INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

DISSERTATION

BAYESIAN MODEL AVERAGING AND SPATIAL PREDICTION

Submitted by
Sandra E. Thompson
Department of Statistics

In partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Colorado State University
Fort Collins, Colorado
Spring 2001

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

COLORADO STATE UNIVERSITY

December 15, 2000

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY SANDRA E. THOMPSON ENTITLED BAYESIAN MODEL AVERAGING AND SPATIAL PREDICTION BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Advisor

Co-Adviser

Department Head

ABSTRACT OF DISSERTATION

BAYESIAN MODEL AVERAGING AND SPATIAL PREDICTION

I consider the problem of modeling spatial data, addressing the uncertainty involved in both the selection of explanatory variables and in the estimation of parameters. The standard practice of selecting a set of explanatory variables, estimating parameters and predicting the random process at unobserved locations ignores the uncertainty in the entire procedure. This thesis focuses on Bayesian methods to provide a more 'honest' assessment of prediction uncertainty. I also look into the effect the spatial sampling pattern has on estimation and prediction. I adopt the Matern class of autocorrelation functions throughout which allows for flexible modeling of a variety of autocorrelation curves.

Standard practice for model selection for spatially correlated data consists first of choosing a set of explanatory variables assuming independent errors and second modeling the residuals using an autocorrelation function. In this thesis I propose a technique which fits the mean and autocorrelation functions simultaneously. Models are compared using Akaike's Information Corrected Criteria (AICC) (Hurvich and Tsai, 1989). This method removes the ambiguity in determining which set of explanatory variables best fits the data when the process is correlated. Simulation results show an improvement in mean square prediction error and predictive coverage as compared to traditional methods for model selection for spatially correlated data.

This thesis includes three methods for incorporating uncertainty into prediction. First, I detail a Bayesian kriging setup, with proper prior distributions on both

the regression parameters and the spatial parameters. Second, I develop methodology for Bayesian model averaging for spatial data, incorporating all possible subsets of explanatory variables into the predictive distribution. Additionally, I develop a Bayesian model selection criteria for selecting one model while incorporating uncertainty in the estimation procedure. Simulation results for these three methods show improvements in terms of predictive coverage, mean square prediction error, and model selection with the true explanatory variables as compared to traditional methods for spatial prediction.

The configuration of sample locations often plays a key role in modeling the data. There are many types of location patterns that may arise in spatial analysis including grid sampling, sampling over geographic areas such as counties or random sampling, or a mixture of grid and clustering. The simulations in this thesis show that sampling patterns can alter the ability to select explanatory variables, estimate parameters, and predict the response at unobserved locations.

Finally, this thesis compares two methods for spatial parameter estimation, restricted maximum likelihood estimation (REML) and maximum likelihood estimation (ML). REML estimation better approximates the true autocorrelation function and gives more accurate parameter estimates, but results vary depending upon the configuration of sampling locations.

Sandra E. Thompson
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Spring 2001

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisors, Dr. Jennifer A. Hoeting and Dr. Richard A. Davis for sharing their knowledge and their time. Jennifer especially has shown incredible patience, encouragement and mentoring as I pursue my career path, and has provided endless hours of assistance. I would also like to thank the other members of my committee for their interest in my research and my successes. Dr. Geof H. Givens, my masters advisor, has continued to answer my questions.

Also, I would like to thank the faculty, students and staff of the statistics department for their friendship, support and collegiality. Additionally, Dr. Jay M. ver Hoef was generous in providing data used in this thesis. And my thanks go to Dr. Dale Heermann and the Precision Farming research group which provided me with interesting work during my studies.

I would also like to thank my friends and family, without their support I wouldn't be here today. Thanks to Gregory Shusta for regularly making sure that I eat a well balanced mean. To Tina Varga for me maintain my sanity and providing a window to a world outside of my dissertation. Thanks to Kate Catlett and Annie Epperson, they provided friendship as well as priceless editing. Last, but most importantly, my parents Kitty and Dick Thompson and brother Rich. They have been patient and provided unquestionable support.

1 Introduction **1**
1.1 Spatial Data and Models 1
1.2 Prediction with Spatial Data 4
1.3 Issues in Modeling Spatial Data 6
1.3.1 Choice of Functions for Trend and Correlation 7
1.3.2 Impact of Different Sampling Patterns 8
1.4 Accounting for Uncertainty 10

2 Issues in Modeling Spatial Data **12**
2.1 Properties of Spatial Autocorrelation Functions 12
2.1.1 Characteristics of Autocorrelation Functions 13
2.1.2 Spectral Densities 18
2.1.3 Microergodicity 19
2.2 Autocorrelation Function Families 20
2.2.1 Traditional Autocorrelation Functions 20
2.2.2 Matern Autocorrelation Function 24
2.3 Estimation of Spatial Model Parameters 26
2.3.1 Estimation of Autocorrelation 26
2.3.2 Estimation of Autocorrelation Parameters when Mean is Known 30
2.3.3 Estimation of Autocorrelation Parameters when Mean is Unknown . . 31
2.3.4 Comparison between ML and REML Techniques 34
2.4 Model Selection 35
2.4.1 Background and Traditional Methods 35
2.4.2 Spatial AICC 38

3 Bayesian Kriging with Proper Priors **40**
3.1 Prior Distributions 40
3.1.1 Types of Prior Distributions 41
3.1.2 Connection between Bayesian Kriging and Universal Kriging 42
3.1.3 History of Bayesian Kriging 43
3.1.4 Selection of Prior Distributions 46
3.2 Derivation of Distributions 50
3.2.1 Posterior Prior Distributions 50
3.2.2 Posterior Predictive Distribution 52
3.2.3 MLE Approximations 54

4 Bayesian Model Averaging for Spatial Models **56**
4.1 Introduction 56
4.2 History of Combining Models 57
4.3 Implementing Bayesian Model Averaging 58
4.3.1 General Bayesian Model Averaging Framework 59
4.3.2 Approaches for Large Model Space 60
4.3.3 Priors on the Models 62
4.4 BMA for Spatial Models 63
4.4.1 The Posterior Predictive Distribution 64
4.4.2 Posterior Model Probability 65
4.4.3 Laplace and BIC Approximations 66
4.5 Bayesian Model Selection 67

5 Simulation Studies **69**
5.1 Simulation Set-Up 69
5.1.1 Sampling Patterns 69
5.1.2 Explanatory Variables and Models 73
5.1.3 Model Assessment 76
5.2 Model Selection 78
5.2.1 Details 78
5.2.2 Results 79
5.3 Simulation without Nugget 88
5.3.1 Details 88
5.3.2 Results 90
5.4 Simulations with Nugget 102
5.4.1 Details 102
5.4.2 Results 103
5.4.3 Sensitivity 117
5.5 REML versus ML Estimation 124
5.5.1 Details 124
5.5.2 Results 125
5.5.3 Conclusion 131

6 Real Data Examples **132**
6.1 Topographic Data Set 132
6.1.1 Prior Distributions 133
6.1.2 Results 134
6.2 Lizard Data Set 136
6.2.1 Prior Distributions 138
6.2.2 Results 138

7 Conclusion and Further Research **141**
7.1 Discussion 141
7.2 Future Research 141
References 145

LIST OF TABLES

4.1 Bayes Factor Decision Table. 68

5.1 Models for Simulations 75
5.2 Model Selection Results for the Random Pattern 81
5.3 Model Selection results for the Highly Clustered Pattern 83
5.4 Model Selection results for the Lightly Clustered Pattern 84
5.5 Model Selection Results for the Regular Pattern 85
5.6 Model Selection Results for the Grid Design 86
5.7 Summary Statistics for Highly Clustered Pattern without Nugget. 96
5.8 Summary Statistics for Lightly Clustered Pattern without Nugget. 96
5.9 Summary Statistics for Random Sampling Pattern without Nugget. . . . 96
5.10 Summary Statistics for the Regular Pattern without Nugget. 96
5.11 Summary Statistics for Grid Sampling Design without Nugget. 96
5.12 Average PMP (%) for the Simulation with Nugget when $\theta_3 = 0.20$. . . . 104
5.13 Average PMP (%) for the Simulation with Nugget when $\theta_3 = 0.50$. . . . 105
5.14 Number of Simulations and Large Estimates of θ_2 when θ_3 is Non Zero. . 125
5.15 Bias results for Random Pattern 129
5.16 Bias results for Highly Clustered Design 130
5.17 Bias results for Lightly Clustered Design 130
5.18 Bias results for Regular Pattern 130
5.19 Bias results for Grid Design 131

6.1 Models for Topographic Data Set. 133
6.2 Lizard example: Summary Statistics 137
6.3 Top Five Models for the Lizard Data Set by PMP (%). 139
6.4 Prediction statistics for Cross Validated Lizard Data 140

LIST OF FIGURES

1.1 Examples of the Spatial Point Patterns Considered for Analysis 9

2.1 An Example of an Isotropic and an Anisotropic Process. 15
2.2 Examples of the Exponential Autocorrelation Function and Spectral Density. 22
2.3 Examples of the Gaussian Autocorrelation Function and Spectral Density 23
2.4 Examples of the Spherical Autocorrelation Function and the Spherical Spectral Density Function. 25
2.5 Examples of the Matern Autocorrelation Function. 27
2.6 Examples of the Matern Spectral Densities. 28

