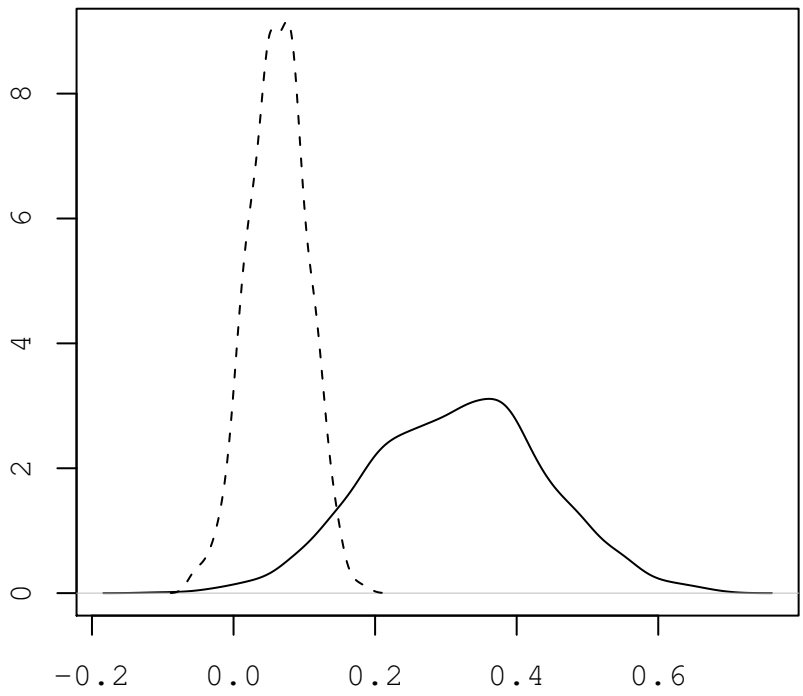

Figure 2.5: Illustration of estimated parameter' posterior distribution for X1 in non-spatial model and spatial model: solid line represents non-spatial model; dashed lines represents spatial model

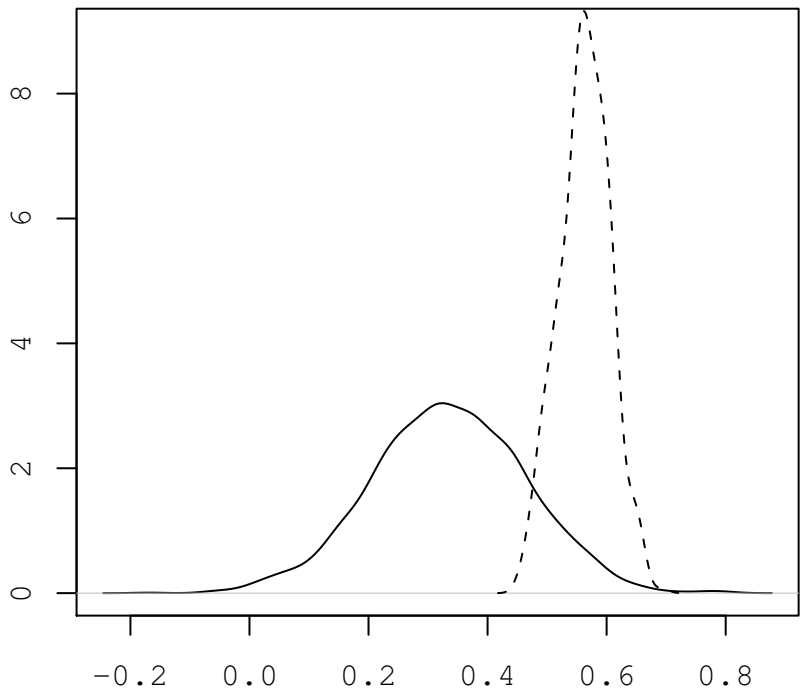

Figure 2.6: Illustration of estimated parameters' posterior distribution for X2 in non-spatial model and spatial model: solid line represents non-spatial model; dashed lines represents spatial model

These differences in the parameters' posterior distributions for the two models are illustrated in Figure 2.4, 2.5 and 2.6. Here, the posterior distribution shift is induced by the violation of model assumptions. For intercept, the posterior distribution of β_0 for the spatial model is wider than that of the non-spatial model. Because the spatial random effect is zero-centered, it in effect assumes the centering of β_0 and its dispersion is reflected in the distribution of β_0. The posterior distributions of β_1 and β_2 for the spatial model are narrower than those corresponding non-spatial model. This is partially because the total

variance in the spatial model is partitioned between $w(s)$ and $\epsilon(s)$, so that variance related to β_1 and β_2 becomes smaller than that of non-spatial model.

The DIC model fit criterion for non-spatial and spatial model are summarized in Table 2.3. Compared with the non-spatial model, the approximate decrease in DIC of $\sim$2000 units suggest the addition of the spatial random effects greatly improve model fit. The substantially larger pD of the spatial model reflects the addition of the spatial random effects.

Table 2.3: Model comparison using Deviance Information Criterion between spatial model and spatial model.

Model	pD	DIC
Non-spatial	3.9	1582.4
Spatial	379	-489

This example highlights the inadequacy of the non-spatial model to capture the extraneous structured variation. Failing to specify a model able to accommodate these residual dependence structures, i.e., meet basic model assumptions, can result in misleading inferences, poor model fit, and erroneous prediction.

2.3 Methods for modeling large datasets

Full inference and accurate assessment of model (2.2) will involve matrix decompositions of either Σ_y or Σ_w, depending on the model parametrization. The computational complexity of these factorizations are in the order of $O(n^3)$, where n is the number of locations. Hence, we encounter computational in feasibility ,a.k.a the "big n" problem, for large datasets.

Modeling large spatial datasets has received much attention in the statistical literature. Vecchia (1988) proposed approximating the likelihood with a product of appropriate conditional distributions to obtain maximum-likelihood estimates. Stein et al. (2004) adapt

this to restricted maximum likelihood estimation. Another possibility is to approximate the likelihood using spectral representations of the spatial process (Fuentes 2007). These likelihood approximations yield a joint distribution, but not a process that facilitates spatial interpolation. Rather than approximations, one could build models especially geared towards handling of large spatial datasets. These are representations of the spatial process in a lower-dimensional subspace and are often referred to as low-rank or reduced-rank spatial models. Here, the idea is to represent the spatial process over the n locations using a smaller number of locations referred to as "knots" (Higdon 2002; Kamman and Wand 2003; Stein 2007, 2008; Cressie and Johannesson 2008; Banerjee et al. 2008; Crainiceaniu et al. 2008). In practice, however, it is often difficult for analysts to choose a reasonable knot intensity and distribution across the domain. Yet another approach, referred to as *covariance tapering*, considers compactly supported correlation functions (Furrer et al. 2006; Kaufman et al. 2008; Du et al. 2009) that yield sparse correlation matrices which, given sparse solvers, are used for computationally efficient kriging and variance estimation.

For the simple spatial regression model (e.g., 2.2) this approach requires minimal input from the analyst beyond evaluation of empirical semivariogram. Hence, from the standpoint of production-level implementation, this approach can provide substantial computational gain without requiring a high-level of analyst modeling expertise, relative to other modeling approaches. In this work, I propose this covariance tapering method to approximate the likelihood and tackle the big "n"problem in this setting.

Chapter 3

Covariance tapering

This chapter introduces covariance tapering as an approach to reduce the computational demand associated with evaluating the likelihood in model (2.2). As noted in Chapter 2.3, the iterative factorization of the $n \times n$ spatial covariance matrix is the computational bottleneck in estimating this model's parameters within a Bayesian modeling framework.

The premise of covariance tapering is that the spatial covariance matrix is dominated by near-zero values. This occurs when the range of residual spatial dependence is short relative to the size of the domain. The idea of covariance tapering is to replace those near-zeros entries of the covariance matrix with zeros – effectively making the covariance matrix sparse. Sparse matrix structures, specially designed to only maintain the information about non-zero elements, can be used to dramatically reduces computational costs associated with data movement and storage. A suite of matrix algorithms specially designed to work on these sparse matrix structures are available to efficiently compute factorizations for the determinant and inverse.

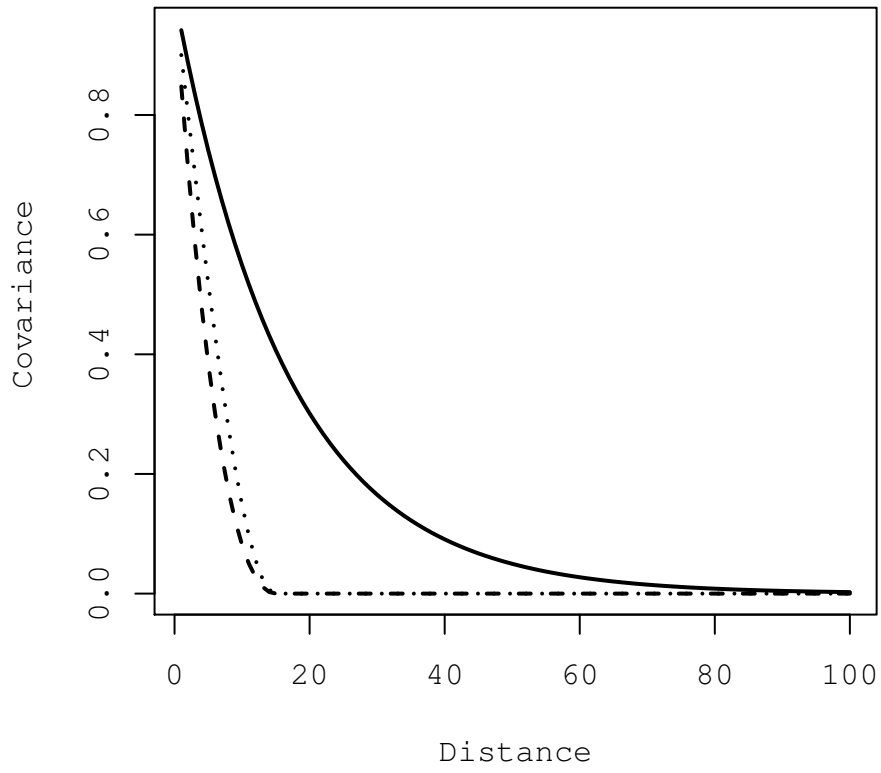

Figure 3.1: Covariance tapering with different ratios of taper range and effective range. Curves of original covariance function (solid line), taper correlation function (fine dotted) and tapered covariance function (long dashed). Ratio of taper range over effective range is 30%.

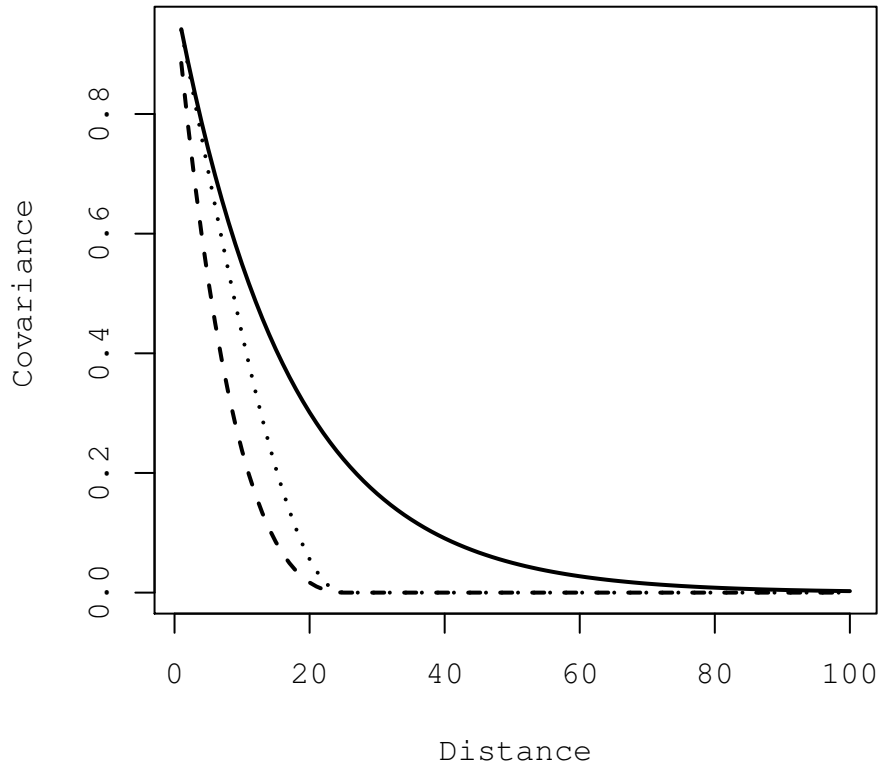

Figure 3.2: Covariance tapering with different ratios of taper range and effective range. Curves of original covariance function (solid line), taper correlation function (fine dotted) and tapered covariance function (long dashed). Ratio of taper range over effective range is 50%.

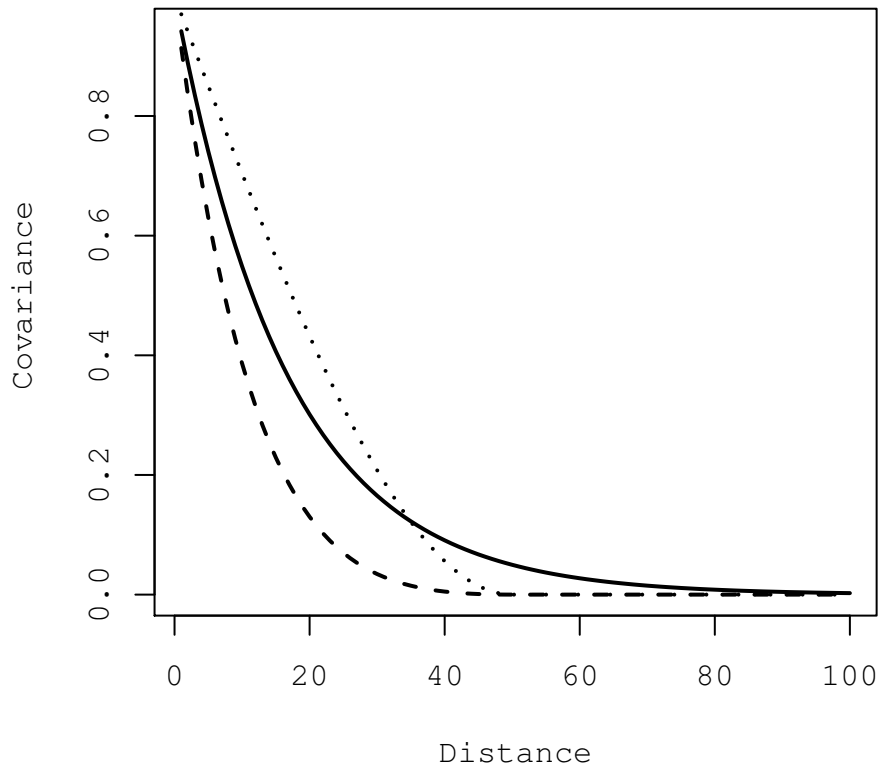

Figure 3.3: Covariance tapering with different ratios of taper range and effective range. Curves of original covariance function (solid line), taper correlation function (fine dotted) and tapered covariance function (long dashed). Ratio of taper range over effective range is 100%.

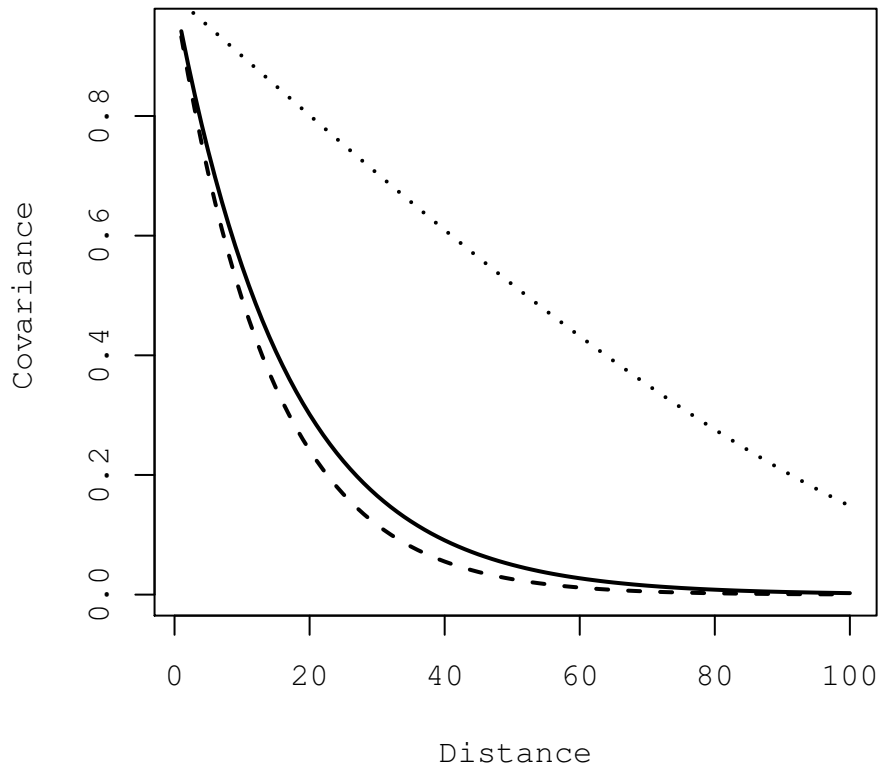

Figure 3.4: Covariance tapering with different ratios of taper range and effective range. Curves of original covariance function (solid line), taper correlation function (fine dotted) and tapered covariance function (long dashed). Ratio of taper range over effective range is 300%.

The critical question is: *how do we introduce zeros into the original covariance matrix while ensuring the resulting matrix remains positive definite and maintains desirable properties of the original covariance matrix.* This can be accomplished by taking the element-wise product of two spatial correlation matrices – the original *dense* correlation matrix and a sparse correlation matrix created using a particular type of correlation function, referred to as a compactly supported correlation function, which produce zeros beyond a specified spatial range. This idea is illustrated in Figure 3.1, 3.2, 3.3 and 3.4. Here, the three lines correspond to the original *dense* correlation function, compactly supported correlation function, and their product. It can be seen that their product retains several features of the original correlation function, but also "tapers" to zero at the taper range of the compactly supported function. From both a computational and model specification standpoint this figure also highlights the need to balance the ratio of the taper range and effective range, which will be more fully explored in the remainder of this thesis.

3.1 Compactly supported correlation functions

A compactly supported correlation function, K_θ, is an isotropic correlation function that equals zero beyond a particular spatial range of θ. This is referred to as the *taper range* and it determines the sparseness of the resulting tapered covariance matrix. There are several approaches to specifying compactly supported correlation functions, see, e.g., Wu, 1995; Gaspari and Cohn, 1999; and Gneiting, 2002). For practical application, Furrer et al. (2006) suggested that K_θ be the spherical or Wendland function (Wendland, 1995, 1998).

For the application in this work, I consider the spherical correlation function

$$K_\theta(t;\phi_\theta) = \begin{cases} 1 - \frac{3}{2}\phi_\theta t + \frac{1}{2}(\phi_\theta)^3 t, & 0 < t < \frac{1}{\phi_\theta} \\ 0, & \text{otherwise,} \end{cases} \tag{3.1}$$

where $t = \|\mathbf{s}_1 - \mathbf{s}_2\|$. The spherical correlation function is commonly used in environmental and biological sciences, and produces zero values beyond the taper range of $1/\phi_\theta$.

Given a compactly supported correlation function K_θ, a *tapered covariance function* is formed by taking the product of the covariance function K and K_θ.

The subsequent development and analyses assumes an exponential spatial correlation function for K; however, any valid spatial correlation function could be applied. Here, the (i,j)-th element of Σ (where Σ is either Σ_y or Σ_w) is $K(\|\mathbf{s}_i - \mathbf{s}_j\|, \sigma^2, \phi) = \sigma^2 \exp(-\phi\|\mathbf{s}_i - \mathbf{s}_j\|)$. Then given this dense Σ and sparse $\Sigma_\theta = \left[K_\theta(\|\mathbf{s}_i - \mathbf{s}_j\|;\phi_\theta)\right]_{i,j=1}^n$, the tapered covariance matrix $\Sigma_{tap} = \Sigma \circ \Sigma_\theta$, where "$\circ$" is the direct or Schur product, i.e., the element-wise matrix product. Importantly, the schur product of two symmetric and positive definite covariance matrices is itself symmetric and positive definite.

The ingredients of covariance tapering are illustrated in Figure 3.5. Here, the curves represent the output from the original exponential covariance function, compactly supported correlation function with $1/\phi_\theta = 40$ distance units as the spatial taper range, and resulting tapered covariance function. As shown by the long dashed curve, this tapered covariance maintains the desired exponential type decay until it is truncated at 40 distance units.

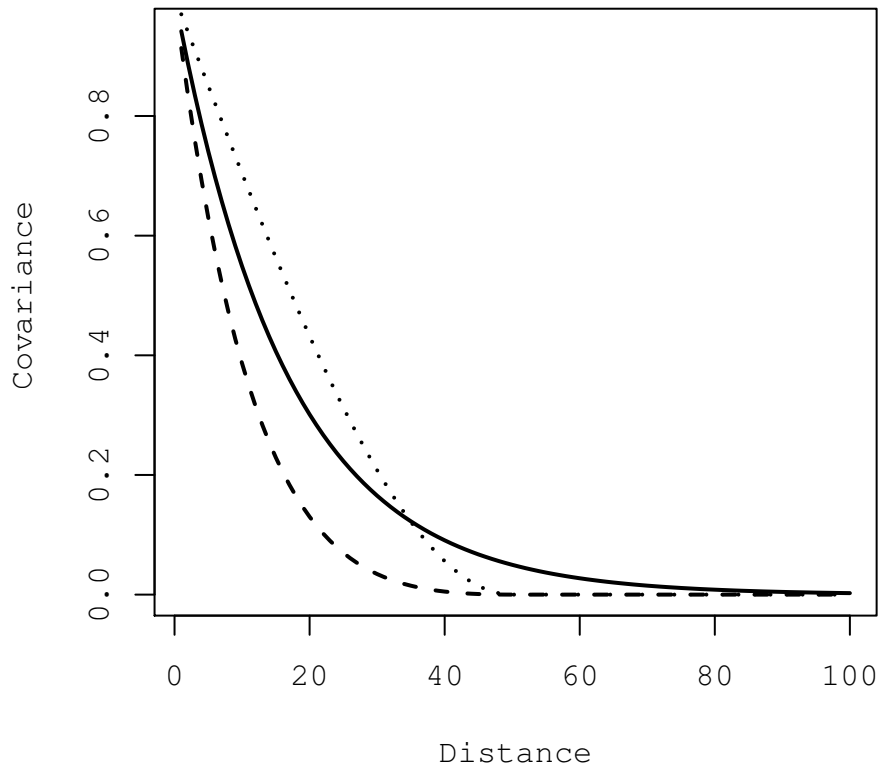

Figure 3.5: Illustration of covariance tapering: exponential covariance function with $\phi = 0.06$ and $\sigma^2 = 1$ (solid); compactly supported correlation function (3.1) with $\phi_\theta = 0.025$ (fine dotted), and; resulting tapered covariance function (long dashed).

3.2 Benefits of covariance tapering method

Stein (1988, 1990, 1997, and 1999) has written extensively about the consequences of misspecifying the spatial covariance function. Here misspecification refers to using a covariance function that is not similar to the *true* covariance function that gave rise to the data. Stein and others identify correlation functions from the Matérn family (2.3) as the most versatile among the common valid correlation functions (see, e.g., Banerjee et al. 2004, p. 27 for a list). Therefore, it is important to show that even after tapering these correlation functions

maintain their versatility and optimal predictive characteristics. Much of the theoretical foundation for justifying covariance tapering is laid out by Furrer et al. (2006), Kaufman et al. (2008), Zhang and Du (2007), and Du et al. (2009).

In the context of a fixed domain, isotropic and stationary Matérn covariance model, asymptotic equivalence of the mean squared prediction error of the corresponding spectral densities can be obtained based on the tail condition (Furrer et al. 2006). Zhang (2004) showed several important results about the Matérn class to prove the equivalence of two mean-zero measures and convergence of the estimators under those different, but equivalent, measure. Kuwfman (2009) summarized the theoretical proof about the equivalent measures under tapering and convergence of the tapering estimators. Zhang and Du (2004) also proved almost sure convergence of Maximum Likelihood Estimation (MLE) when using a Matérn covariance model and several not overly stringent assumptions.

The covariance tapering method can also achieve a linear predictor that is nearly the same as the non-tapered covariance function. Moreover, the asymptotic mean squared error of prediction will converge to the minimal error within a fixed domain size and using tapering functions that are second order stationary and isotropic. This was also empirically demonstrated by Furrer et al. (2006) using data from the US climate data record.

3.3 Sparse matrix algorithms

The goal of introducing zeros into the covariance matrix is to be able to use sparse matrix algebra routines to more efficiently evaluate the likelihood (2.6). The most computationally demanding matrix operation in this evaluation is a factorization that facilities the computation of the determinant (or log-determinant) and inverse. One of the most efficient ways

to compute the inverse of a symmetric and positive definite matrix Σ is to use a Cholesky factorization $\Sigma = LL'$, where L is a lower triangular matrix square root, also called the Cholesky factor. Then solve a system of equations with one or more right-hand sides to find the inverse of matrix Σ. The log-determinant of Σ is simply 2 times the sum of the log diagonal elements of L.

As an aside, there is actually no need to calculate the full inverse of Σ to evaluate the likelihood (2.6). Rather, using the identity $\mathbf{z}'\Sigma^{-1}\mathbf{z} = (L^{-1}\mathbf{z})'(L^{-1}\mathbf{z})$ we can work exclusively with the Cholesky factorization. This approach also provides additional computational advantage because calculating the inverse of a triangular matrix and subsequent multiplication with $\mathbf{z}$ is more efficient compared to forming the *dense* Σ^{-1} and subsequent multiplication with $\mathbf{z}$.

The efficiency of the sparse matrix routines used to compute the Cholesky factorization of Σ depend on the number of non-zero values and where these values occur within the matrix. In the covariance tapering literature, analyses are conduced with matrices consisting of less than 1% non-zero values. Moreover, one notable attribute of a sparse matrix is that the Cholesky factor of a sparse matrix will also be sparse, provided locations of its non-zero values are properly (deliberately) permuted prior to factorization. There are several ways to permute these matrices such as the reverse Cuthill-Mckee or minimum-degree ordering. The analyses presented in this thesis use the `spam` (**Spa**rse **M**atrix algebra) `R` package. This package provides an interface to lower-level Sparskit FORTRAN routines written by Saad (1994). The `spam` package provides two permutation options – multiple-minimum degree and reverse Cuthill-Mckee from George and Lui (1981). The default is multiple-minimum degree ordering method, which has been shown to produce a permutation that provides *good*

results in most settings.

In addition to more efficient matrix operations, a sparse matrix storage format also allows for increased capacity to work with larger datasets. Sparse matrix storage formats only maintain the non-zero values and their position within the matrix to reduce both storage and arithmetic needs. For example, to store a $n \times n$ matrix with q non-zero elements in the standard row oriented format, a sparse matrix need only to store q real numbers and $q + n + 2$ integers.

Chapter 4

Numerical experiments using covariance tapering

In this chapter, I explore several aspects of the practical implementation of covariance tapering within the Bayesian modeling framework detailed in Chapter 2. Specifically, I use synthetic datasets to address guiding questions that focus on the balance between gain in computational efficiency and information loss. These guiding questions are

Question 1 What is the reduction in computation time when using covariance tapering?

Question 2 How much information is lost as a result of covariance tapering?

Question 3 How should the taper range of the compactly supported correlation function be chosen to effectively balance computational efficiency and information loss?

4.1 Analysis of computational efficiency

This simulation study focuses on *Question 1* which concerns the computational efficiency of the covariance tapering method compared to the analysis using the dense $n \times n$ covariance matrix. The synthetic data were generated at n random locations within a $[0, 100] \times [0, 100]$ unit domain. The synthetic observations at each location were drawn from the multivariate normal distribution $MVN(\mathbf{0}, \Sigma)$, where the dispersion matrix Σ was formed using an exponential correlation function with variance and spatial decay parameters $\sigma^2 = 1$ and $\phi = 0.04$, respectively. Note, because the exponential correlation function never equals zero, it is convenient to define an *effective spatial range*, which again is the distance at which the correlation drops below some value, e.g., 0.05. Then the effective spatial range for the synthetic data is 75 units i.e., $-\log(0.05)/\phi \approx 3/\phi$.

Given the domain size and fixed ϕ used to generate the data, the variables that influence the computational efficiency are *i*) choice of decay parameter ϕ_θ in taper correlation function K_θ (3.1) and *ii*) n, the dimension of the dataset. The larger the value of ϕ_θ, i.e., the shorter the taper range, the more zeros in the tapered covariance matrix and hence the shorter the computation time.

As detailed in Chapter 2, the computational bottleneck for iteratively evaluating the likelihood (2.6) is computing Σ^{-1}. Therefore, the first experiment explores the relationship between the computation time for Σ^{-1}, n, and the taper range $1/\phi_\theta$. Here, I consider values of n from 500 to 5000 units and taper ranges from 3 to 40. These taper ranges and resulting percent of non-zero values are summarized in Table 4.1.

The experiment results are shown in Figure 4.1. Here the plot axes are n and the computing time of the tapered Σ^{-1} in seconds. Each of the six taper ranges detailed in

Non-zero Percent	ϕ_θ	Taper Range
0.5%	0.300	3.333
1%	0.160	6.250
6%	0.060	16.667
10%	0.050	20.000
20%	0.030	33.333
34%	0.025	40.000

Table 4.1: Computation time experiment variables with the resulting percent of the spatial covariance matrix non-zero values given a taper range of 75.

Table 4.1 are plotted with different line symbols in Figure 4.1. The corresponding plot labels are prefaced with `spam` in the plot legend, which indicates they were generated using the sparse matrix algebra routines. Also plotted in this figure is the timing results which assume a dense Σ – labeled LAPACK in the plot legend. As the label indicates these results were generated using the LAPACK (**L**inear **A**lgebra **PACK**age; www.netlib.org/lapack) FORTRAN library of dense matrix routines.

For values of n greater than $\sim$1500, Figure 4.1 shows that the sparse matrix routines for inverting Σ only have an advantage over LAPACK's dense matrix routines for matrices with non-zero values of $\sim$1% or smaller. Referring to Table 4.1, this level of sparseness corresponds to a taper range of between $\sim$6.25-10 units. This result is consistent with `spam`'s documentation and similar routines found in other sparse matrix libraries. These routines are highly tuned to work with very sparse and well structured (i.e., optimally permuted) matrices. When the location of non-zero values is not well structured (i.e., poor structure is when non-zero values are randomly distributed across the matrix) and/or the percent of non-zero values is larger than $\sim$5 percent, then sparse matrix routines will generally not perform well compared to equivalent routines tuned for dense matrices.

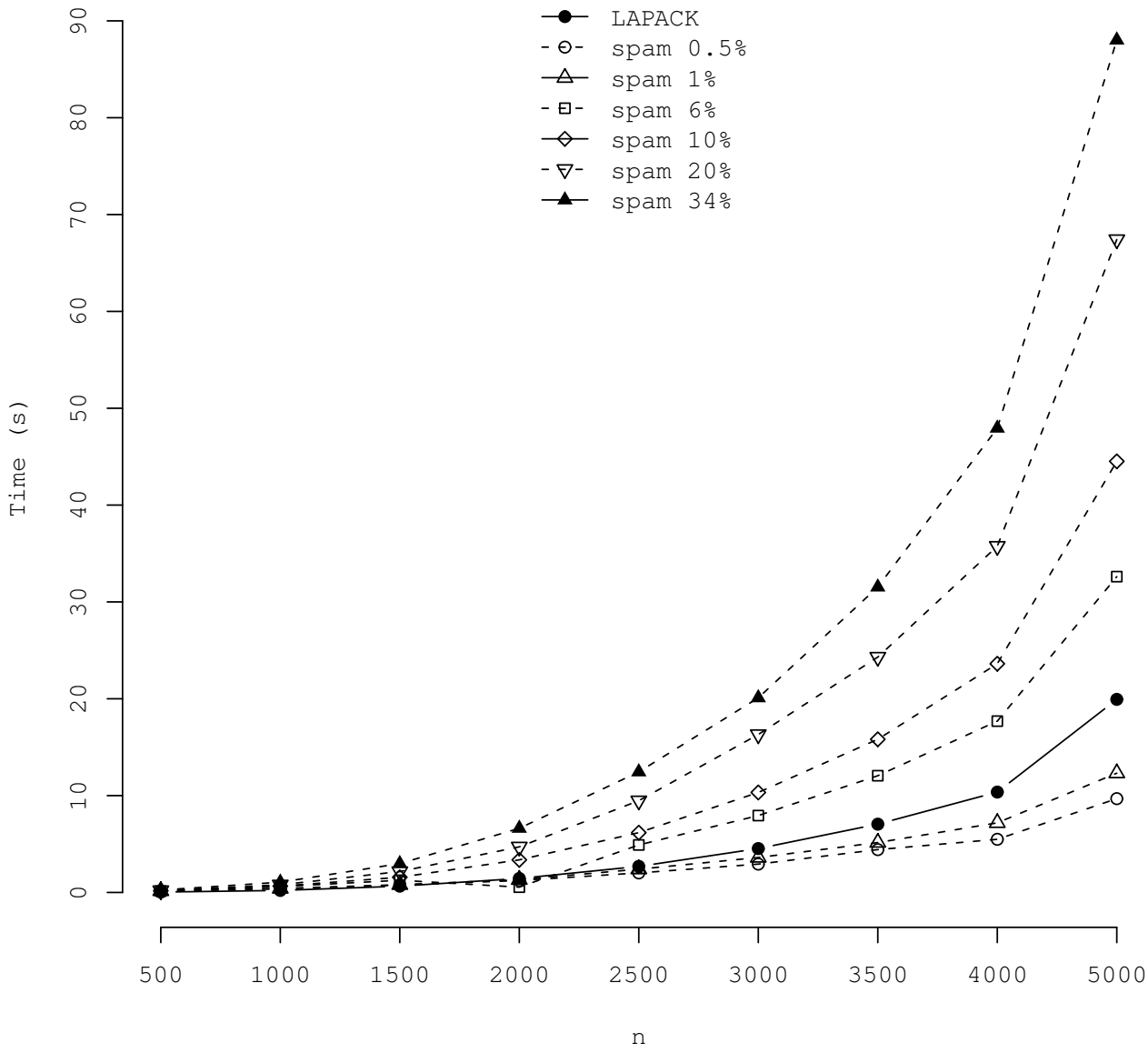

Figure 4.1: Dense and sparse matrix routine computing times for the inverse of tapered covariance matrices of varying dimension and number of non-zero values.