5.1 Spatial Point Patterns for the Estimation Set 71
5.2 Spatial Point Pattern for the Prediction Set 74
5.3 Distance Between Locations for the Spatial Point Patterns. 80
5.4 MSPE for the random pattern when the data does not include a nugget. 91
5.5 Comparison of MSPE for Different Spatial Designs. 95
5.6 Relative MSPE for Different Spatial Designs. 98
5.7 EPC for Different Spatial Designs. 99
5.8 PC for Different Spatial Designs. 100
5.9 Comparison of Estimates of θ_1 for Matern Nugget and Matern no Nugget for all Spatial Patterns when the True Nugget is 0.20. 107
5.10 Comparison of Estimates of θ_2 for Matern Nugget and Matern no Nugget for all Spatial Patterns when the true Nugget is 0.20. 108
5.11 Comparison of Estimates of θ_1 for Matern Nugget and Matern no Nugget for all Spatial Patterns when the true Nugget is 0.50. 109
5.12 Comparison of Estimates of θ_2 for Matern Nugget and Matern no Nugget for all Spatial Patterns when the true Nugget is 0.50. 110
5.13 Comparison of Estimates of θ_3 for Matern with Nugget for all Spatial Patterns when the True Nugget is 0.20 and 0.50. 111
5.14 Comparison of Median Autocorrelation Functions for the Matern with Nugget and Matern without Nugget when $\theta_3 = 0.2$. 113
5.15 Comparison of Median Autocorrelation Functions for the Matern with Nugget and Matern without Nugget when $\theta_3 = 0.5$. 114
5.16 Comparison of MSPE for Matern Nugget and Matern no Nugget when the True Nugget is 0.20. 115
5.17 Comparison of MSPE for Matern Nugget and Matern no Nugget when the True Nugget is 0.50. 116

5.18 Comparison of EPC for Matern Nugget and Matern no Nugget when the True Nugget is 0.20. 118
5.19 Comparison of EPC for Matern Nugget and Matern no Nugget when the True Nugget is 0.50. 119
5.20 Comparison of EPC for Matern Nugget and Matern no Nugget when the True Nugget is 0.50 for the Regular and Grid Patterns. 120
5.21 Prior Distributions for θ_1. 121
5.22 Prior Distributions for θ_2. 122
5.23 Prior Distributions for θ_3. 123
5.24 Average Autocorrelation Curves for Five Sampling Patterns. 127
5.25 Average Autocorrelation Curve for the Grid Pattern at Small Distances. 128

6.1 Locations for the Topographic Data Set. 133
6.2 Autocorrelation Curves for the Topographic Data Set. 135
6.3 Locations for the Lizard Data Set. 137

Chapter 1

INTRODUCTION

1.1 Spatial Data and Models

Prediction and the assessment of prediction errors are often the primary goal in many areas of statistics. In a spatial statistics context, one wants to predict the process at unobserved locations. The prediction paradigm typically involves the following four steps. First, a class of candidate models must be selected. For example in a general linear model framework, this would include the choice of explanatory variables and a covariance matrix for the errors. The second stage involves estimation of the parameters in the model. Third, the estimated model is used to predict the random field at unobserved locations. Finally, the model should provide an assessment of the prediction error at the unobserved locations. Generally, the prediction intervals are based on the assumption that the model is correct and the parameter estimates are known.

The first step in prediction requires the selection of one set of explanatory variables. In general, after the set of explanatory variables has been chosen, the analysis proceeds without considering other sets of explanatory variables that may fit the data equally well. In some cases, traditional methods of model selection for spatially correlated data do not select the best set of explanatory variables for prediction due to the autocorrelation structure of the response. In this thesis I propose a technique which fits the mean and autocorrelation functions simultaneously. Models are compared using Akaike's Information Corrected Criteria (AICC) (Hurvich and Tsai,

1989). This method removes the ambiguity in determining which set of explanatory variables best fits the data when the process is correlated.

This thesis will focus on Bayesian methods to provide a more 'honest' assessment of prediction uncertainty. These procedures will be compared to traditional methods of spatial prediction. This thesis includes three methods for incorporating uncertainty into prediction. First, I detail a Bayesian kriging setup expanding on work by Handcock and Stein (1993) and Kitanidis (1986), with proper prior distributions on both the regression parameters and the spatial parameters. Second, I develop methodology for Bayesian model averaging for spatial data, incorporating all possible subsets of explanatory variables into the predictive distribution. Additionally, I develop a Bayesian model selection criteria for selecting one model while incorporating uncertainty in the estimation procedure.

Spatial data analysis is a growing field of statistics that deals with modeling and inference procedures for data that are collected in space. The random field of interest, Z, is defined for some fixed area of interest, $D \subset \mathbb{R}^d$, where $d \geq 1$ is the dimension of the space. Typically, $d = 2$ or $d = 3$. In the case $d = 1$, then the spatial analysis is analogous to that used in time series.

Some fields of study which have benefited from the use of spatial statistics include epidemiology, geology, and hydrology with examples such as mapping pollutants in ground water (Cressie, 1993, p.211), rates of SIDS deaths in North Carolina (Cressie and Reed, 1989), and estimating topography (Davis, 1986). Griffith and Layne (1999) provide a wide variety of applications and analyses of spatial data. The standard reference for spatial data analysis is Cressie (1993).

The configuration of the spatial locations of the data often plays a key role in modeling the data. There are many types of location patterns that may arise in spatial analysis including grid sampling, sampling over geographic areas such as counties or random sampling, or a mixture of grid and clustering. Although

this thesis does not investigate optimal sampling locations for spatial analysis, it will focus on the effect of sampling patterns on the four steps in the prediction paradigm.

There are two ways to incorporate the spatial locations into the model. First, the locations may be included in the deterministic mean function also known as the large scale variation. For example, in the Northern Hemisphere temperature decreases while traveling to the north so the latitude is a good predictor of temperature. Second, there may be a correlation structure of the random process dependent on the distance between the locations. In this situation, the locations, or the distance between locations, may be used to model the autocorrelation structure of the process, the small scale variation.

For the moment, consider the case when there are n observations at locations $s_1, \ldots, s_n$. Denote these observations by $\boldsymbol{Z} = (Z(s_1), \ldots, Z(s_n))'$ where $\boldsymbol{Z}$ is a partial realization of a random field $\boldsymbol{Z}(s), s \in D$ and D is a fixed finite area under study. If there are explanatory variables that are also measured at these n sites, a model for the random field at any location s is given by

$$Z(s) = \sum_{j=0}^{p} X_j(s)\beta_j + \delta(s) = \boldsymbol{X}'(s)\boldsymbol{\beta} + \delta(s), \tag{1.1}$$

where p is the number of explanatory variables. $\boldsymbol{X}(s)$ is a vector of observed explanatory variables at location s. $\boldsymbol{\beta}$ is a vector of unknown coefficients, and $\delta(s)$ is the unobserved error at location s. In most cases, $X_{i0} = 1$ for all i so that the first element of $\boldsymbol{\beta}$ is the intercept. In modeling the correlated error structure, certain simplifying assumptions are made. The error is assumed to be a stationary, isotropic Gaussian process with mean zero and covariance function $\mathrm{Cov}(Z(s), Z(t)) = \sigma^2\rho(|s-t|)$, where σ^2 is the variance of the process, and $\rho(\cdot)$ is the autocorrelation function. Consequences of these assumptions and further characteristics of the autocorrelation function will be discussed in Section 2.2.

In this dissertation, I will focus on the spatial linear model in (1.1), describing a variety of methods for model selection and estimation as well as presenting methods for modeling the correlated error structure for $\boldsymbol{\delta}$. The goal in the prediction paradigm is accurate prediction and assessment of the variability of the prediction. One technique for improving the estimated prediction variability is to account for the uncertainty in the first three steps of the paradigm. Bayesian model averaging (BMA) will be used to account for the uncertainty in selecting explanatory variables (Chapter 4). I will place prior distributions on the parameters of the model to account for variability in parameter estimates (Chapter 3). The next section describes the spatial prediction model.

1.2 Prediction with Spatial Data

Spatial prediction consists of predicting the random quantity Z at an unobserved location s_0, using the data $\boldsymbol{Z}$. Universal kriging is a method of spatial prediction which yields the best linear unbiased predictor (BLUP) (Huijbregts and Matheron, 1971). Define the linear predictor for an unobserved location s_0 in terms of the components of $\boldsymbol{Z}$ by

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i). \tag{1.2}$$

The best linear unbiased predictor corresponds to the $\hat{Z}(s_0)$ that minimizes the mean square prediction error (MSPE). $E\left[\left(Z(s_0) - \hat{Z}(s_0)\right)^2\right]$ subject to the unbiasedness condition

$$E\left[\hat{Z}(s)\right] \;=\; E\left[\boldsymbol{\lambda}'\boldsymbol{Z}\right] = \boldsymbol{\lambda}'\boldsymbol{X}\boldsymbol{\beta} \;= E\left[Z(s)\right] = \boldsymbol{X}_0'\boldsymbol{\beta}$$

for all $s \in D$ and for all $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ where $\boldsymbol{X}$ is the n by $(p+1)$ design matrix. This implies $\boldsymbol{\lambda}'\boldsymbol{X} = \boldsymbol{X}_0'$ where $\boldsymbol{X}_0$ is the vector of explanatory variables at location s_0, and $\boldsymbol{\lambda}$ is a vector of weights (Cressie, 1993, Section 3.4).

The following results are from Cressie (1993, Section 3.4). Combining the spatial linear model (1.1) and minimizing the MSPE subject to the unbiasedness constraint results in

$$\boldsymbol{\lambda} = \left(\boldsymbol{\psi} + \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}\right)\right)'\boldsymbol{\Psi}^{-1}. \tag{1.3}$$

where $[\boldsymbol{\Psi}]_{ij} \equiv \rho(|s_i - s_j|)$ for $i, j \in (1, \ldots, n)$ is the assumed known matrix of autocorrelations between the n observations. $\boldsymbol{\psi}' \equiv (\rho(|s_0 - s_1|), \ldots, \rho(|s_0 - s_n|))$ is the vector of autocorrelations between the n observed values and the prediction location. The resulting universal kriging predictor is

$$\hat{Z}(s_0) = \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} + \left(\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}\right)'\hat{\boldsymbol{\beta}}\,. \tag{1.4}$$

where $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimate of $\boldsymbol{\beta}$. The prediction variance for universal kriging is

$$\sigma^2\left(1 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi} + \left(\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}\right)'\left(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}\right)\right). \tag{1.5}$$

where σ^2 and $\boldsymbol{\Psi}$ are assumed known.