In the contemporary geostatistical literature, some of which is referenced in Chapter 2.3, a *large* dataset has a n in the tens or even hundreds of thousands. Typically with these large values of n the domain is also large. For many of the environmental outcome variables considered in forestry and ecology, the spatial range of dependence is small relative to the extent of the domain – as illustrated in Chapter 5. If this is indeed the case, then a sparseness of $\sim$2% or less might be realistic. Therefore the next experiment considers a smaller percent of non-zero values but much larger values of n. The results for this analysis are detailed in the four figures correspond to n of 10000, 14400, 19600, and 22500.

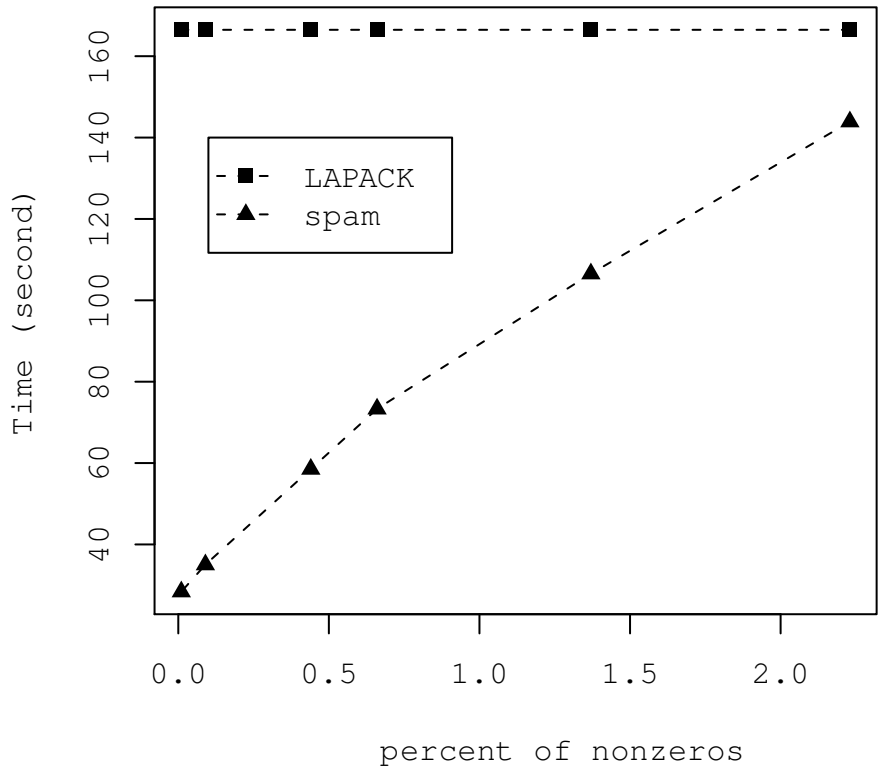

Figure 4.2: Dense and sparse matrix routine computing times for the inverse of tapered covariance matrices of large dimension and sparseness. Dimensions of each matrix $n = 10000$.

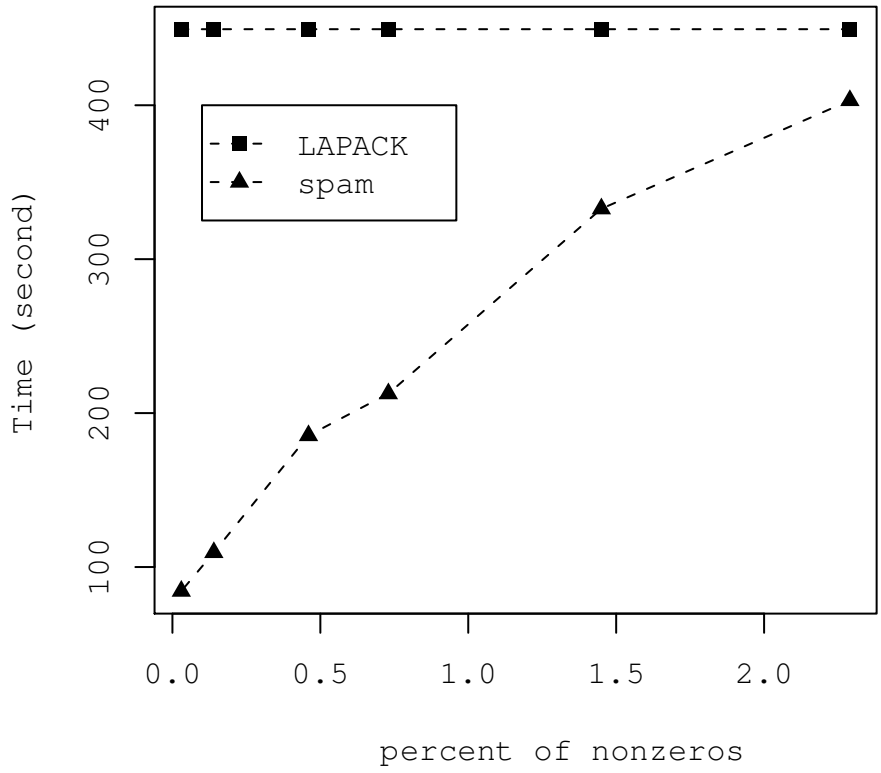

Figure 4.3: Dense and sparse matrix routine computing times for the inverse of tapered covariance matrices of large dimension and sparseness. Dimensions of each matrix $n = 14400$.

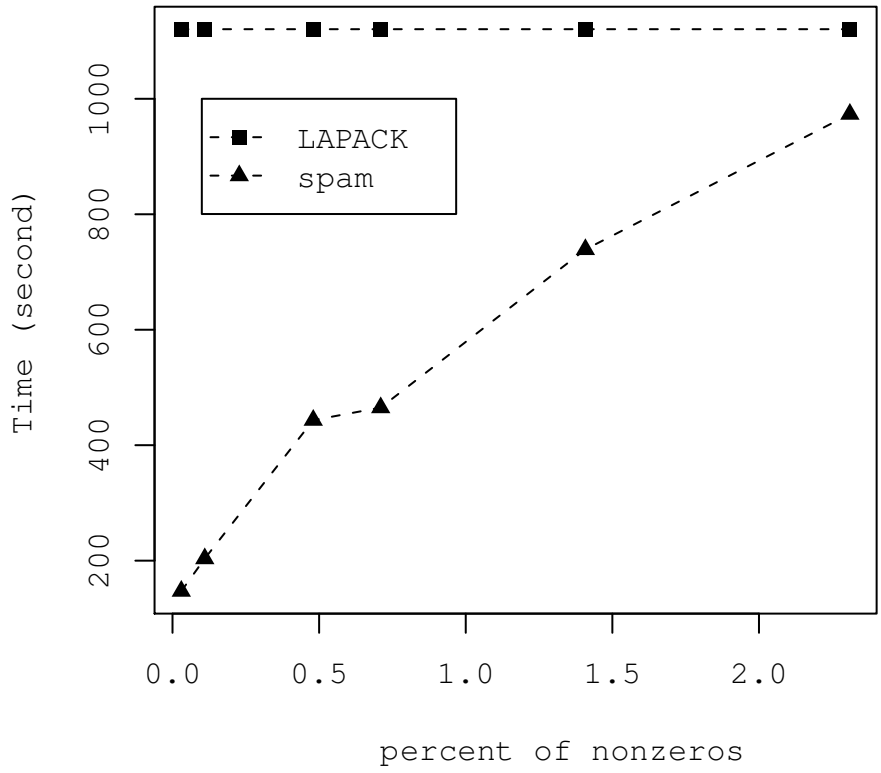

Figure 4.4: Dense and sparse matrix routine computing times for the inverse of tapered covariance matrices of large dimension and sparseness. Dimensions of each matrix $n = 19600$.

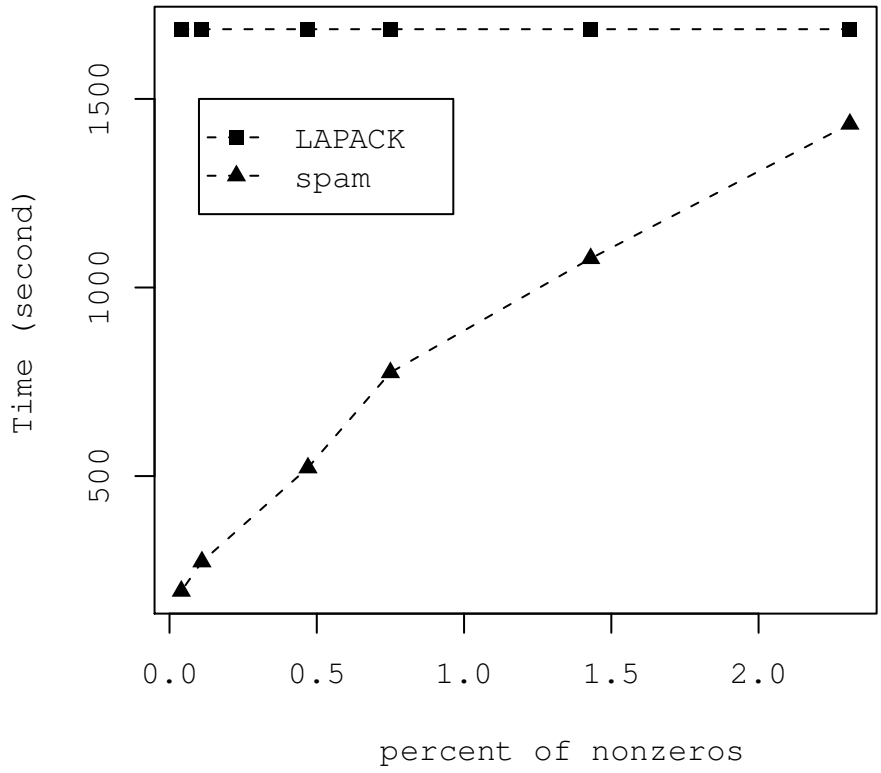

Figure 4.5: Dense and sparse matrix routine computing times for the inverse of tapered covariance matrices of large dimension and sparseness. Dimensions of each matrix $n = 22500$.

These sub-figures more clearly show when the sparse routines has the computational advantage. Here the vertical plot axis corresponds to the time in seconds it takes to calculate the matrix inverse and the horizontal axis denotes the percent non-zero values. As in the first experiment, beyond $\sim$2% non-zero values the dense routines have the advantage. However, for large matrices dominated by zero, the sparse routines offers substantial computational gains. These gains are on the order of several weeks in computing time when we consider this matrix operation must occur several thousand times to achieve adequate MCMC chain

sampling from parameters' posterior distribution.

In general, covariance tapering and the sparse matrix algorithms can achieve significant computational gain for large dimension sparse matrices with a small number of non-zero values. However, as always, "there is no free lunch." Effectively ignoring data that lie beyond the taper range results in varying degrees of information loss and can seriously compromise the usefulness of the modeling exercise. This information loss can be substantial when the taper range is chosen to be short relative to the range of dependence in the observed outcome variables.

Therefore, the next set of experiments attempt to quantify one important aspect of information loss. The results can help inform the choice for the degree of tapering to balance the computational gain and the model's predictive performance.

4.2 Analysis of information loss

This simulation study focuses on exploring *Question 2* and *Questions 3* which consider how the choice of taper range can effect the model's inferential performance. The results of this study are also used to suggest some general guidelines for setting a *defensible* taper range.

In the following experiment I quantify the information loss of model's (2.2) predictive performance given varying degrees of spatial dependence in the underlying data and covariance tapering. Here predictive performance is measured using Mean Square Prediction Error (MSPE), which is defined as

$$MSPE(\hat{\gamma}) = E[(\hat{\gamma} - \gamma)^2], \text{ where } \hat{\gamma} \text{ is the estimator of } \gamma. \tag{4.1}$$

In this experiment, I consider MSPE calculated using the vector of predicted values $\mathbf{y}_0$ (see equation 2.10 and subsequent development) and the vector of true values $\mathbf{y}$ i.e., $\hat{\gamma}$ and γ, respectively. Lower values of MSPE indicate greater predictive performance.

Synthetic datasets of size n equal 300, 600, and 1100 were generated using model (2.2). The data were produced using 10 different effective spatial ranges (i.e., $\sim 3/\phi$) in the set $1, 10, \ldots, 90, 100$. The additional model parameters were fixed at $\beta_0 = 0$, $\sigma^2 = 5$, and $\tau^2 = 1$. From each dataset, 100 locations were randomly chosen as a holdout set for model validation and the remaining locations were used to fit the model and make predictions. Figure 4.2 illustrates these holdout and modeling data for one of the synthetic datasets. Here too, the realization of $\mathbf{y}$ based on $\phi = 0.0375$ over the observed locations is depicted in Figure 4.2 and holdout locations in Figure 4.2. An example of the predicted $\mathbf{y}_0$ used to calculate MSPE in the following experiment is provided in Figure 4.2.

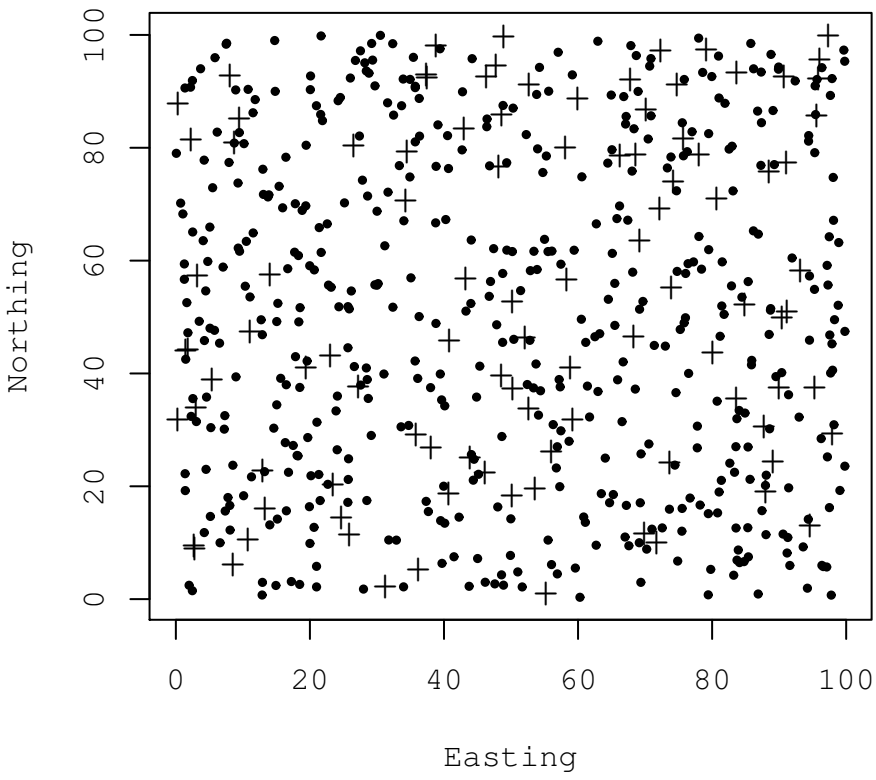

Figure 4.6: A synthetic dataset used to assess the model (2.2) predictive performance under different effective spatial ranges and degrees of covariance tapering. This figure shows locations of the observed ("•") and holdout ("+") locations.

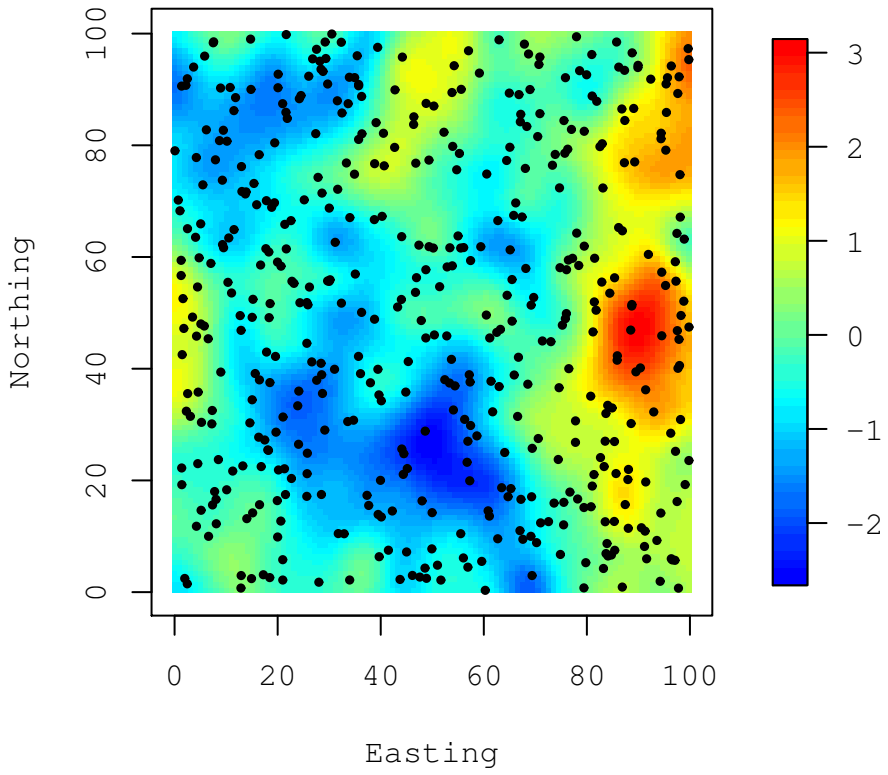

Figure 4.7: A synthetic dataset used to assess the model (2.2) predictive performance under different effective spatial ranges and degrees of covariance tapering. This figure shows realization of $\mathbf{y}$ over the observed locations.

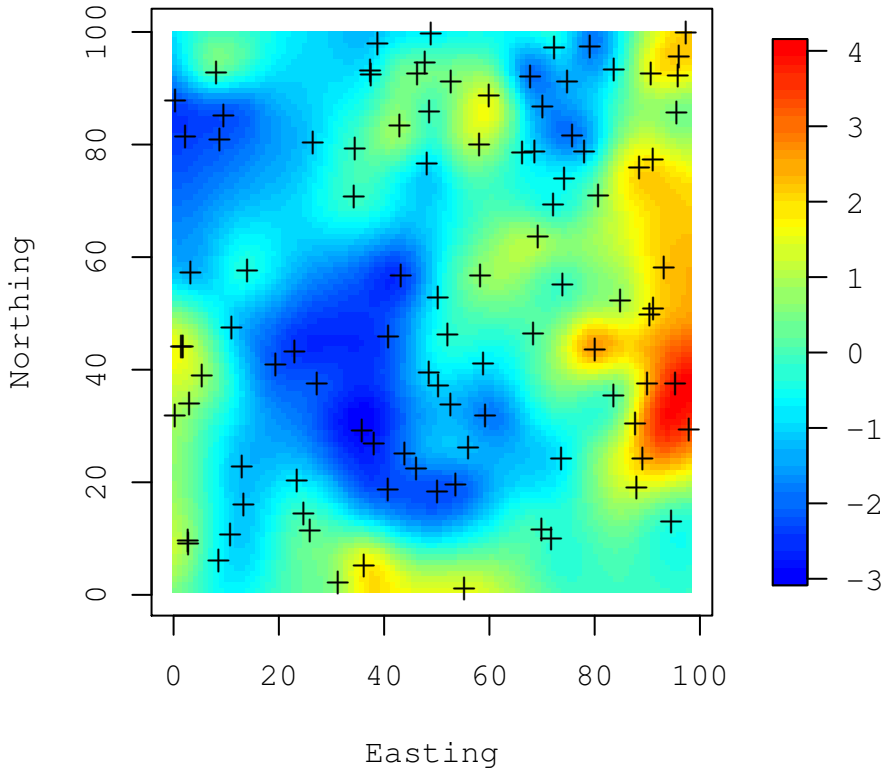

Figure 4.8: A synthetic dataset used to assess the model (2.2) predictive performance under different effective spatial ranges and degrees of covariance tapering. This figure shows realization of $\mathbf{y}_0$ over the holdout locations.

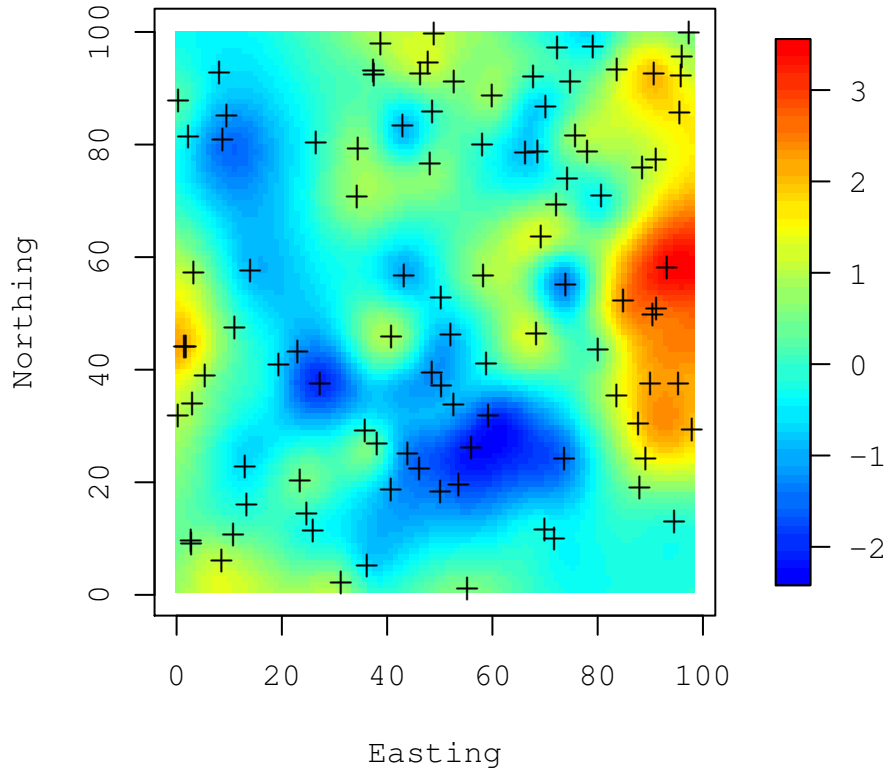

Figure 4.9: A synthetic dataset used to assess the model (2.2) predictive performance under different effective spatial ranges and degrees of covariance tapering. This figure shows model predicted values of $\mathbf{y}_0$.

Model predictions were based on 10 different taper ranges in the set $1, 10, \ldots, 90, 100$. Given the 10 effective spatial ranges used to generate the data, the analysis produced 100 values of MSPE for each of the three values of n. These results are provided in Figure 4.10, 4.11 and 4.12. Here, each surface is interpolated over the 100 combinations of ϕ and ϕ_θ.

In Figure 4.10, 4.11 and 4.12, the vertical axis of the sub-figures index the taper range used in the covariance tapering and the horizontal axis is the effective spatial range of dependence in the synthetic dataset. Recall, Figure 3.4 shows the original covariance curve

is close to the tapered covariance curve when the ratio of taper range and effective range is ∼300%. Therefore, Figure 4.12 depicts MSPE corresponding to ratios of taper range to effective range less than 300% and the sample size n. These surfaces highlight several useful results:

1. According to Figure 4.10, 4.11 and 4.12, the vertical band of high MSPE along the sub-figures' left margin is a result of poor prediction due to short effective spatial ranges in the data itself. That is, because the range of spatial dependence in the data is short, prediction is generally poor, regardless of the taper range. Prediction would be poor in this region even under the full, non-tapered, covariance matrix. Importantly, however, the MSPE in this region decrease as the n increases. This is because there is simply more locations within a shorter spatial distances that can be used for prediction.

2. In each Figure 4.10, 4.11 and 4.12, beyond an effective spatial range of ∼30, the horizontal band of high MSPE along the sub-figures' lower margin is the result of information loss from covariance tapering. Here, the taper range is short relative to the range of dependence in the data and hence prediction is based on only a fraction of the *useful* information in the dataset.

3. For Figure 4.10, 4.11 and 4.12, at effective spatial ranges of ∼40, there is negligible difference in MSPE based on taper ranges of ∼20 and greater. Further, as n increases this lower bound on the taper range decrease. For instance, at $n = 1100$, one might use a taper range of ∼10 without much impact on predictive performance. This trend also relates to the greater amount of information within proximity to prediction locations which is afforded by the larger dataset.

4. In Figure 4.13, when n is small, a taper range to effective range ratio less than $\sim$0.5 results in high MSPE. This is because too much information is lost when the taper range is small compared to the effective range. As n increases, the ratio can become smaller while still resulting in the same MSPE. This is because there is more information (i.e., observed locations) within the short taper range from which to borrow information to use for prediction. Interestingly, when n is small and the ratio is large (i.e., the original covariance function is not changed much) predictive performance is actually worse than when the ratio is near 1 – tapering actually improves prediction even though the data was not generated from a tapered correlation function. This effect decreases at n increases.

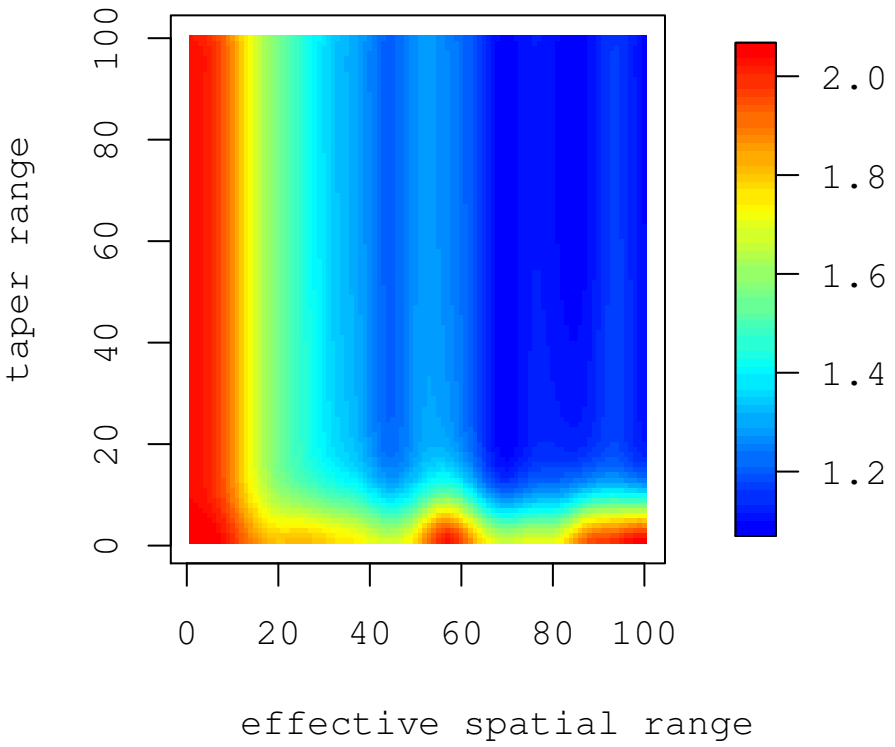

Figure 4.10: Surfaces of Mean Squared Prediction Error (MSPE) for combinations of ϕ, ϕ_θ, and dataset size is 300.

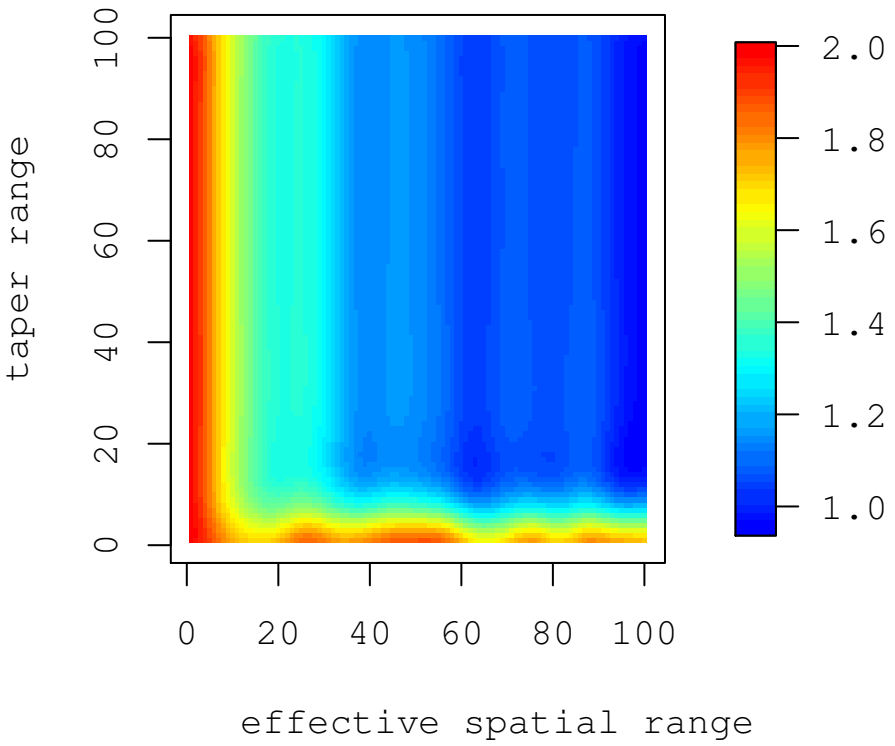

Figure 4.11: Surfaces of Mean Squared Prediction Error (MSPE) for combinations of ϕ, ϕ_θ, and dataset size is 600

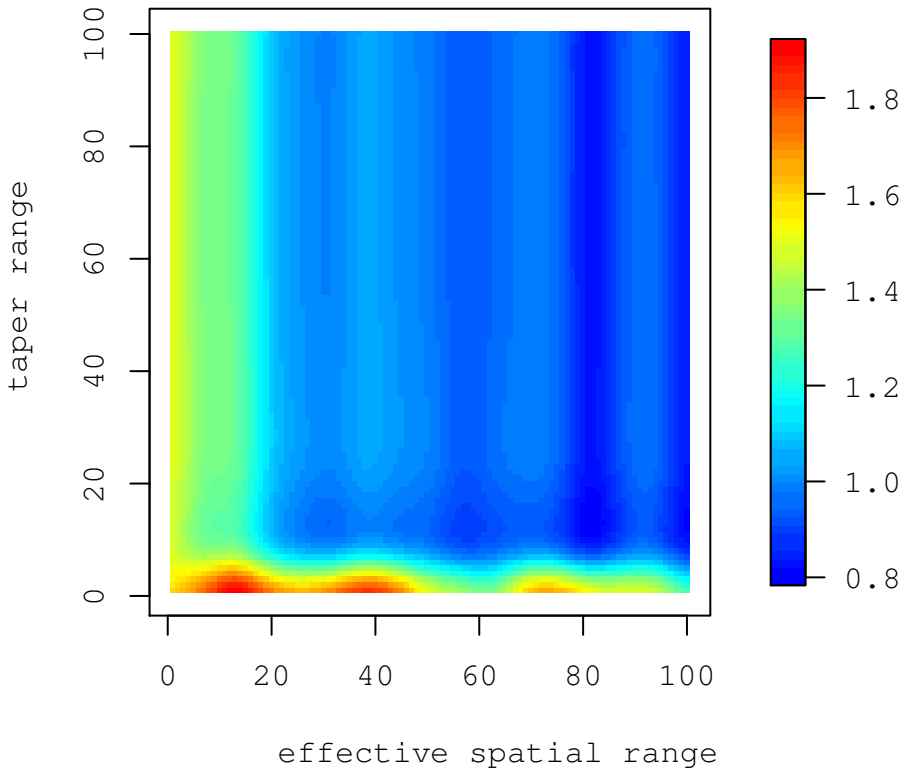

Figure 4.12: Surfaces of Mean Squared Prediction Error (MSPE) for combinations of ϕ, ϕ_θ, and dataset size is 1100.