Ordinary kriging is a simplification of the spatial linear model (1.1). The ordinary kriging model does not include explanatory variables other than the constant vector and is given by

$$Z(s) = \mu + \delta(s).$$

The expected value of the predictor is

$$E\left[\hat{Z}(s_0)\right] = E\left[\sum_{i=1}^{n}\lambda_i Z(s_i)\right] = \sum_{i=1}^{n}\lambda_i E\left[Z(s_i)\right] = \mu,$$

subject to the corresponding unbiasedness constraint $\sum_{i=1}^{n}\lambda_i = 1$. When the error process $\boldsymbol{\delta}$ is assumed Gaussian, the kriging predictor is equal to the optimal unbiased (non-linear) predictor under squared error loss (Cressie, 1993, p.110). The universal kriging equations produce the optimal prediction assuming the autocorrelation

function and variance are known. Standard practice requires the estimation of the autocorrelation function which is then plugged into equation (1.5). The uncertainty in this estimate of autocorrelation is not accounted for in the prediction variance formula.

1.3 Issues in Modeling Spatial Data

There are a wide variety of issues in modeling spatial data. They include the trade-off between modeling the mean function and modeling the autocorrelation function, model selection in terms of selecting explanatory variables, selection of an appropriate family of autocorrelation functions, and methods for parameter estimation. The kriging framework for modeling spatial data requires the selection of a mean function and a family of autocorrelation functions as well as estimating the parameters of both the mean and autocorrelation functions. The dependence in modeling the mean and autocorrelation function can result in mean functions that dramatically alter the autocorrelation function. One extreme is a mean function that removes the spatial autocorrelation and results in independent residuals.

Some common autocorrelation function families used in spatial statistics are discussed in Section 2.2. A variety of methods exist for estimating parameters of the correlation functions, some of the most common will be described in Section 2.3.1. The question of model selection for spatial models can be challenging. Cressie (1993, p 25) states that "Explanatory variables...should be included in the mean structure first, and great care should be taken to find all of them." Failing to include important variables may lead to errors in modeling both small scale and large scale variability. The interaction between parameter estimation and model selection will be discussed in Section 2.4.1.

1.3.1 Choice of Functions for Trend and Correlation

For the classical linear model in which the errors are independent, there are a variety of order selection methods for the explanatory variables. These methods include techniques such as Akaike's Information Criteria (AIC) and its corrected version (AICC) (Akaike, 1973; Hurvich and Tsai, 1989), the Bayesian Information Criteria (BIC) (Schwartz, 1978), and likelihood ratio tests for nested models (Neter *et al.*, 1996, Section 14.5). Similar procedures can be implemented for spatial data.

In order to implement the kriging equations a correlation function for the process must be specified. The goal is to select a family of autocorrelation functions that are capable of modeling a wide range of autocorrelations, yet only depend on a few parameters. For a given problem, the autocorrelation function family is chosen using a variety of different methods. First, it could be chosen based on previous information about the autocorrelation structure, such as a previously completed study. Second, the autocorrelation function family could be chosen via 'fit by eye', equating the empirical function to one of the families. Third, likelihood based methods such as AICC could be used to compare the quality of fit for a variety of families.

The interaction between the trend and autocorrelation functions makes it problematic to fit both functions simultaneously. The inclusion of some explanatory variables may remove the spatial autocorrelation from the data, resulting in a model that has uncorrelated or nearly uncorrelated residuals. Similarly, some autocorrelation functions may reduce the importance of the explanatory variables so that once the correlation function family is taken into account, none of the explanatory variables are significant using standard significance tests. These two possibilities are extremes, but care should be taken in every step of the analysis. In fact, Cressie (1993, p.25) states "What is one person's (spatial) covariance structure may be another person's mean structure."

In this thesis, I propose a technique which fits the mean and autocorrelation functions simultaneously by using the likelihood based method Akaike's Information Corrected Criteria (AICC) (Hurvich and Tsai, 1989). This approach removes the ambiguity in determining which set of explanatory variables best fits the data when the process is correlated.

1.3.2 Impact of Different Sampling Patterns

The configuration of sampling locations of the observations can have a large effect on the ability to model both the large and small scale variability. For example, if the sampled locations are sparse, then it is difficult to model the small scale variation. In contrast, if the sampled locations are clustered and do not provide complete coverage of D, it may be easier to model the small scale variation at the expense of modeling the large scale variation.

One component of this thesis considers the impact of various configurations of sampled points on modeling and prediction performance. Figure 1.1 contains a variety of the spatial point patterns considered in this thesis. The patterns range from a grid pattern where the locations are equally spaced throughout the area D, to highly clustered data where the locations are grouped together. Section 5.1.1 includes further discussion of spatial point patterns.

Frequently, the sampling pattern for the locations of observed data is not under the control of the researcher. Instead, it may be determined by location accessibility and cost of gathering additional observations. For example, it may be difficult to get to many random locations throughout the area of interest. Instead, the samples may be clustered. An example of this would be biological studies in mountainous areas, where logistical considerations may force the choice of sampling scheme. On the other hand for modeling insect populations, it may be advantageous to place the traps on a regular grid design, to ease the placement of a large number of traps.

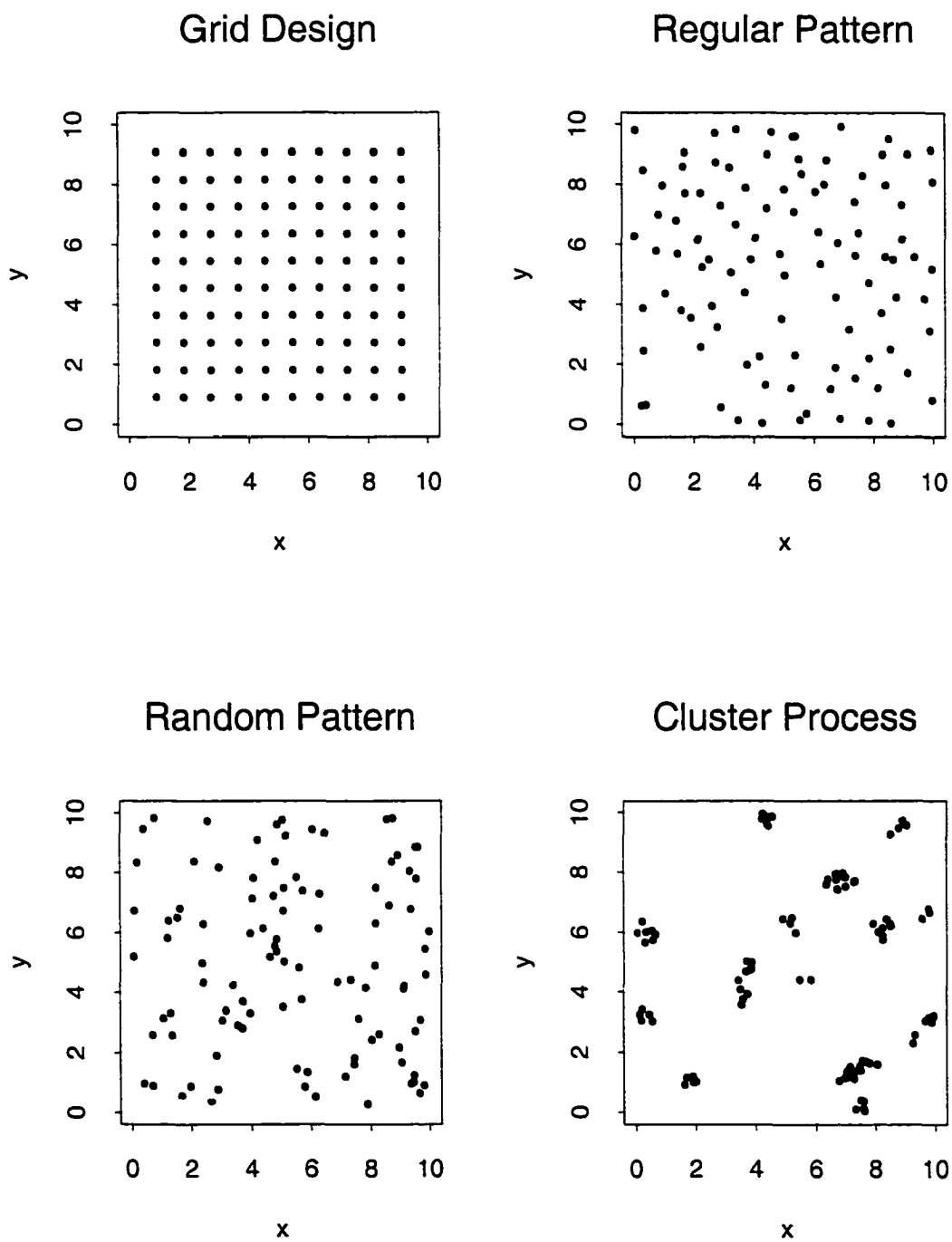

Figure 1.1: Examples of the Spatial Point Patterns Considered for Analysis. The patterns are a) Grid, b) Regular, c) Random, and d) Cluster.

With the advent and accessibility of new technologies such as a Global Positioning System (GPS), it is easier to implement a random location pattern throughout the area of interest.

One question of interest is what sampling pattern is best for prediction? And, if the sampling pattern is fixed due to forces outside of the researcher's control, what is the best procedure to model the data in order to get the most information out of a particular location configuration? Results of this thesis will show that differences in the location of sampling points can effect the quality of prediction and the assessment of prediction error.

1.4 Accounting for Uncertainty

Traditionally, spatial data are modeled using a two step procedure. First, the mean function is chosen using either a model selection procedure such as Akaike's AICC (Hurvich and Tsai, 1989) ignoring the spatial autocorrelation or from previous expert opinion about the behavior of the process of interest. Second, the residuals are modeled as a zero mean process with some correlation structure.

Once the spatial model has been chosen, the parameters are estimated and values of the random field at unobserved locations are predicted. Classical prediction intervals for the unobserved locations account for the variability in the parameter estimates of the mean function, but assume that the chosen model is the correct model for the data. Classical prediction also assumes the autocorrelation function is known. The uncertainty inherent in the model selection procedure and in the estimation of the covariance parameters as well as selecting the autocorrelation function are ignored.