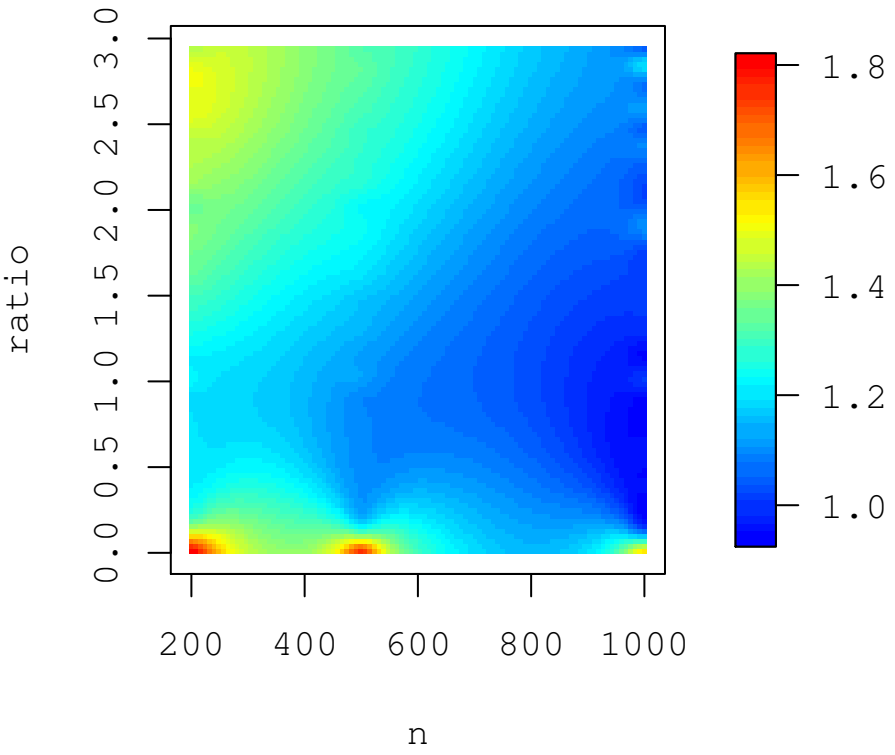

Figure 4.13: Surfaces of Mean Squared Prediction Error (MSPE) for combinations of ϕ, ϕ_θ with different ratio of taper range and effective range.

The contour lines on Figure 4.10, 4.11 and 4.12 also provide some guidance on how to specify the taper range given the effective spatial range of the data. In practice, and as demonstrated in Chapter 5, estimates of this effective spatial range can be obtained from an empirical semivariogram fit to the non-spatial model's residual values. This experiment also helps to identify the taper range that balances computational efficiency and information loss. Specifically, one would chose the smallest possible taper range given some predefined limit of acceptable information loss. The allowable taper range to effective range ratio is best explored using Figure 4.13. This surface suggests a defensible taper range can be chosen by keeping the ratio between $\sim 0.5 - 1$ for small to moderately sized datasets and, given this allowable ratio decreases as n increases, large datasets could be modeled using much smaller ratios without compromising predictive performance. Given the results of this synthetic data analysis and the MSPE objective function, one might safely consider taper ranges at the lower bound of the blue colored zone in the surface in Figure 4.10, 4.11 and 4.12. Although, the taper range and effective range ratio as well as n are important variables, predictive performance will depending on additional dataset characteristics. Therefore, I recommend using an exploratory analysis, similar to the one conduced here, when choosing a taper range.

This experiment only considered predictive inferences based on the MSPE criterion. Other criterion could be considered and might reveal different aspects of information loss. Gneiting and Raftery (2007) detail several such criterion that might be considered. Further, this analysis only focused on predictive performance. Depending on the inferential focus, one might consider looking at information loss in the model parameters, such as the regression slopes and/or spatial and residual variances.

Chapter 5

Illustrative analyses of forest complexity and biomass

As described in Chapter 1, a central objective of this thesis is to define modeling methods and create software to help meet the research objectives of a larger project entitled "Forest complexity in the Lake States: Implications for carbon storage." The goal of the larger project is to identify forest management opportunities that might aid in climate change mitigation and adaptation. The premise of this project is that forest managers are seeking strategies to create stands that can both adapt to new climatic conditions and help mitigate increases in atmospheric CO_2. One approach to adaptation involves maximizing forest complexity, which can be characterized as both compositional complexity in species diversity and structural complexity in tree sizes. Enhanced forest complexity may contribute to climate change adaptation because trees of different species and/or sizes may respond differently to weather fluctuations or disturbances. However, forests must also provide mitigation of climate change by maintaining carbon storage and/or sequestration and forest complexity

may be negatively related to carbon storage or sequestration because more complex forests may contain a higher proportion of smaller and/or less productivity trees. Although the importance of forest complexity for climate change adaptation and mitigation is clear, few studies have examined how land ownership influences forest complexity across large areas or directly related complexity to carbon cycling.

In the following sections, I present several illustrative analyses that will help guide the more thorough exploration of the larger study's research objectives. These analyses are a subset of the models that will be explored in the larger study. The analyses show the advantage of using a hierarchical spatial model and compliment the synthetic data analyses in Chapter 4 by providing additional insights into covariance tapering.

5.1 Data

The data considered in the following analyses was collected by the USDA Forest Service Forest Inventory and Analysis (FIA) program. This program has established field plots in permanent locations by using an equal probability sampling design (Bechtold and Patterson 2005). In each forested plot, they calculated total tree biomass based on established equations that use tree-level variables, such as species, diameter at breast height (DBH), and total height. Additional plot-level variables such as stand age and ownership were also recorded. The dataset consists of the 8670 FIA database observations measured between 1999-2006 that were designated as completely forested and fell within Minnesota, Wisconsin, and Michigan. Here too, I only considered plots that fell within a single private ownership category. The ownership categories considered are defined in Table 5.1.

Table 5.1: USDA Forest Service Forest Inventory and Analysis private ownership categories.

Variable name	Description
OWN-41	Corporate
OWN-42	Non-governmental conservation/Natural resources organization
OWN-43	Unincorporated local partnership/association/club
OWN-44	Native American (Indian)
OWN-45	Individual

5.2 Analyses and results

The subsequent analyses use the five forest ownership categories defined in Table 5.1 to explain the variability in an outcome variable of interest. This analysis of variance is conducted using both the spatial model (2.2) and its non-spatial equivalent (i.e., $\mathbf{w} = \mathbf{0}$). For all analyses, OWN-41 was arbitrarily chosen as the baseline category. This choice should obviously change to accommodate specific hypotheses to be tested in the larger project. Given the computationally prohibitive size of the dataset, I again taper the spatial random effects covariance matrix. I consider several different taper ranges for each analysis to assess the robustness of the results.

In addition to comparing the parameter estimates, the popular *Deviance Information Criterion*, or DIC is used to measure the difference between models and lower DIC indicates *better* fit. Moreover, an empirical semivariogram was fit to the non-spatial models' residuals. These semivariograms served to help guide the choice prior distributions for the model parameters and the taper range.

To complete the model specification, the regression slope parameters were assumed to have the customary *flat* prior, the variance parameters were assigned inverse-Gamma $IG(shape, scale)$ where the shape hyperparameter was set to 2 and the scale was taken as the semivariogram's partial sill and nugget value for σ^2 and τ^2, respectively. Note, with

a shape of 2, the mean of the IG equals the scale value and the distribution has infinite variance, see, e.g., Gelman et al., (2004) for the IG parametrization. Again assuming an exponential spatial correlation function, the prior support for the spatial decay parameter ϕ followed a Uniform $U(0.00025,\ 0.03)$, which corresponds to 100 to 12,000 m.

Three MCMC chains were run for 1,500 iterations each. Convergence was diagnosed using the `CODA` package in `R` by monitoring mixing of chains and the Gelman-Rubin statistic (Gelman and Rubin 1992). Satisfactory convergence was diagnosed within 1,000 iterations for all models and posterior inference was based on a post-burn-in sample of 1,500 iterations (500 from each chain).

5.2.1 Standard deviation of DBH and ownership

The standard deviation of DBH (SD-DBH) is one measure of a forest stand's complexity. Stands with large SD-DBH will have a greater distribution of diameters. For example, stands with uneven age structure and/or multiple tree species with different shade tolerances will likely exhibit SD-DBH larger than stands that consist of trees within a single age class. One would expect single species and even-aged plantations to have some of the lowest SD-DBH within the dataset. For this analysis the variability in the outcome variable SD-DBH is described by the ownership categories defined in Table 5.1. The distribution of the outcome variable was approximately normal and no transformation was considered.

Table 5.2 summarizes parameters' posterior distribution estimated using the non-spatial model. Here, β_0 is the mean standard deviation of DBH for the baseline ownership category. Subsequent β estimates are the offset from the baseline mean. Ownership categories with 95% credible intervals that do not include zero are statistically different from the baseline at

the 0.05 level. Given this definition, these estimates suggest that the four ownerships have statistically higher average SD-DBH than the baseline OWN-41 category.

This non-spatial regression model is adequate in the absence of extraneous structured variation, beyond what is explained by the ownership. However, when observations are spatially indexed, we might expect similar outcomes in proximate locations, possibly resulting from common environmental conditions, disturbance regimes, and/or management practices. As illustrated in Chapter 2.2.1, ignoring this spatial dependence can result in falsely precise estimates of model parameters and erroneous predictions. Hoeting (2009) offers a nice discussion on consequences of not meeting the assumption of uncorrelated model residuals. As detailed in Chapter 2, a common solution to spatial dependence among the residuals is to include a spatially-varying model intercept through the addition of a spatial random effect. Beyond helping to ensure the statistical validity of the model, the addition of a spatial random effect to the intercept allows for partitioning of residual uncertainty into a spatial and non-spatial component which can improve model fit.

Table 5.2: Non-spatial model's parameter credible intervals for standard deviation of DBH and forest ownership. Parameters that are statistically significant at the 0.05 level are indicated with an "*".

Parameter	2.5%	25%	50%	75%	97.5%
β_0^*	3.22	3.27	3.30	3.33	3.38
β_{OWN-42}^*	0.09	0.39	0.54	0.70	0.98
β_{OWN-43}^*	0.13	0.29	0.39	0.47	0.63
β_{OWN-44}^*	0.40	0.54	0.62	0.69	0.83
β_{OWN-45}^*	0.28	0.34	0.37	0.40	0.46
τ^2	2.39	2.44	2.46	2.49	2.54

Figure 5.1 shows the semivariogram of the non-spatial model's residuals. Here, the long effective spatial range of ∼85 km and large size of the partial sill compared to the nugget,

suggests there is substantial residual dependence. This encourages further analysis using the full spatial model (2.2). Again, given the size of this dataset, I tapered the random effects covariance matrix to reduce the computational burden. As in the synthetic data analyses, I again used the spherical compactly supported correlation function (3.1). I considered taper ranges (i.e., $1/\phi_\theta$) of 30%, 50% and 100% of the estimated effective spatial range $\sim$85 km, which correspond to 26, 42, and 85 km, respectively. These taper ranges result in spatial covariance matrices with 0.5%, 1%, and 2% non-zero values. Give the relatively large dimension of the dataset, these levels of sparseness result in substantial computational gains over the dense matrix computations – on the order of days.

The model parameter estimates for the three taper ranges are presented in Table 5.3. At a taper range of 26 km, the spatial model produces regression slope parameter estimates that do not differ in sign or significances from those of the non-spatial model. Only considering dependence up to 26 km would likely leave a substantial portion of residual structured dependence intact that could negatively impact the model parameter estimates. At a taper range of 42 and then 85, this residual spatial dependence is substantially reduced. With the model assumptions better met, the regression slope parameters associated with OWN-42 and OWN-43 become non-significant. This could be a true signal or perhaps the result of an insufficient sample size for these two ownership categories. OWN-42 and OWN-43 have the lowest number of observations among the ownerships. The addition of spatial random effects essentially reduces the *effective* number of observations (by acknowledging spatially redundant information) and hence the regression parameters may shrink toward zero. Figure 5.2, 5.3, 5.4, 5.5 and 5.6 summary the posterior distributions of non-spatial and spatial model with different taper ranges. It is obvious that including the spatial random

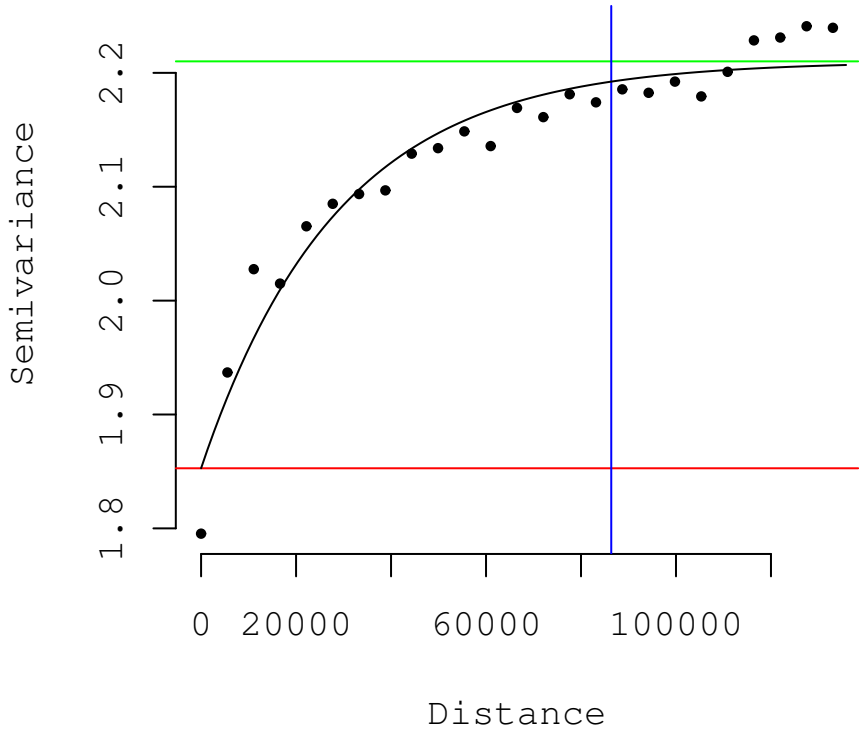

Figure 5.1: Empirical semivariogram of residuals from the non-spatial version of model (2.2) for standard deviation of DBH and forest ownership along with the maximum likelihood estimates of the exponential correlation function's nugget (lower horizontal line), sill (upper horizontal line), and effective spatial range (vertical line).

effects influences coefficient estimates, even changes their significance under 95% confidence level.

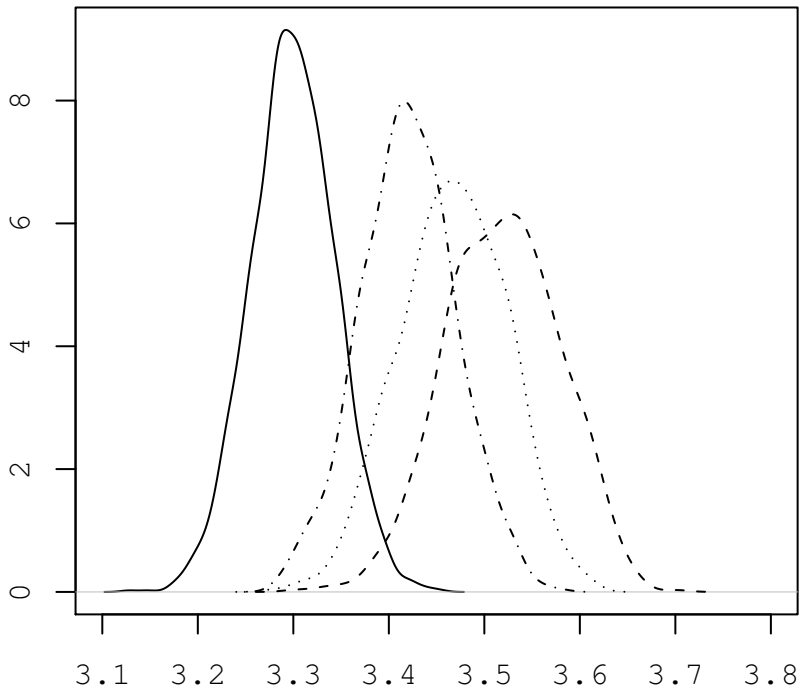

Figure 5.2: Illustration of estimated parameter's posterior distribution for intercept in non-spatial model and spatial model for for standard deviation of DBH and forest ownership: solid line represents the non-spatial model; dashed line, dotted line, and mixture line represents spatial models with taper ranges of 100%, 50% and 30% of effective range, respectively.