Bayesian kriging models have been proposed by Handcock and Stein (1993), Omre (1989), and others. Building on model (1.1) these models use Bayesian methodology to incorporate prior distributions on the regression and covariance

parameters and to estimate the parameters using the posterior distribution. The corresponding prediction intervals include the uncertainty in the parameter estimates.

Bayesian model averaging (BMA) provides a framework where the results from many models can be combined and the uncertainty in the selection of explanatory variables can be included in the prediction intervals. BMA has been applied to general linear models (Hoeting *et al.*, 1999) and to survival analysis models (Volinsky, 1997). This dissertation incorporates the spatial model and correlation function families into the BMA methodology to improve prediction intervals so as to provide more accurate assessments of the coverage associated with these prediction intervals.

Chapter 2

ISSUES IN MODELING SPATIAL DATA

Traditionally, the autocorrelation function is selected based on prior information such as previous studies and scientific knowledge about the process, as well as estimation using the observed data. In Section 1.2, the prediction equation and prediction variance conditional on the assumed autocorrelation function ρ and variance σ^2 were described. This chapter focuses on methods for modeling the autocorrelation process. Two primary issues in modeling the spatial process are selection of an autocorrelation function family and estimation of the parameters indexing the family.

Section 2.1 describes some of the characteristics associated with autocorrelation functions. Four families of autocorrelation functions, the spherical, the exponential, the Gaussian, and the Matern are discussed in Section 2.2. Many techniques exist to estimate the parameters of autocorrelation functions, some of the most common are described in Section 2.3.1. Section 2.4 discusses the traditional methods for model selection in spatial models and proposes an improved method which combines model selection with spatial parameter estimation.

2.1 Properties of Spatial Autocorrelation Functions

If n locations of a spatial process are sampled, there are $\frac{n(n-1)}{2}$ pairs of observations. For each pair of locations, there is an autocorrelation to be estimated. A common procedure for consolidating the number of estimated parameters in the autocorrelation is to model it as a parametric function indexed by $\boldsymbol{\theta}$.

2.1.1 Characteristics of Autocorrelation Functions

In this section I consider several commonly used families of isotropic functions. The properties of general correlation functions and those of these various families will also be described. Much of this section is drawn from Cressie (1993). In this thesis, I focus on autocorrelation functions that are weakly stationary and isotropic.

Weak Stationarity

A spatial process $\delta(\boldsymbol{s})$ is defined to be second order (weakly) stationary if the following two conditions hold:

$$E[\delta(\boldsymbol{s})] = \mu, \ \forall \ \boldsymbol{s} \in D \tag{2.1}$$

$$\text{Cov}[\delta(\boldsymbol{s}), \delta(\boldsymbol{t})] = \sigma^2 \rho(\boldsymbol{s} - \boldsymbol{t}), \ \forall \ \boldsymbol{s}, \boldsymbol{t} \in D. \tag{2.2}$$

where ρ is an autocorrelation function. Condition (2.1) states that the mean of the spatial process is location invariant while condition (2.2) specifies that the autocorrelation between two values of the spatial process depends only on the distance and direction between the locations.

When modeling the autocorrelation function of a weakly stationary spatial process the following three properties follow:

$$\begin{aligned} \text{Cov}(\mathbf{0}) &> 0. \\ \text{Cov}(\boldsymbol{h}) &= \text{Cov}(-\boldsymbol{h}) \text{ and} \\ |\text{Cov}(\boldsymbol{h})| &\leq \text{Cov}(\mathbf{0}) \end{aligned} \tag{2.3}$$

where $\boldsymbol{h}$ is the vector of distances between locations and $\text{Cov}(\cdot) = \sigma^2 \rho(\cdot)$ is defined to be the covariance function. Equations (2.3) states that the variance of the spatial process at all locations must be greater than zero, the autocovariance function based on the distance is a symmetric function, and the covariance of the spatial process

between any two locations must be less than the variance of the spatial process (Stein, 1999, p.19).

A necessary and sufficient condition for the existence of a weakly stationary process with autocorrelation function ρ is that ρ must be nonnegative definite, i.e., if

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i \delta(\boldsymbol{s}_i)\right) = \sum_{i,j=1}^{n} a_i a_j \sigma^2 \rho(\boldsymbol{s}_i - \boldsymbol{s}_j) \geq 0 \tag{2.4}$$

for all $a_i, a_j \in \mathbb{R}$ and for all $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathbb{R}^k$.

Isotropy

The process is said to be isotropic if $\rho(\cdot)$ is a function of the Euclidean distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ (Cressie, 1993, p. 53). Isotropic functions have the same autocorrelation structure in the north-south direction as the east-west direction. An anisotropic process is a function of the distance and direction between locations. For example, the autocorrelation function may be twice as strong in the north-south direction compared to the east-west direction. Figure 2.1 shows the contour curves for an isotropic process as well as an anisotropic process.

Weak stationary and isotropy simplify the modeling of the autocorrelation function and hence restricts the family of autocorrelation functions to be used. There are three additional characteristics of the autocorrelation function that we will consider: the smoothness of the correlation function near the origin, the range, and the nugget effect. These characteristics are useful in distinguishing between different families of autocorrelations.

Smoothness

Stein (1999, p.12) states "that properties of spatial interpolants depend strongly on the local behavior of the random field". Smoothness describes the behavior of

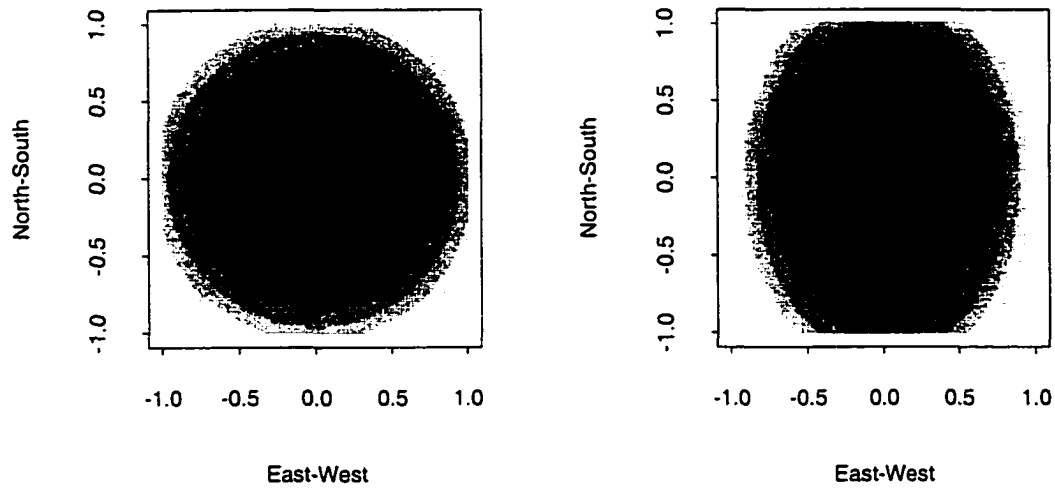

Figure 2.1: An Example of an Isotropic and an Anisotropic Process. The plot on the left represents a sample isotropic process where the autocorrelation is a function of the distance between locations. The plot on the right represents a sample anisotropic process where the autocorrelation is a function of both the distance and the direction between locations, with the north-south direction twice as strong as the east-west direction.

the autocorrelation function near the origin. The larger the smoothness value, the larger the number of derivatives at the origin. Ripley (1981, p.50-51) elaborates on smoothness for Gaussian processes as well as for general autocorrelation functions. For any autocorrelation function ρ which is continuous at zero, the smoothness can be quantified using

$$|\rho(\mathbf{0}) - \rho(\boldsymbol{h})| \leq c\,|\boldsymbol{h}|^{\alpha} \tag{2.5}$$

as $\boldsymbol{h}$ approaches zero where c is a constant. If $\alpha > 0$ then the surface is continuous and if $\alpha > 2n$, the surface is differentiable n times.

Smoothness is also connected with mean square differentiability. A process δ on $\mathbb{R}$ with finite second moments is called mean square differentiable at t if

$$\frac{1}{h_n}\{\delta(t + h_n) - \delta(t)\}$$

converges in L^2 to $\delta'(t)$ for all sequences $\{h_n\}$ converging to 0 as $n \to \infty$ (Stein, 1999, p.21). The process δ is α times mean square differentiable if it is $(\alpha - 1)$ times mean square differentiable and $\delta^{(\alpha-1)}$ is mean square differentiable.

For most of the traditional autocorrelation function families, each member has the same smoothness value. One exception is the Matern autocorrelation family which will be described in Section 2.2.2 and does include a parameter for smoothness.

Range

The range is the minimum distance between observations such that the points can be considered uncorrelated. Since most models used in practice include an autocorrelation function that is non-zero for all distances, the range is often described as the minimum distance h such that $|\mathrm{Cov}[\delta(\boldsymbol{s}), \delta(\boldsymbol{s}+\boldsymbol{h})]| < \epsilon$ where $\epsilon > 0$ is a preassigned small number. Some autocorrelation function families such as the spherical family described in Section 2.2.1 have $\mathrm{Cov}[\delta(\boldsymbol{s}), \delta(\boldsymbol{s}+\boldsymbol{h})] = 0$ for all $|\boldsymbol{h}|$ greater than the range. This definition simplifies the analysis by creating block diagonal autocorrelation matrices and reduces the computing time necessary for inverting the autocorrelation matrix. Most other autocorrelation function families such as the Gaussian, Matern and have non zero values but decay exponentially to zero as the distance goes to infinity. All autocorrelation function families described in this dissertation include a parameter for range.

Nugget

If $\lim \rho(\boldsymbol{h}) \to 1 - c_0 > 0$ as $|\boldsymbol{h}| \to 0$ for an isotropic correlation function $\rho(\cdot)$, then c_0 is the nugget term (Cressie, 1993, p.59). The nugget, also known as the discontinuity at the origin, is due to two sources of variability, measurement error and microscale variation. Measurement error is the variability in the process used to measure the observations. For example, if the observations were repeatedly sampled

at the same location, it is usually assumed that any variability present is due to measurement error.