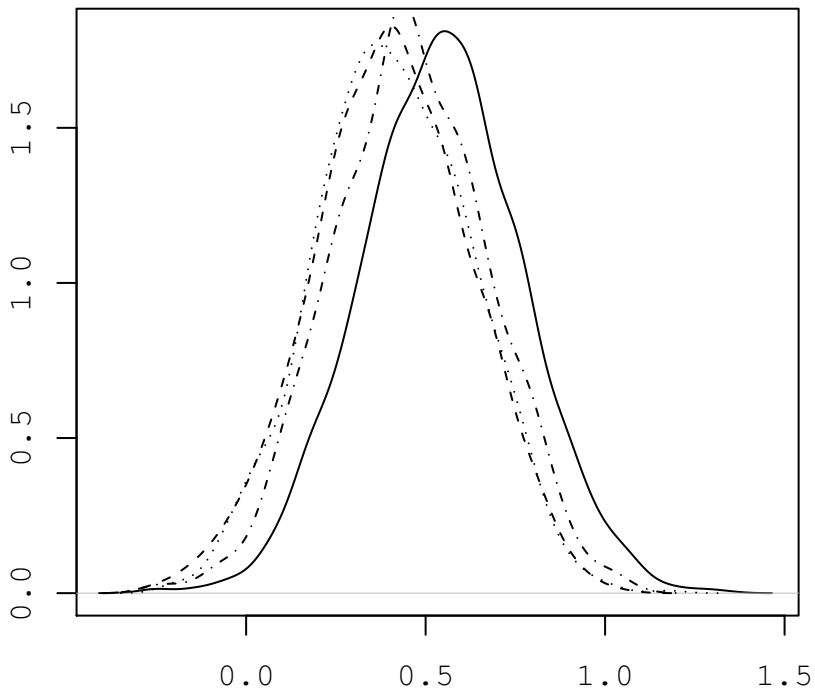

Figure 5.3: Illustration of estimated parameter's posterior distribution for OWN-42 in non-spatial model and spatial model for for standard deviation of DBH and forest ownership: solid line represents the non-spatial model; dashed line, dotted line, and mixture line represents spatial models with taper ranges of 100%, 50% and 30% of effective range, respectively.

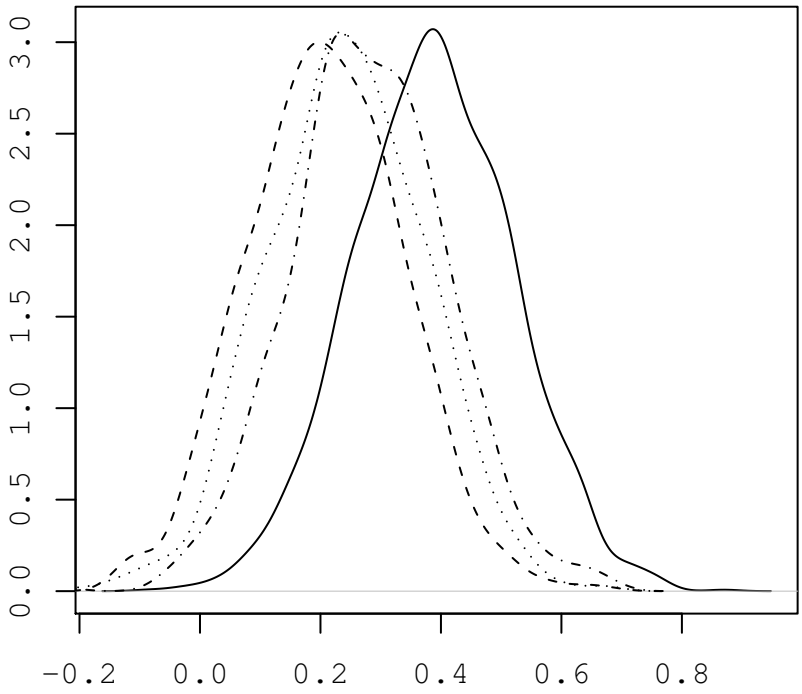

Figure 5.4: Illustration of estimated parameter's posterior distribution for OWN-43 in non-spatial model and spatial model for for standard deviation of DBH and forest ownership: solid line represents the non-spatial model; dashed line, dotted line, and mixture line represents spatial models with taper ranges of 100%, 50% and 30% of effective range, respectively.

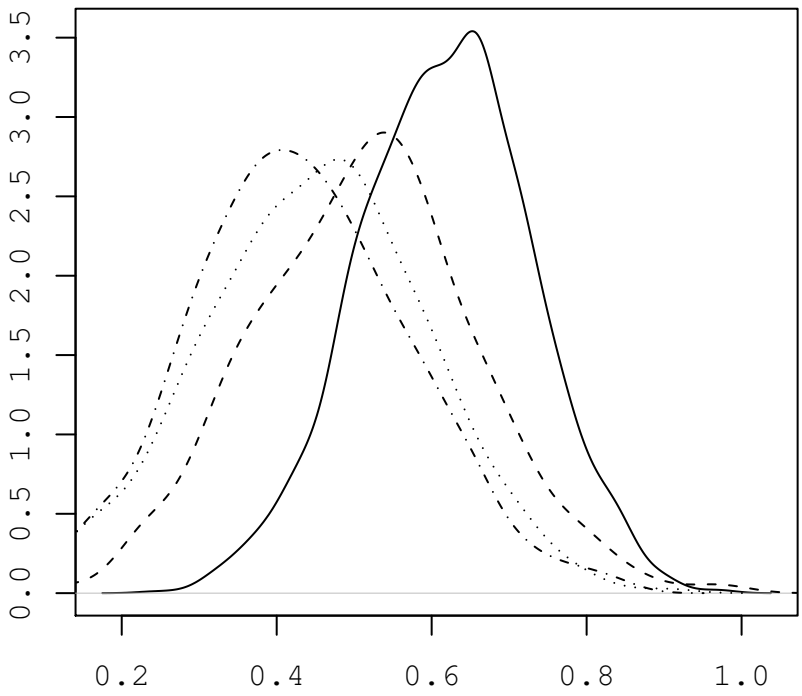

Figure 5.5: Illustration of estimated parameter's posterior distribution for OWN-44 in non-spatial model and spatial model for for standard deviation of DBH and forest ownership: solid line represents the non-spatial model; dashed line, dotted line, and mixture line represents spatial models with taper ranges of 100%, 50% and 30% of effective range, respectively.

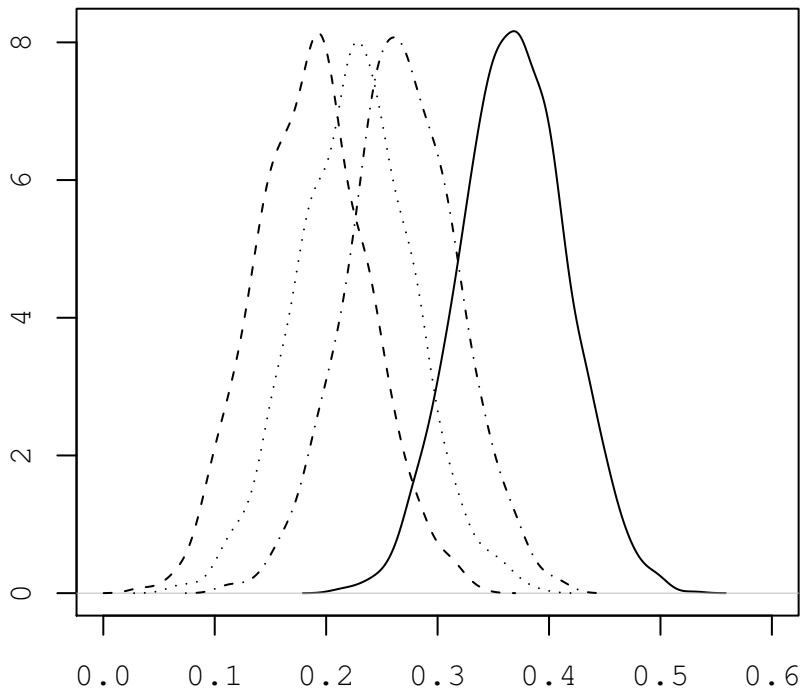

Figure 5.6: Illustration of estimated parameter's posterior distribution for OWN-45 in non-spatial model and spatial model for for standard deviation of DBH and forest ownership: solid line represents the non-spatial model; dashed line, dotted line, and mixture line represents spatial models with taper ranges of 100%, 50% and 30% of effective range, respectively.

The DIC model fit criterion for the four models is summarized in Table 5.4. Compared with the non-spatial model, the approximate decrease in DIC of 1000 units suggest that the addition of the spatial random effects greatly improve model fit. The DIC values among the spatial models are comparable.

Table 5.3: Spatial model's parameter credible intervals for standard deviation of DBH and forest ownership. Parameters that are statistically significant at the 0.05 level are indicated with an "*".

	2.5%	25%	50%	75%	97.5%
Taper range equals 26 km (30% of effective spatial range)					
β_0^*	3.32	3.39	3.42	3.46	3.52
β_{OWN-42}^*	0.05	0.31	0.46	0.61	0.87
β_{OWN-32}^*	0.04	0.20	0.28	0.36	0.53
β_{OWN-44}^*	0.15	0.33	0.43	0.52	0.71
β_{OWN-45}^*	0.17	0.23	0.27	0.30	0.37
ϕ	3.062e-05	3.138e-05	3.287e-05	3.52e-05	4.0e-05
σ^2	0.41	0.48	0.52	0.54	0.60
τ^2	1.86	1.91	1.94	1.97	2.03
Taper range equals 42 km (50% of effective spatial range)					
β_0^*	3.36	3.43	3.47	3.51	3.57
β_{OWN-42}^*	0.01	0.27	0.41	0.56	0.83
β_{OWN-43}	-0.01	0.15	0.24	0.33	0.50
β_{OWN-44}^*	0.16	0.35	0.45	0.54	0.72
β_{OWN-45}^*	0.12	0.19	0.23	0.26	0.33
ϕ	3.05e-05	3.1e-05	3.18e-05	3.28e-05	3.63e-05
σ^2	0.39	0.42	0.45	0.47	0.52
τ^2	1.93	1.97	1.99	2.02	2.07
Taper range equals 85 km (100% of effective spatial range)					
β_0^*	3.40	3.48	3.52	3.56	3.63
β_{OWN-42}	-0.03	0.26	0.40	0.55	0.82
β_{OWN-43}	-0.04	0.12	0.21	0.30	0.46
β_{OWN-44}^*	0.24	0.41	0.52	0.60	0.79
β_{OWN-45}^*	0.09	0.15	0.19	0.22	0.29
ϕ	3.03e-05	3.05e-05	3.11e-05	3.16e-05	3.31e-05
σ^2	0.33	0.36	0.39	0.41	0.46
τ^2	1.97	2.02	2.04	2.07	2.11

Model	pD	DIC
Non-spatial	5	16486
Spatial (taper range equals 26 km)	1236	15535
Spatial (taper range equals 43 km)	1186	15967
Spatial (taper range equals 85 km)	966	15908

Table 5.4: Model comparison using Deviance Information Criterion for standard deviation of DBH and forest ownership.

5.2.2 Biomass and ownership

This analysis focuses on explaining the variability in the outcome variable metric tons of live standing biomass per hectare (BIO) using the five ownership categories defined in Table 5.1. The outcome variable was log transformed to better approximate a normal distribution.

Table 5.5 summarizes parameters' posterior distribution estimated using the non-spatial model. Again, those ownership categories with 95% credible intervals that do not include zero are considered statistically significant at the 0.05 level. Given this definition, these estimates suggest that OWN-42 and OWN-44 have statistically higher average BIO than the baseline OWN-41 category.

Table 5.5: Non-spatial model's parameter credible intervals for biomass and forest ownership. Parameters that are statistically significant at the 0.05 level are indicated with an "*".

	2.5%	25%	50%	75%	97.5%
β_0^*	7.79	7.89	7.95	8.00	8.11
β_{OWN-42}^*	0.21	0.74	1.02	1.36	1.85
β_{OWN-43}	-0.04	0.29	0.46	0.63	0.97
β_{OWN-44}^*	0.46	0.73	0.88	1.03	1.29
β_{OWN-45}	-0.26	-0.14	-0.09	-0.02	0.09
τ^2	9.09	9.25	9.34	9.44	9.63

The semivariogram of the non-spatial model's residuals is given in Figure 5.7. Here, similar to the analysis for SD-DBH, the semivariogram shows clear spatial structure among the residuals which indicates a violation of the model assumptions. This violation could produce misleading inference about the significance of the regression slope parameters and hence erroneous statements about the average volume of biomass compared to the baseline ownership. The vertical line on the semivariogram indicates an estimated effective spatial range of $\sim$75 km.

Given the results of the semivariogram analysis, I am again encouraged to explore a model

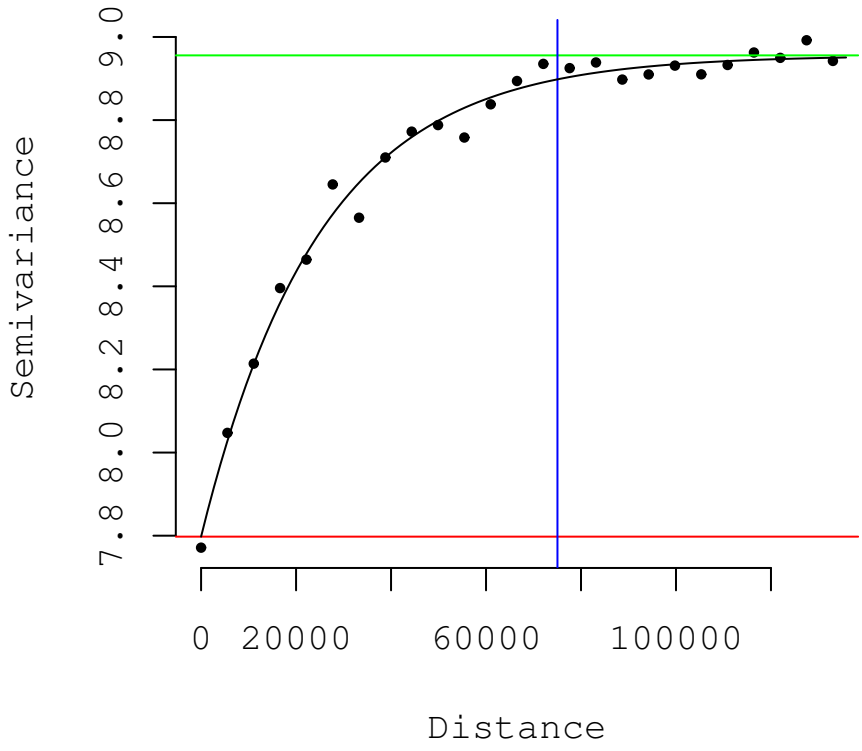

Figure 5.7: Empirical semivariogram of residuals from the non-spatial version of model (2.2) for biomass and forest ownership along with the maximum likelihood estimates of the exponential correlation function's nugget (lower horizontal line), sill (upper horizontal line), and effective spatial range (vertical line).

that includes spatial random effects which arise from a multivariate normal distribution with a tapered covariance matrix. I considered taper ranges of 50% and 100% of the estimated effective spatial range i.e., 38 km and 75 km, respectively. Both of these taper ranges produce a sparse covariance matrix with less than 3% non-zero values. Parameter estimates for these two spatial models are summarized in Table 5.6. Contrary to the non-spatial model results, both spatial models suggest the average biomass in ownership category 42 (i.e., OWN-42) does not differ from the baseline ownership. Figure 5.8, 5.9, 5.10, 5.11 and 5.12 represent

the correspond posterior distributions of the non-spatial and spatial model. Here too, the addition of the spatial random effects improves model fit as reflected by the lower DIC values in Table 5.7.

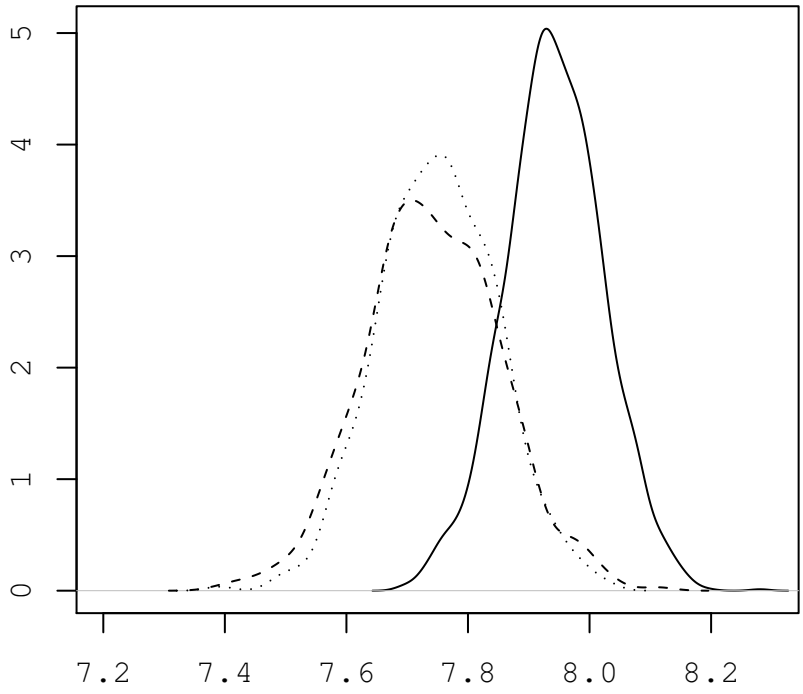

Figure 5.8: Illustration of estimated parameter for intercept' posterior distribution in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line and dotted line represents spatial models with taper ranges are 100%, and 50% of effective range, respectively.