The microscale variability refers to variability at distances smaller than the observed distances between locations. If the process is assumed to be L_2 continuous, i.e.,

$$E_{||\boldsymbol{h}||\to 0}\left[(\delta(\boldsymbol{s}+\boldsymbol{h})-\delta(\boldsymbol{s}))^2\right] \to 0. \tag{2.6}$$

then the microscale variability is zero. This is equivalent to continuity of the autocorrelation function ρ at zero. In general, the spatial process is assumed continuous at the microscale, which restricts any nonzero c_0 to be measurement error.

Nugget terms can be incorporated into any autocorrelation function family. Generally, real data examples include some amount of measurement error. Situations where the nugget term is zero as well as situations when there is a significant amount of measurement error in the data will be investigated in the simulations in Chapter 5.

Potential Difficulties

Identifiability can be a problem in modeling a spatial process. For example, when there is a non-zero nugget, it may be difficult or impossible to accurately estimate the smoothness of the process based on observed data. If the data are observed on a grid, then there are no pairs of observations at distances less than one unit apart. In this situation, holding the range constant, as the smoothness of the process at the origin increases, the nugget also increases, resulting in very similar autocorrelation curves at distances greater than one unit.

Another challenge is that the parameter space for the model may need to be restricted in order for the correlation matrix to be positive definite (Cressie, 1993, p.89.94). This is not a serious issue for the autocorrelation functions described in

this thesis. A general rule that characterizes permissible parameters which result in valid covariance models does not exist. Instead, each parameter set should be checked to be sure the resulting covariance matrix is valid.

2.1.2 Spectral Densities

One tool for examining the behavior of the autocorrelation function of a spatial process is through the corresponding spectral density. The behavior of the autocorrelation function in terms of smoothness and range corresponds to behavior of the spectral density. The smoothness of an autocorrelation function is related to the tail behavior of a spectral density and the range of an autocorrelation function is related to the height of the spectral density at zero.

The spectral density can be found from the autocorrelation function using the inversion formula (Brockwell and Davis, 1991, p.152) which states that if $\delta(\boldsymbol{h}), \boldsymbol{h} \in \mathbb{R}^d$, is a zero mean stationary process with covariance function $\mathrm{Cov}(\cdot) = \sigma^2\rho(\cdot)$, which is continuous at zero, then there exists a spectral distribution function $F(\boldsymbol{t})$ such that

$$\sigma^2\rho(\boldsymbol{s}) = \int_{\mathbb{R}^d} \exp\{i\boldsymbol{s}\boldsymbol{\lambda}\}\, dF(\boldsymbol{\lambda}). \tag{2.7}$$

The spectral density f is defined to be

$$f(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp\{-i\boldsymbol{\omega}'\boldsymbol{h}\}\, \mathrm{Cov}(\boldsymbol{h}) d\boldsymbol{h} \tag{2.8}$$

if the integral exists (Stein, 1999, p.25). If the covariance function is assumed isotropic and integrable then letting $\tau = ||\omega||$, the equations above simplifies to

$$f(\tau) = \frac{1}{2\pi} \int_0^\infty J_0(r\tau)\mathrm{Cov}(r) r dr \tag{2.9}$$

$$\mathrm{Cov}(r) = 2\pi \int_0^\infty J_0(r\tau) f(\tau) r d\tau \tag{2.10}$$

in $\mathbb{R}^2$, where J_0 is the Bessel function of the first kind (Abramowitz and Stegun, 1965, Chapter 9).

2.1.3 Microergodicity

Stein (1999, p.163) defines a function g on $\boldsymbol{\theta}$, the parameters, to be microergodic if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, then $g(\boldsymbol{\theta}) \neq g(\boldsymbol{\theta}')$ implies $P_{\boldsymbol{\theta}} \perp P_{\boldsymbol{\theta}'}$ where $P_{\boldsymbol{\theta}}$ is defined to be a class of probability models for a random field. Two measures P_0 and P_1 are orthogonal ($P_0 \perp P_1$) if there exists a set A such that $P_0(A) = 1$ and $P_1(A) = 0$.

Microergodicity is related to the ability to find both asymptotically optimal predictors as well as asymptotically correct mean square prediction error (Stein, 1999, Section 6.1).

When the error structure is assumed to be a Gaussian random field, equivalence and orthogonality of the measures are helpful in determining microergodicity. Let m_0 and m_1 be continuous functions on $\mathbb{R}^d$, K_0 and K_1 are continuous, positive definite on $\mathbb{R}^d$ and D is a closed subset of $\mathbb{R}^d$, where d is the dimension of the space. Stein (1999, Section 4.2) proves that Gaussian measures $P_0 = G_D(m_0, K_0)$ and $P_1 = G_D(m_1, K_1)$ are either equivalent or orthogonal where $G_D(m, K)$ is the Gaussian measure for a random field on D with mean m and covariance K.

Microergodicity is determined from the assumptions about the asymptotic behavior of the process. Spatial statistics uses both infill asymptotics where the observations are becoming more and more dense within a fixed area D, and increasing domain asymptotics where the maximum distance between locations tends to infinity (Cressie, 1993, Section 2.6.3). Infill asymptotics allow for more information about the small scale variation, while increasing domain asymptotics are more useful in large scale variation. If a parameter is non-microergodic under infill asymptotics, then it is not possible to estimate that parameter consistently and the asymptotic behavior of the MLE does not converge to normality (Stein, 1999, p.174).

2.2 Autocorrelation Function Families

As mentioned in Section 2.1.1, the autocorrelation function depends upon a parameter set θ, a vector of unknown parameters which describes the shape of the autocorrelation function. Traditional autocorrelation functions that will be addressed in this thesis include the exponential, Gaussian, and spherical classes. The Matern class is more flexible than the traditional autocorrelation functions in the variety of behaviors it can describe.

The general form of an autocorrelation function is

$$\underset{\sim}{\rho}(h) = \theta_3 I_{(h=0)} + (1 - \theta_3)\,\rho(h) \tag{2.11}$$

where θ_3 is the proportion of variance due to the nugget term; $I_{(h=0)}$ is an indicator function which equals one when distance equals zero and 0 otherwise; and $\rho(h)$ is a member of some autocorrelation function class. When a nugget term is not included in modeling, $\theta_3 = 0$, the general autocorrelation function reduces to $\underset{\sim}{\rho}(h) = \rho(h)$.

2.2.1 Traditional Autocorrelation Functions

The following sections describe three of the most common autocorrelation function families used for modeling spatial processes. All three include a parameter for the range, but each has a fixed smoothness term specific for the family. There are other traditional autocorrelation functions such as the linear, rational quadratic, wave, and power model which will not be addressed in this thesis. See Cressie (1993, p.61-63) and Haining (1990, p.97) for further details about autocorrelation functions not discussed in this thesis.

Exponential Class

The exponential autocorrelation family has one parameter $\boldsymbol{\theta} = \theta_1$. the range of the autocorrelation as defined in Section 2.1.1 with $\epsilon = \exp(-1)$. and has autocorrelation function

$$\rho_{exp}(h) = \exp\left(\frac{-h}{\theta_1}\right). \tag{2.12}$$

where h is the distance between locations. The corresponding spectral density is

$$f_{exp}(\omega) = \sigma^2\left(\theta_1^2 + \omega^2\right)^{-1}. \tag{2.13}$$

This family is continuous but not differentiable at the origin. Examples of the exponential autocorrelation function as well as the spectral density can be seen in Figure 2.2. In the figure. as the range increases in the autocorrelation function plot. the corresponding spectral densities are taller at the origin.

Gaussian Class

The Gaussian autocorrelation family also has one parameter $\boldsymbol{\theta} = \theta_1$. where θ_1 is the range of the autocorrelation. has autocorrelation function

$$\rho_{gau}(h) = \exp\left(\frac{-h^2}{\theta_1^2}\right). \tag{2.14}$$

and spectral density

$$f_{gau}(\omega) = \sigma^2\left(\frac{\theta_1^2}{4\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{4}\omega^2\theta_1^2\right). \tag{2.15}$$

Examples of both the autocorrelation function and the spectral density can be seen in Figure 2.3.

The Gaussian autocorrelation family is continuous and infinitely differentiable at the origin. The Gaussian name arises from the similarity in form of the autocorrelation function to the Gaussian distribution. The smooth behavior of this