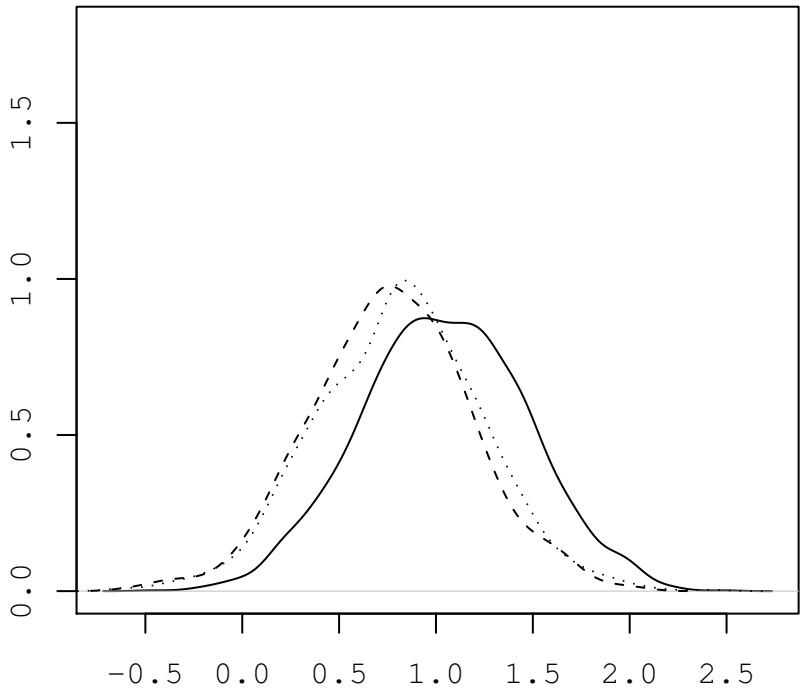

Figure 5.9: Illustration of estimated parameter's posterior distribution for OWN-42 in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line and dotted line represents spatial models with taper ranges are 100%, and 50% of effective range, respectively.

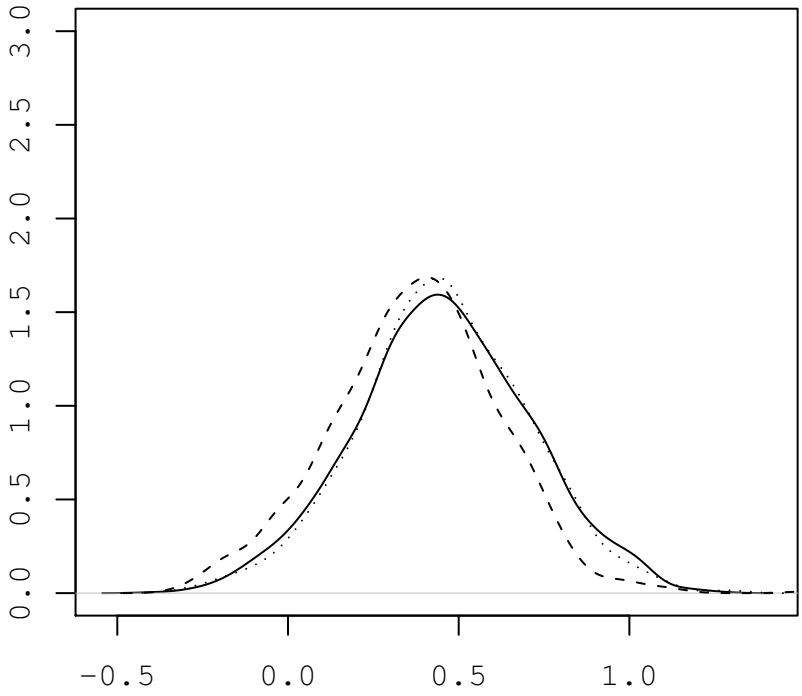

Figure 5.10: Illustration of estimated parameter's posterior distribution for OWN-43 in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line and dotted line represents spatial models with taper ranges are 100%, and 50% of effective range, respectively.

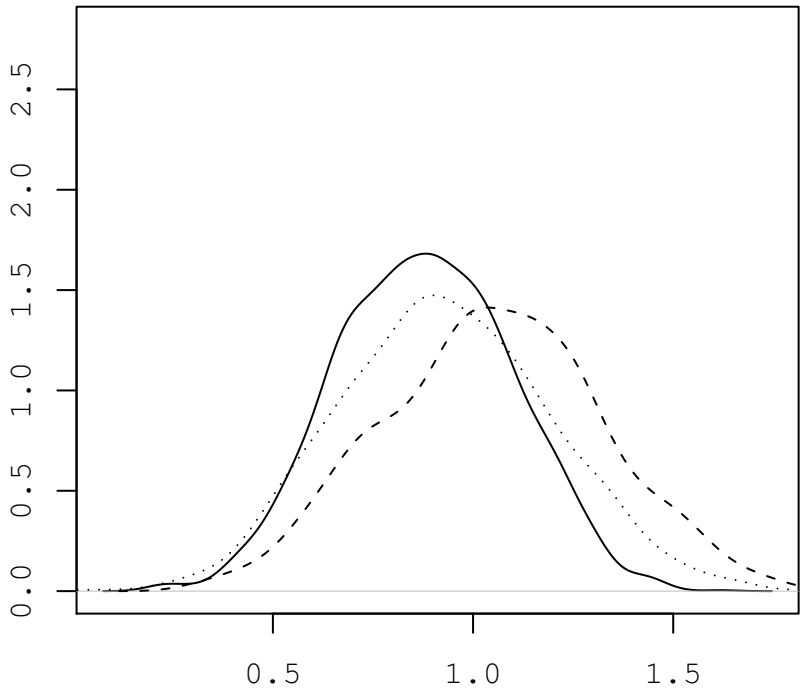

Figure 5.11: Illustration of estimated parameter's posterior distribution for OWN-44 in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line and dotted line represents spatial models with taper ranges are 100%, and 50% of effective range, respectively.

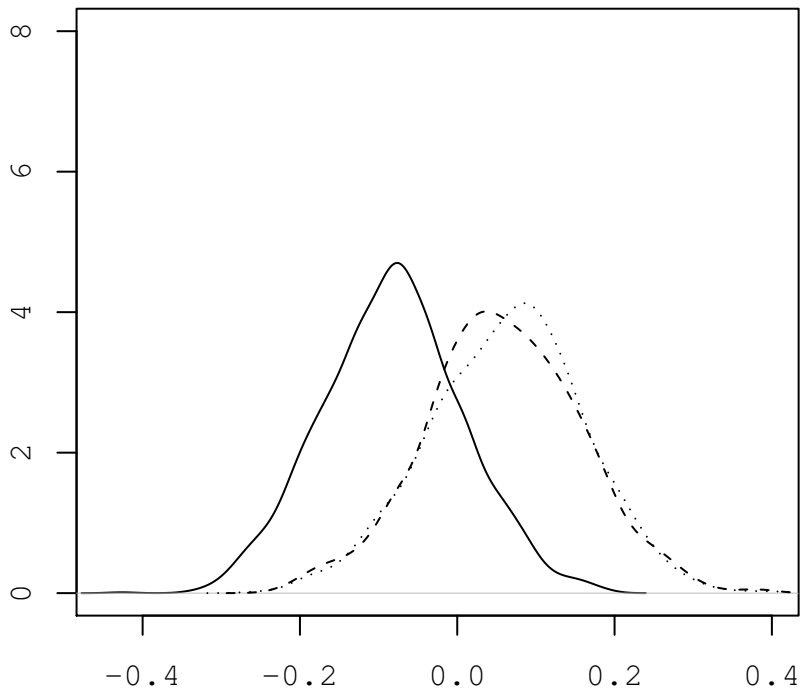

Figure 5.12: Illustration of estimated parameter's posterior distribution for OWN-45 in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line and dotted line represents spatial models with taper ranges are 100%, and 50% of effective range, respectively.

Table 5.6: Spatial model's parameter credible intervals for biomass and forest ownership. Parameters that are statistically significant at the 0.05 level are indicated with an "*".

	2.5%	25%	50%	75%	97.5%
Taper range equals 38 km (50% of effective spatial range)					
β_0^*	7.55	7.68	7.74	7.81	7.94
β_{OWN-42}	-0.01	0.51	0.80	1.10	1.64
β_{OWN-43}	-0.04	0.28	0.45	0.61	0.94
β_{OWN-44}^*	0.38	0.73	0.92	1.10	1.44
β_{OWN-45}	-0.12	0.00	0.07	0.13	0.26
ϕ	3.07e-05	3.13e-05	3.20e-05	3.31e-05	3.59e-05
σ^2	1.25	1.45	1.54	1.62	1.74
τ^2	7.41	7.55	7.65	7.77	8.00
Taper range equals 75 km (100% of effective spatial range)					
β_0^*	7.54	7.67	7.74	7.82	7.97
β_{OWN-42}	-0.04	0.47	0.75	1.03	1.60
β_{OWN-43}	-0.12	0.21	0.38	0.53	0.82
β_{OWN-44}^*	0.50	0.87	1.05	1.24	1.62
β_{OWN-45}	-0.15	-0.01	0.05	0.12	0.25
ϕ	3.03e-05	3.05e-05	3.112e-05	3.24e-05	3.499e-05
σ^2	1.11	1.23	1.30	1.38	1.56
τ^2	7.60	7.75	7.85	7.94	8.10

Table 5.7: Model comparison using Deviance Information Criterion for biomass and forest ownership.

Model	pD	DIC
Non-spatial	6	28052
Spatial (taper range equals 38 km)	1043	27419
Spatial (taper range equals 75 km)	802	27409

5.2.3 Stand age and ownership

This analysis focuses on explaining the variability in the outcome variable stand age (AGE) using the five ownership categories defined in Table 5.1. The distribution of the outcome variable was approximately normal and no transformation was considered. Table 5.8 summarizes parameters' posterior distribution estimated using the non-spatial model. These parameter estimates suggest that stands on all ownerships except for OWN-45 are on average older than the baseline ownership OWN-41.

Table 5.8: Non-spatial model's parameter credible intervals for stand age and forest ownership. Parameters that are statistically significant at the 0.05 level are indicated with an "*".

β_0^*	53.70	54.54	54.97	55.44	56.33
β_{OWN-42}^*	1.24	5.75	8.14	10.55	15.24
β_{OWN-43}^*	2.90	5.62	6.94	8.40	10.75
β_{OWN-44}^*	11.58	13.78	14.99	16.22	18.61
β_{OWN-45}	-0.47	0.56	1.07	1.56	2.45
τ^2	616.11	627.72	634.58	640.95	653.95

The semivariogram of the non-spatial model's residuals is given in Figure 5.13. Here, similar to the other analyses, the semivariogram shows clear spatial structure among the residuals which indicates a violation of the model assumptions. The vertical line on the semivariogram indicates a long estimated effective spatial range of ∼96 km.

Again, given the results of the semivariogram analysis, I added spatial random effects to the model. For this analysis, I specified a taper range equal to the estimated effective spatial range. This resulted in a tapered covariance matrix with ∼5% non-zero values. Despite this level of sparseness, the time required to fit the model was substantially less than it would take using dense matrix routines. Parameter estimates for the spatial model are summarized in Table 5.9. Unlike the previous analyses, the addition of the random effects actually made

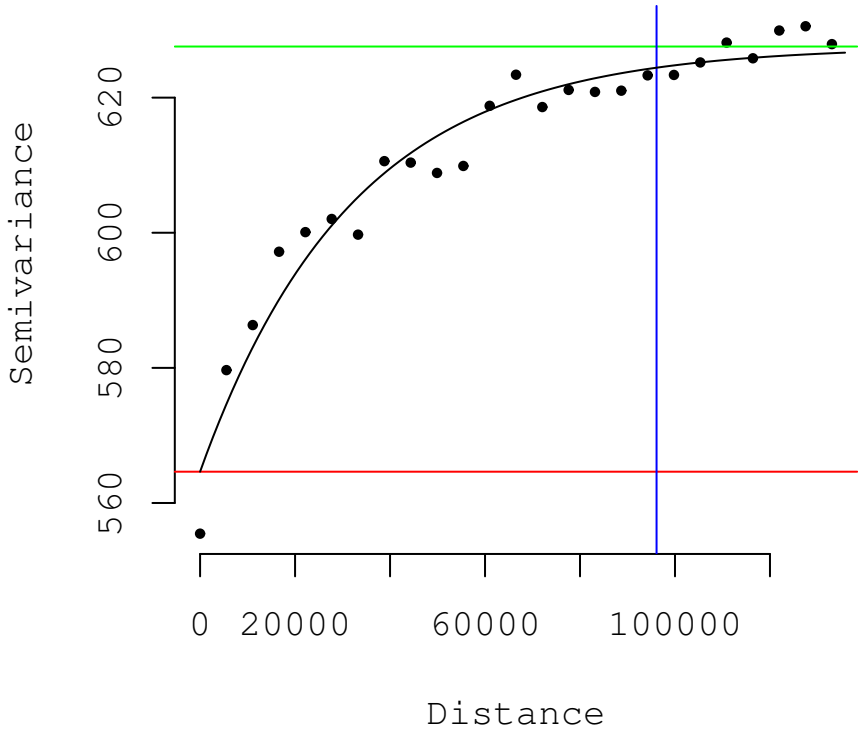

Figure 5.13: Empirical semivariogram of residuals from the non-spatial version of model (2.2) for stand age and forest ownership along with the maximum likelihood estimates of the exponential correlation function's nugget (lower horizontal line), sill (upper horizontal line), and effective spatial range (vertical line).

one of the slope parameters (OWN-45) statistically different from zero. Again, changes in parameter estimates between the non-spatial and spatial models is expected, due to the non-spatial model's violation of assumptions. The posterior distributions of non-spatial and spatial model parameters are summarized in Figure 5.14, 5.15, 5.16, 5.17 and 5.18. Here too, the addition of the spatial random effects improves model fit as reflected by the lower DIC values in Table 5.10.

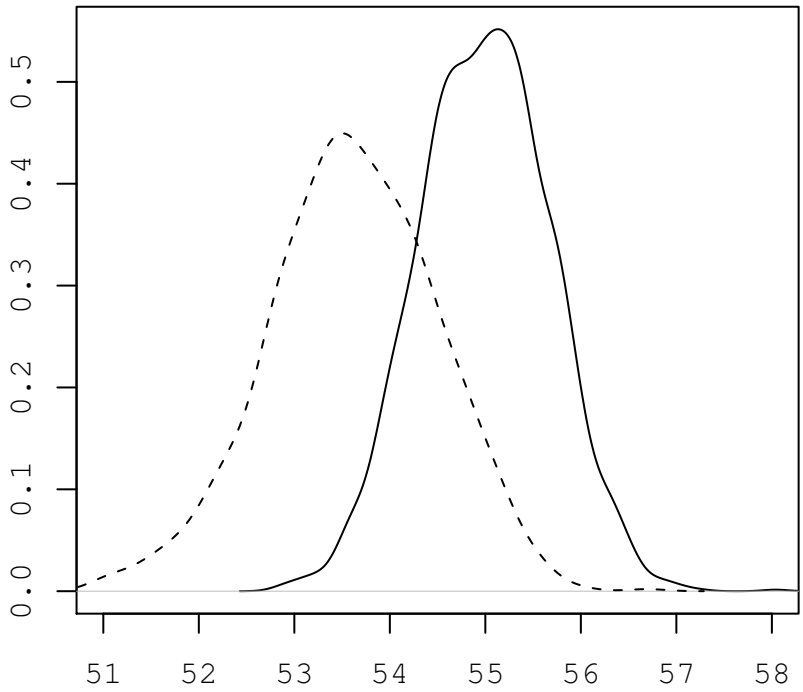

Figure 5.14: Illustration of estimated parameter's posterior distribution for intercept in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line spatial model with the taper range equal to the effective range

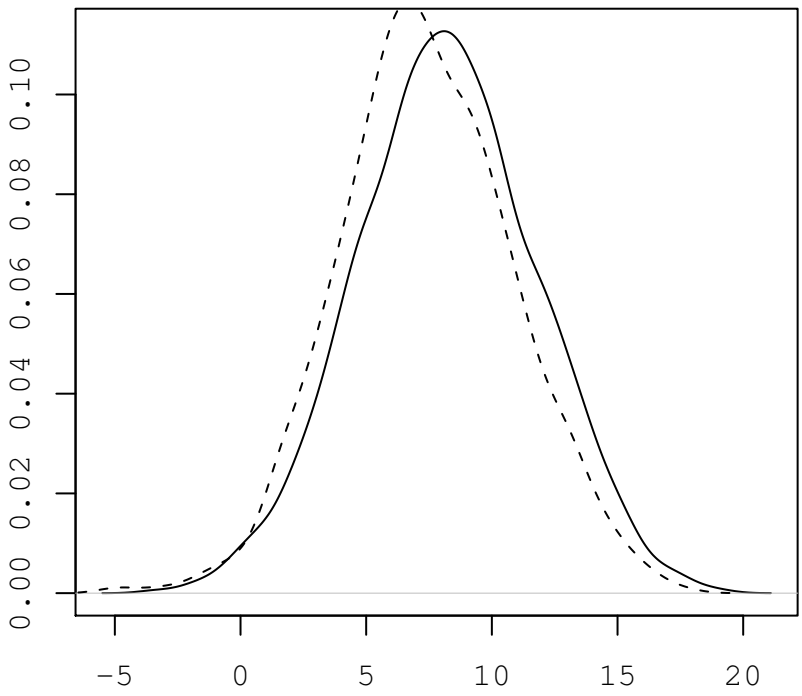

Figure 5.15: Illustration of estimated parameter's posterior distribution for OWN-42 in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line spatial model with the taper range equal to the effective range