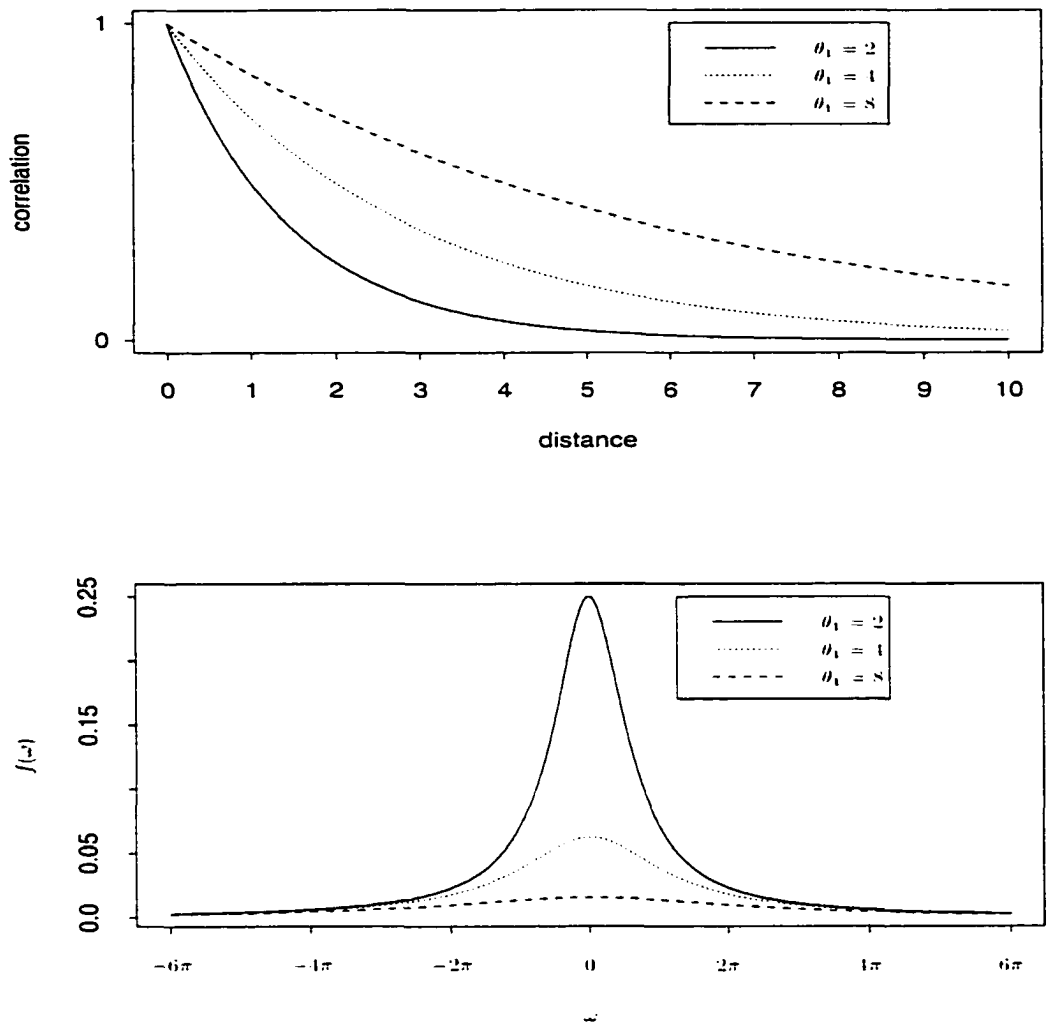

Figure 2.2: Examples of the Exponential Autocorrelation Function and Spectral Density. The first figure plots the autocorrelation function for $\theta_1 = (2, 4, 8)$ and the second figure shows examples of the corresponding spectral densities.

autocorrelation function leads to unnatural behavior in the spatial process. As Stein states, "it is possible to predict $Z(t)$ perfectly for all $t > 0$ based on observing $Z(s)$ for all $s \in (-\epsilon, 0]$ for any $\epsilon > 0$" (1999, p.30). There are few, if any, applications for which this assumption holds.

Furthermore, Stein (1999, Section 3.5) investigated the behavior of the Gaussian autocorrelation function under model (1.1) with zero mean. Data were simulated

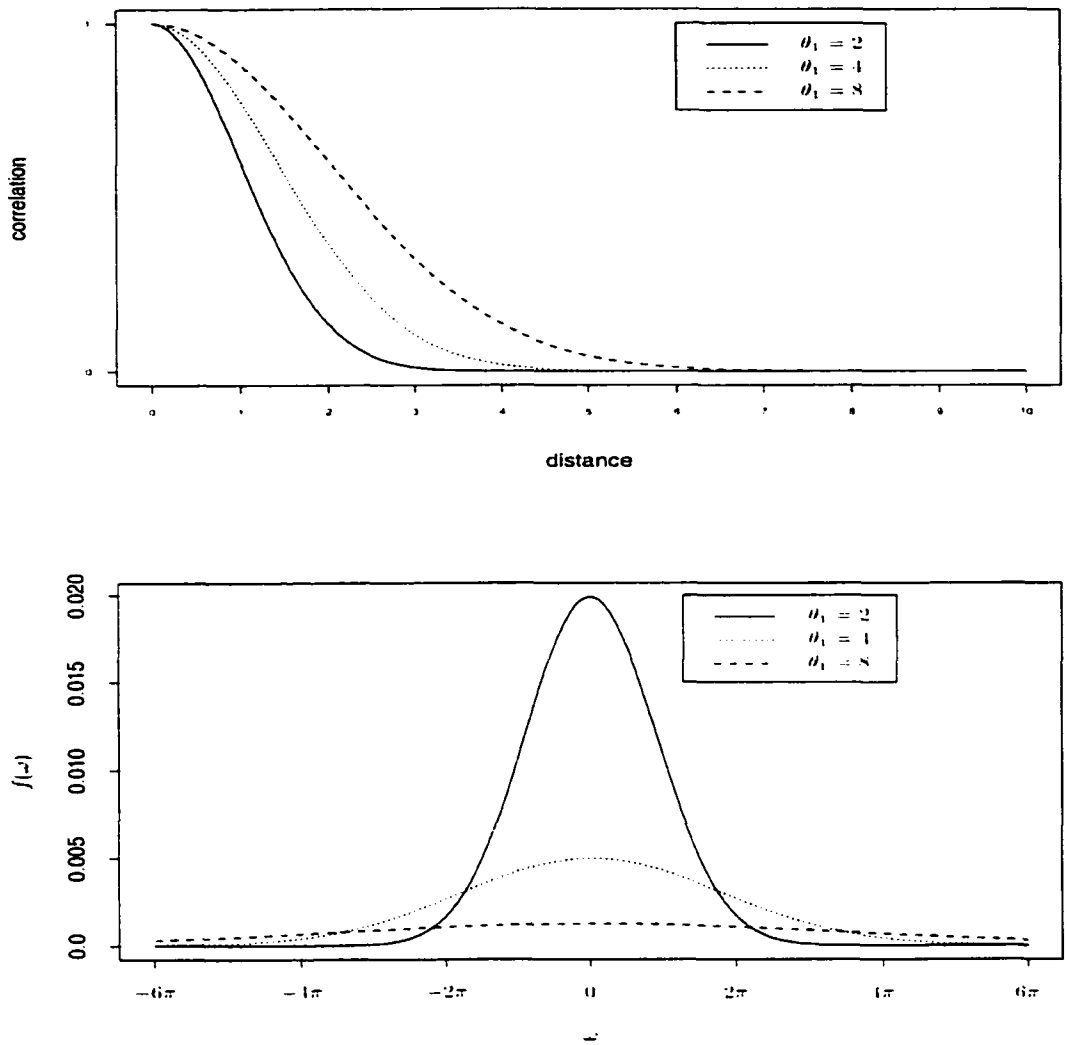

Figure 2.3: Examples of the Gaussian Autocorrelation Function and Spectral Density. The corresponding ranges are (2.4.8).

in $\mathbb{R}$ at locations $Z(\pm\epsilon j)$ for $j = 1, \ldots, n$ from either the Gaussian autocorrelation function (2.14) or from the autocorrelation function $\rho(h) = \exp\{-|t|\}(1 + |t|)$. The first two terms of the Taylor series expansion for both autocorrelation functions are the same. Results from this experiment show that the prediction estimates are reasonable, but the mean squared prediction error was underestimated. This underestimation of the prediction error can indicate over confidence in the prediction.

Stein concludes by recommending against the use of the Gaussian autocorrelation function.

Spherical Class

The spherical autocorrelation family described by Cressie (1993, p.61) is parameterized by $\theta = \theta_1$, where θ_1 is the range of the autocorrelation, and has autocorrelation function

$$\rho_{sph}(h) = \begin{cases} 1 - \frac{3h}{2\theta_1} + \frac{h^3}{2\theta_1^3}, & 0 \leq h. \leq \theta_1. \\ 0. & h > \theta_1. \end{cases} \tag{2.16}$$

This autocorrelation function is valid in $\mathbb{R}^1$, $\mathbb{R}^2$, and $\mathbb{R}^3$. Examples of the spherical autocorrelation function (2.16) are shown in Figure 2.4.

The spherical autocorrelation function is the only autocorrelation function I consider which has zero autocorrelation at distances greater than the range θ_1. This simplifies the matrix of autocorrelations and can reduce the amount of calculations due to block diagonal structure of the autocorrelation matrix. The spherical autocorrelation family is continuous, but not differentiable at the origin.

The spectral density of autocorrelation function (2.16) in $\mathbb{R}$ is

$$f_{sph}(\omega) = 1.5\frac{2 + \omega^2\theta_1^2 - 2\cos(\theta_1\omega) - 2 * \theta_1\omega\sin(\theta_1\omega)}{\omega^4\theta_1^3} \tag{2.17}$$

Figure 2.4 also displays the spectral density of the spherical autocorrelation function. Note the slight fluctuation of the spectral density when $\theta_1 = 8$. This behavior does not correspond to any physical process, and therefore the use of the spherical class is questionable (Stein, 1999, p.31).

2.2.2 Matern Autocorrelation Function

The Matern autocorrelation function is parameterized by $\theta = (\theta_1, \theta_2)$, where θ_1 is the range of autocorrelation and θ_2 is the smoothness. This autocorrelation