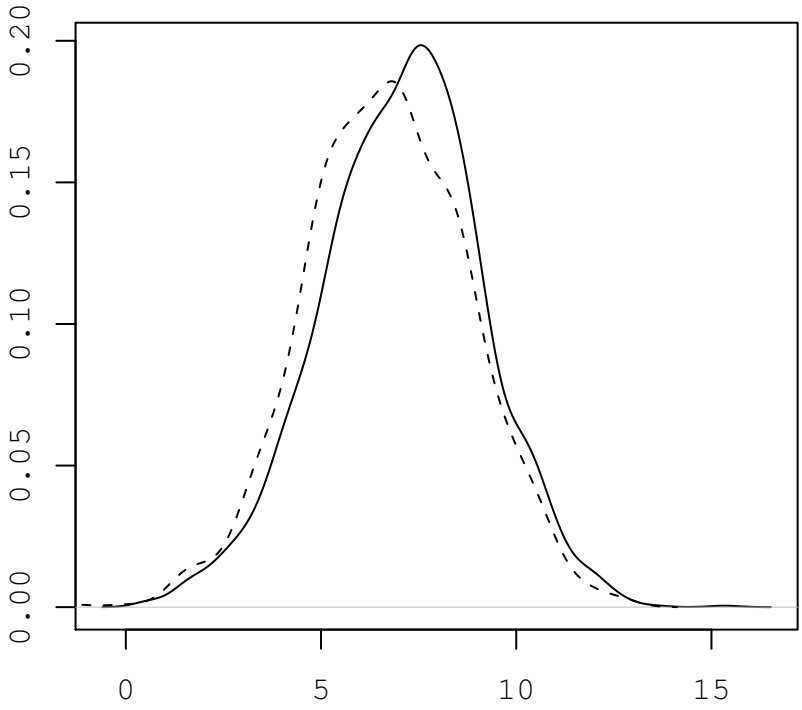

Figure 5.16: Illustration of estimated parameter's posterior distribution for OWN-43 in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line spatial model with the taper range equal to the effective range

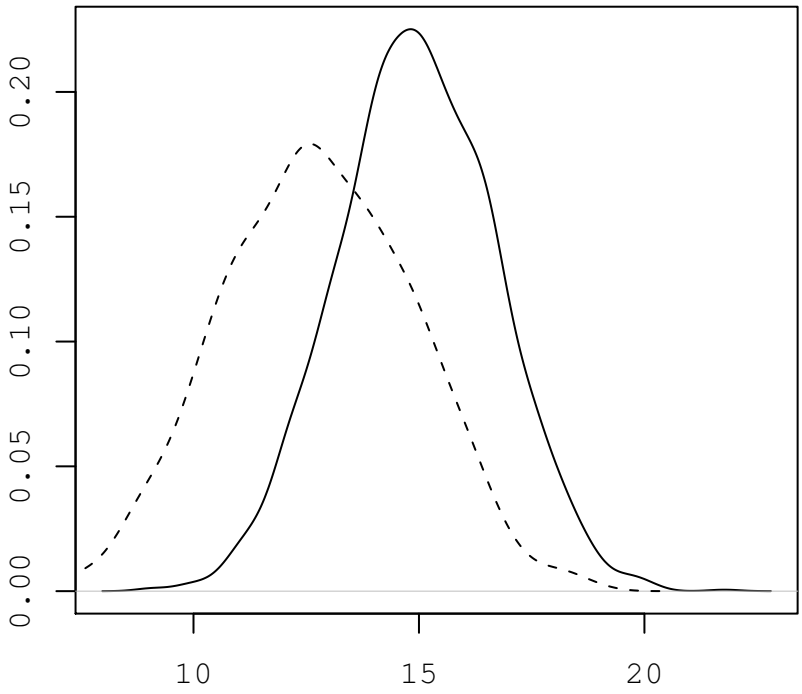

Figure 5.17: Illustration of estimated parameter's posterior distribution for OWN-44 in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line spatial model with the taper range equal to the effective range

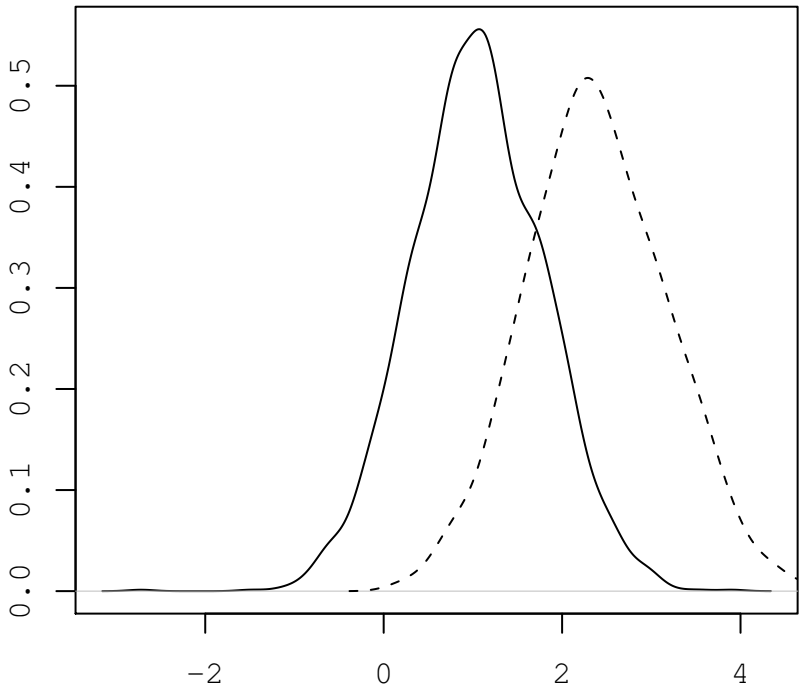

Figure 5.18: Illustration of estimated parameter's posterior distribution for OWN-45 in non-spatial model and spatial model for for biomass and forest ownership: solid line represents non-spatial model; dashed line spatial model with the taper range equal to the effective range

Table 5.9: Spatial model's parameter credible intervals for stand age and forest ownership. Parameters that are statistically significant at the 0.05 level are indicated with an "*".

	2.5%	25%	50%	75%	97.5%
Taper range equals 38 km (50% of effective spatial range)					
β_0^*	51.90	53.07	53.61	54.21	55.28
β_{OWN-42}^*	0.84	5.17	7.35	9.82	14.10
β_{OWN-43}^*	2.46	5.24	6.65	8.02	10.53
β_{OWN-44}^*	8.60	11.17	12.70	14.13	16.83
β_{OWN-45}^*	0.85	1.85	2.35	2.92	3.93
ϕ	3.05e-05	3.12e-05	3.2e-05	3.5e-05	4.3e-05
σ^2	55.07	64.20	68.44	73.22	86.33
τ^2	540.29	552.37	558.76	565.66	577.33

Table 5.10: Model comparison using Deviance Information Criterion for stand age and forest ownership.

Model	pD	DIC
Non-spatial	6	64617
Spatial (taper range equals 96 km)(effective range)	704	64307

Chapter 6

Summary and Discussion

6.1 Summary

The goal of this thesis was to explore the properties of some promising new modeling tools that will enable researchers to extract more information from large spatial datasets. Despite their power and flexibility, practitioners have been slow to adopt these methods, due in part to lack of statistical/mathematical training, computationally efficient software, and illustration of these methods within their area of interest. To help move motivate this community to engage new and more applicable modeling tools, this thesis work also yielded software and associated documentation (`http://blue.for.msu.edu/data/huirong-ms-code.pdf`) for the highlighted geostatistical models.

A central objective of this thesis was to illustrate how the proposed methods could be applied to generate valid inference about patterns in forest ownership, forest complexity, and biomass that will ultimately help in meeting the goals of the larger project. Results from the analysis of synthetic and forestry datasets, Chapter 4 and Chapter 5 respectively,

clearly demonstrate how model misspecification, specifically ignoring spatial dependence among model residuals, can result in incorrect inference about regression parameters of interest. These results have important implications for hypothesis testing and ultimately forest management and policy decisions. Specifically, the forest ownership analyses presented three cases where the wrong conclusion would be made if practitioners used a simple non-spatial analysis of variance. Applying a hierarchical Bayesian spatial model, that better meets model assumptions, provided different conclusions that more fully acknowledged the information content of the data and sources of uncertainty.

The fundamentals for fitting the Bayesian spatial regression model were offered in the Chapter 2. Here too, I detailed the computational challenges to fitting this model when the dataset is large. There are several possible approaches to reducing the dimensionality of this problem without sacrificing model richness or flexibility. Covariance tapering is a particularly appealing solution to this challenge. Chapter 3 reviewed covariance tapering and its ingredients including compactly supported correlation functions, sparse matrix structures, and software routines for sparse matrix linear algebra computations. With the reduced computational demand comes a potential loss of information and hence compromised inference – "there is no free lunch." Some care is need in selecting a taper range that balances computational gains with information loss. The synthetic data analyses in Chapter 4 provided several insights which should be useful to practitioners who want to apply covariance taper.

The models explored in this thesis assumed *stationary* and *isotropic* spatial processes. Obviously, as researchers consider larger domains, and hence are confronted with the modeling challenges partially addressed here, these assumptions are likely to be violated. There has been considerable advances for specifying nonstationary and anisotropic covariance struc-

tures (see e.g., Higdon et al. 1999, Fuentes and Smith 2001, Wood et al. 2002, Kim et al. 2005, Christopher et al. 2006). Covariance tapering can be incorporated into some of these specialized frameworks; however, model specification does become more challenging.

There are of course other options for reducing the complexity or dimensionality of the models explored in this thesis – beyond covariance tapering. Several of these contemporary approaches can be applied alone or in combination with covariance tapering. For instances, my future work will consider combining covariance tapering with a reduced rank or *predictive process* representation of the underlying spatial process, see, e.g., Banerjee et al. (2008) and Finley et al. (2009). The predictive process projects the n realization of the spatial process to a lower subspace that is defined on a smaller number of *knots*. This approach will gain greater computational efficiency by leveraging a sparse matrix specification of the random spatial effects over the smaller dimension knot-based predictive process.

Although there are still many open questions concerning the application of hierarchical spatial model to large datasets, this thesis should be helpful for guiding some decisions about when and how to apply covariance tapering to reduce computational challenges. Further, the software tools, implementation guidelines, and illustrative analyses will be instrumental to completing a full analysis of the FIA database to meet the objectives of the overarching study.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC Press.

[2] Banerjee, S., Gelfand, A.E. and Sirmans, C.F (2003) Directional Rates of Change Under Spatial Process Models. *Journal of the American Statistical Association*, **98**, 946–954.

[3] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

[4] Carlin, B.P. and Louis, T.A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall/CRC, New York.

[5] Chilés J.P. and Delfiner, P. (1999) *Geostatistics: Modelling Spatial Uncertainty*. Wiley, New York.

[6] Crainiceaniu, C.M., Diggle, P.J., and Rowlingson, B. (2008) Bivariate binomial spatial modeling of Loa Loa prevalence in Tropical Africa, *Journal of the American Statistical Association*, *103*, 21–37.

[7] Cressie, N.A.C. (1993) *Statistics for Spatial Data*, 2nd edition. New York: Wiley.

[8] Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M. Ver, and C.K. Wikle (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications* **19**, 553–570.

[9] Cressie, N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society*, Series B. **70**, 209–226.

[10] Cressie, N.A.C. and Johannesson, G. (2007) Spatial prediction for massive datasets. *Journal of the Royal Statistical Society Series B*, **70**, 209–226.

[11] Noel Cressie, Catherine A. Calder, James S. Clark, Jay M. Hoef, Christopher K. Wikle. (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*,**19**(3), 553–570.

[12] Du, J., Zhang, H., and Mandrekarm, V.S. (2009) Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *The Annals of Statistics*, **37**, 3330–3361.

[13] Finley, A.O., Sang, H., Banerjee, S. and Gelfand, A.E. (2009) Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis*, **53**, 2873–2884.

[14] Fuentes, M. (2007) Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, **102**, 321–331.

[15] Furrer, R., Genton, M.G., and Nychka, D. (2006) Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, **15**, 502–523.

[16] Furrer, R and Sain, S. (2009) Spatial model fitting for large datasets with applications to climate and microarray problems. *statistics and computing*, **19**, 113–128.

[17] Furrer, R. Spam:sparse matrix algebra. (2007) `http://www.mines.edu/~rfurrer/software/spam`

[18] Gaspari, G. and Cohn, S.E. (1999) Construction of Correlation Functions in Two and Three Dimensions, *The Quarterly Journal of the Royal Meteorological Society*, **125**, 723–757.

[19] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*, 2nd edition. Boca Raton, FL: Chapman and Hall/CRC Press.

[20] Geman, S. and Geman, D. (1984) "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(6), 721–741.

[21] Galfend, A. E. and Smith, A. F. M. (1990) "Sampling-based approaches to calculating marginal densities". *Journal of the American Statistical Association*, **85**(401), 398–409.

[22] Gneiting, T. (2002) Compactly supported correlation functions. *J. Multivariate Anal.* **83**, 493508.

[23] George, A. and Liu, J. (1981) Computer Solution of Large Sparse Positive Definite Systems, Prentice Hall.

[24] Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359-378.

[25] Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

[26] Higdon, D. (2002) Space and Space-Time Modeling Using Process Convolutions. In *Quantitative methods for current environmental issues*, eds. C. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi, 37–56. Springer-Verlag.

[27] Hoeting, J.A. (2009) The importance of accounting for spatial and temporal correlation in analyses of ecological data, .*Ecological Applications* **19**(3) , 574–577

[28] Kaufman, C., Schervish, M. and Nychka, D. (2008) Covariance tapering for likelihood-based estimation in large spatial datasets.*J. Amer. Statist. Assoc*, **103**, 1545–1555.

[29] Kamman, E.E. and Wand, M.P. (2003) Geoadditive models. *Applied Statistics* **52** 1–18.

[30] Lee, Thomas C. M. (2004) Improved Smoothing Spline Regression by Combining Estimates of Different Smoothness, *Statistics & Probability Letters* **67**, 133–140.

[31] Lee, Thomas C. M. and Oh, Hee-Seok (2004) Automatic Polynomial Wavelet Regression, *Statistics and Computing*, **14**, 337–341.

[32] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953) Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

[33] Saad, Y. (1994) Sparskit: A basic tool kit for sparse matrix computations; Version 2, `http://www.cs.umn.edu/Research/arpa/SPARSKIT/sparskit.html`

[34] Stein, M. L. (1988) Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Statist*, **16**, 55–63.

[35] Stein, M. L. (1990a) Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *Ann. Statist*, **18**, 850–872.

[36] Stein, M. L. (1990b) Bounds on the efficiency of linear predictions using an incorrect covariance function. *Ann. Statist*, **18**, 1116–1138.

[37] Stein, M. L. (1990c) A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Statist*, **18**, 1139–1157.

[38] Stein, M. L. (1997) Efficiency of linear predictors for periodic processes using an incorrect covariance function. *Journal of Statistical Planning Inference*, **58**, 321–331.

[39] Stein, M. L. (1999a) Predicting random fields with increasing dense observations. *Ann. Statist*, **9** 242273.

[40] Stein, M. L. (1999b) *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York.

[41] Stein, M.L., Chi, Z., and Welty, L.J. (2004) Approximating likelihoods for large spatial datasets. *Journal of the Royal Statistical Society, Series B* **66** 275–296.

[42] Stein, M.L. (2007) Spatial variation of total column ozone on a global scale. *Annals of Applied Statistics* **1** 191–210.

[43] Stein, M.L. (2008) A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society* **37** 3–10.

[44] Vecchia, A.V. (1988) Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B*, **50**, 297–312.

[45] Wu, Z. M. (1995) Compactly supported positive definite radial functions. Adv. *Comput. Math.* **4** 283–292.

[46] Wendland, H. (1995) Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4** 389–396.

[47] Wendland, H. (1998) Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *J. Approx.* Theory 93 258272.

[48] Zhang, H and Du, J. (2008) Covariance tapering in spatial statistics. In J.Mateu and E.Porcu, editors, *Positive definite function: From schoenberg to space-time challenge*,pages 181–196.