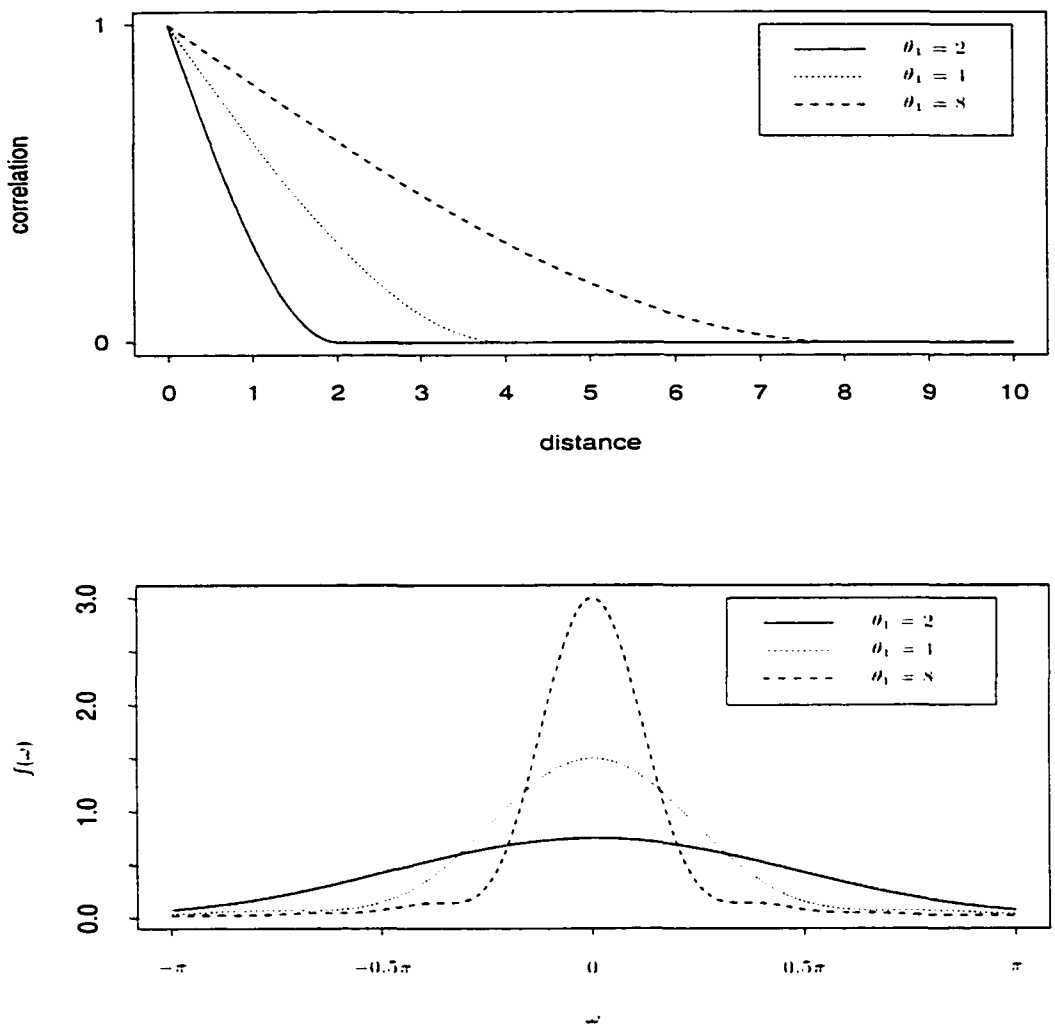

Figure 2.4: Examples of the Spherical Autocorrelation Function and the Spherical Spectral Density Function.

function is continuous for $\theta_2 \geq 1$ and differentiable n times for $\theta_2 \geq n$. The autocorrelation function is

$$\rho_{mat}(h) = \frac{1}{2^{\left(\theta_2 - \frac{d-3}{2}\right)}\Gamma\left(\theta_2 - \frac{d-1}{2}\right)} \left(\frac{h}{\theta_1'}\right)^{\left(\theta_2 - \frac{d-1}{2}\right)} \mathcal{K}_{\left(\theta_2 - \frac{d-1}{2}\right)} \left(\frac{h}{\theta_1'}\right). \qquad (2.18)$$

where $\theta_1' = \theta_1 / \left(2\sqrt{\theta_2 - \frac{d-1}{2}}\right)$ and $\mathcal{K}_{\theta_2}$ is the modified Bessel function of order θ_2 (Abramowitz and Stegun, 1965). This autocorrelation function is valid for $d \geq 1$.

in this thesis I focus on $d = 1$. The spectral density is (Stein, 1999, p.48-50)

$$f_{mat}(\omega) = \frac{\Gamma(\theta_2 + d/2)(4\theta_2)^{\theta_2}}{\pi^{d/2}\Gamma(\theta_2)\theta_1^{2\theta_2}} \left(\frac{4\theta_2}{\theta_1^2} + \omega^2\right)^{-(\theta_2+d/2)}. \quad (2.19)$$

This parameterization of the Matern autocorrelation function is selected because the interpretation of the parameters is similar to the other autocorrelation functions (Stein, 1999, p.50). The Matern covariance class includes the exponential class as a member when $\theta_2 = 0.5$ and $d = 1$, along with the Gaussian covariance function as a limiting member when $\theta_2 \to \infty$. Examples of the autocorrelation function can be seen in Figure 2.5, while examples of the spectral densities can be seen in Figure 2.6.

The Matern autocorrelation function can be traced back to Whittle (1954). Prior to Handcock and Stein (1993), the Matern autocorrelation function was not commonly used in analyzing spatial data. Currently, there are a variety of applications using the Matern autocorrelation function including Handcock and Wallis (1994), Ecker and Gelfand (1997; 1999), and Stein (1999).

2.3 Estimation of Spatial Model Parameters

There are a wide variety of methods used to estimate parameters of spatial models including ad hoc procedures, likelihood-based methods, and least square based methods. The ad hoc methods, also known as 'fit by eye', compare an empirical autocorrelation curve to a set of known autocorrelation curves. The likelihood-based methods include maximum likelihood estimation and restricted maximum likelihood estimation as well as robust estimators. The least squares family of estimators includes weighted least squares and generalized least squares methods.

2.3.1 Estimation of Autocorrelation

For the stationary spatial error process δ, the variogram is defined to be

$$2\gamma(\boldsymbol{h}) = \mathrm{Var}(\delta(\boldsymbol{s} + \boldsymbol{h}) - \delta(\boldsymbol{s})).$$

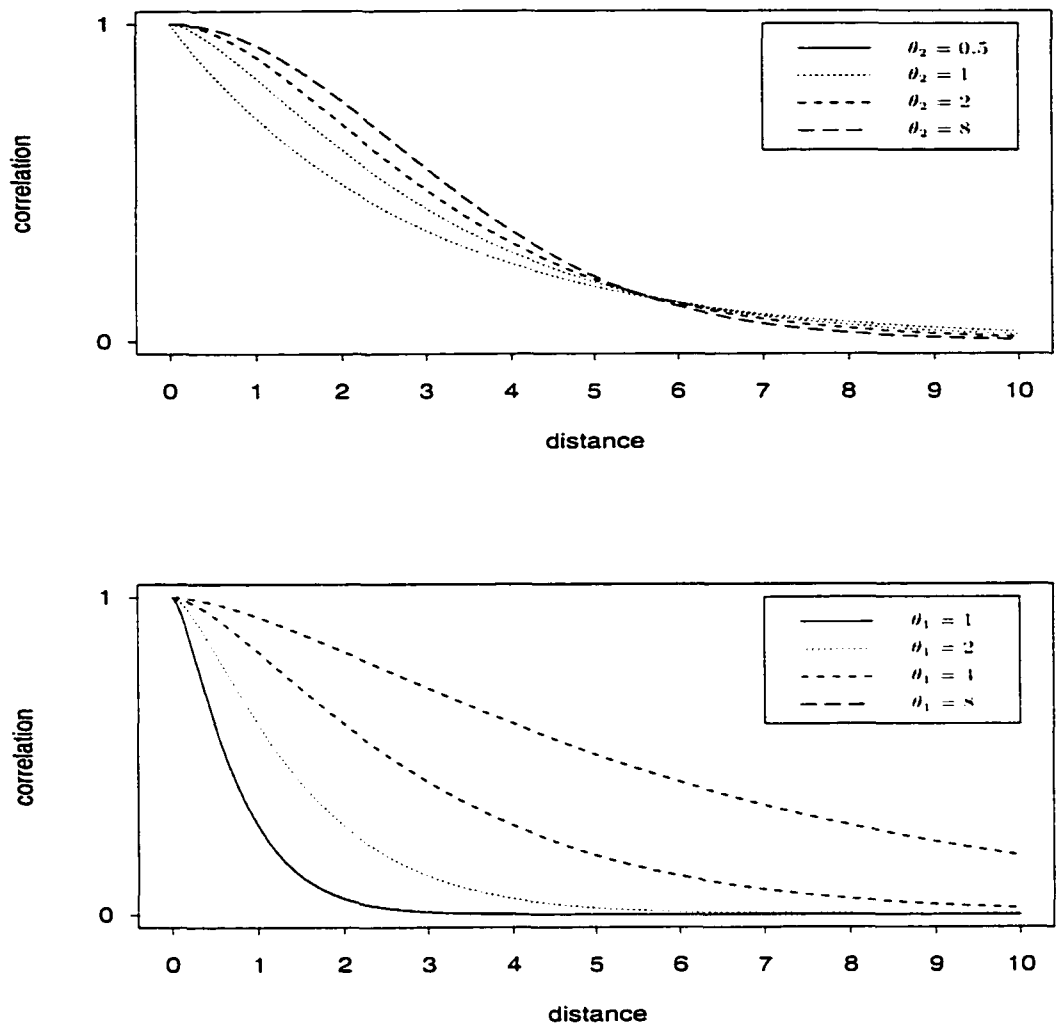

Figure 2.5: Examples of the Matern Autocorrelation Function. The top plot demonstrates the effect of smoothness (.5, 1, 2, 4, 8) on the Matern autocorrelation function for a fixed range 2.4. The bottom plot shows the effect of range on the autocorrelation function, with range values (1, 2, 4, 8) and smoothness 1. In both plots, $d = 1$.

The three functions, $\rho(\cdot)$, $\mathrm{Cov}(\cdot)$, and $\gamma(\cdot)$ are related through the following equations:

$$\rho(\boldsymbol{h}) = \mathrm{Cov}(\boldsymbol{h})/\mathrm{Cov}(\boldsymbol{0}), \text{ and} \tag{2.20}$$

$$2\gamma(\boldsymbol{h}) = 2\mathrm{Cov}(\boldsymbol{0}) - 2\mathrm{Cov}(\boldsymbol{h}), \tag{2.21}$$

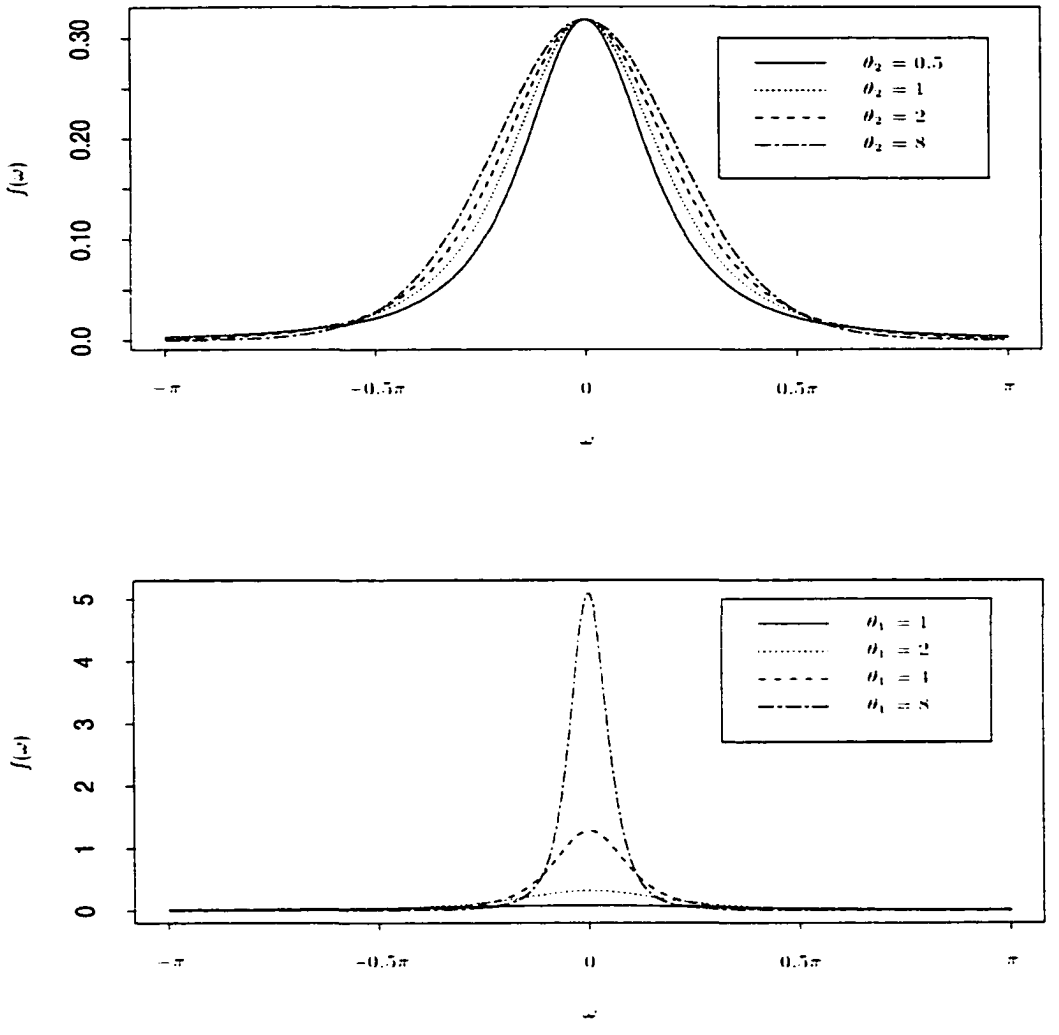

Figure 2.6: Examples of the Matern Spectral Densities. The top plot demonstrates the effect of smoothness (.5, 1, 2, 4, 8) on the Matern spectral density for a fixed range 2.4 when $d = 1$. The bottom plot shows the effect of range on the spectral density, with range values (1, 2, 4, 8) and smoothness 1 when $d = 1$.

where $\boldsymbol{h}$ is a vector of distances. When estimating the spatial parameters, Cressie (1993) argues that variogram estimation is better than correlogram estimation for the following reasons. First, the bias of the variogram estimator is less than the bias for the covariogram estimator (Cressie and Grondona, 1992). Second, if the mean function is not correctly specified and some portion of the mean function is

not modeled, the variogram estimator does a better job at modeling the spatial structure (Cressie, 1993, p.72).

The empirical variogram is one technique for estimating the spatial parameters of the autocorrelation function. Assuming the process is isotropic and second order stationarity (Section 2.1.1), the variogram is a function of the distance between locations, $d_{ij} = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$. The classical variogram estimator (Huijbregts and Matheron, 1971) is defined to be

$$2\hat{\gamma}(h) \equiv \frac{1}{|N(h)|} \sum_{N(h)} (Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j))^2 . \tag{2.22}$$

where $h \in \mathbb{R}$, and

$$N(h) \equiv \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : |\boldsymbol{s}_i - \boldsymbol{s}_j| \in (h + \epsilon, h - \epsilon); i, j = 1, \ldots, n\} . \tag{2.23}$$

where $(h - \epsilon, h + \epsilon)$ form the bins of size 2ϵ and $N(h)$ consists of all pairs of observations with $|\boldsymbol{s}_i - \boldsymbol{s}_j| \in (h \pm \epsilon)$. The classical variogram estimator calculates the average $(Z(s_i) - Z(s_j))^2$ for all pairs of observations for that bin. In general the bins are disjoint and cover the distances $(0, \frac{1}{2}d_{max})$ where d_{max} is the maximum distance between locations.

Estimation of the variogram curve using least squares techniques requires the minimization of

$$(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\boldsymbol{\theta}})' (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\boldsymbol{\theta}}) .$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}(b_1), \ldots, \hat{\gamma}(b_m))'$ is the vector of empirical variogram estimations. b_i denote the i^{th} bin, $i = 1, \ldots, m$ and $\boldsymbol{\gamma}_{\boldsymbol{\theta}} = (\gamma_{\boldsymbol{\theta}}(b_1), \ldots, \gamma_{\boldsymbol{\theta}}(b_m))'$ is the vector of variogram curves parameterized by $\boldsymbol{\theta}$ (Cressie, 1993). Similarly, the weighted least squares technique requires the minimization of

$$(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\boldsymbol{\theta}})' W(\boldsymbol{\theta})^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_{\boldsymbol{\theta}}) .$$

where $W(\boldsymbol{\theta})$ is a diagonal matrix with $[W(\boldsymbol{\theta})]_{ii}$ equal to the variance of the squared differences for all location pairs in bin b_i (Smith, 2000).

Determining the optimal size for the bins can be challenging (Cressie, 1993, Section 2.4). The number of pairs of observations in each bin varies greatly, but there should be enough observations so that the average squared difference does not rely strongly on one observation pair. Conversely, the bins need to be small enough to be able to estimate the spatial process. A general rule for bin size is that there should be at least 30 location pairs in each bin (Journel and Huijbregts, 1978, p.194).

Stein (1999, p.212-214) addressed one of the problems of the classical variogram estimator. The correlation between $\hat{\gamma}(b_i)$ and $\hat{\gamma}(b_{i+1})$ can be extremely high. Therefore, the empirical variogram can look dramatically different from the true variogram of the spatial process. This is also a difficulty with correlogram estimation.

2.3.2 Estimation of Autocorrelation Parameters when Mean is Known

Section 2.3.1 was concerned with estimating the autocorrelation function when the spatial process $\boldsymbol{Z}$ was second order stationary. When the mean function is known, the empirical variogram can be used to model the spatial process by subtracting the known mean from the response and using the residuals as inputs to the variogram.

Maximum likelihood estimation provides another method for estimation of the autocorrelation parameters. Likelihood based estimation requires the additional assumption that the error process δ from (1.1) follows some distribution. This thesis is based on the assumption that δ is a Gaussian random process. The log-likelihood of the parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2)$ given the data, $\boldsymbol{Z}$, is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2; \boldsymbol{Z}) = -\frac{1}{2}\log\left|\sigma^2\Psi\right| - \frac{1}{2\sigma^2}(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta})'\Psi^{-1}(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta}). \qquad (2.24)$$

where $\Psi = [\rho_{\theta}(s)]$ represents the matrix of correlations. In this situation, β is assumed know and is not a parameter of the model. This likelihood is a function of $q + 1$ parameters, where q is the dimension of θ.

Mardia and Marshall (1984) describe conditions required for consistency and asymptotic normality of the maximum likelihood spatial estimators using increasing domain asymptotics. The conditions involve eigenvalues of the covariance matrix and characteristics of the design matrix X, which are not easy to evaluate. Furthermore, problems using the technique for some correlation models such as the spherical family (Warnes and Ripley, 1987; Mardia and Watkins, 1989), where the likelihood is not twice differentiable, can lead to questions about the unimodality of the likelihood. Warnes and Ripley (1987) suggest that in order to investigate possible multimodality of the likelihood, the profile likelihood should be plotted.

One technique to reduce the parameter space by one is to substitute the maximum likelihood estimate of σ^2,

$$\hat{\sigma}^2 = (Z - X'\beta)\,\Psi^{-1}\,(Z - X'\beta)/n \tag{2.25}$$

into (2.24). The resulting log profile likelihood (Smith, 2000) is

$$\ell_{profile}(\theta : \hat{\sigma}^2, Z) = -\frac{1}{2}\log|\Psi| - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2}. \tag{2.26}$$

2.3.3 Estimation of Autocorrelation Parameters when Mean is Unknown

The estimation of the autocorrelation parameters when the mean function is unknown results in an added level of complexity. Techniques include an iterative process combined with the empirical variogram technique, maximum likelihood estimation and restricted maximum likelihood estimation.

Iterative Variogram Estimation

Any of the variogram estimation procedures described in Section 2.3.1 can incorporate an unknown mean function using a two step procedure. This commonly used approach begins with the computation of the residuals from the least squares estimate of $\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}$. Let $\boldsymbol{\Psi}^{-1}$ be the variance-covariance matrix estimated from the variogram.

1. Estimate the variogram curve of the residuals $\boldsymbol{Z} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}$.

2. Estimate the regression coefficients via generalized least squares, $\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}$.

Iteration between the steps one and two are continued until the spatial parameter estimates have converged (Haining, 1990, Section 4.2.1).

Maximum Likelihood Estimation

Maximum likelihood estimation when the mean is unknown is very similar to estimation when the mean is known (Smith, 2000). The log-likelihood, equation (2.24), can be simplified and the number of parameters reduced by examining estimation for both $\boldsymbol{\beta}$ and σ^2. Fixing $\boldsymbol{\theta}$, the maximum likelihood estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}$. Fixing both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the maximum likelihood estimate of σ^2 is $\hat{\sigma}^2 = (\boldsymbol{Z} - \boldsymbol{X}'\boldsymbol{\beta})\boldsymbol{\Psi}^{-1}(\boldsymbol{Z} - \boldsymbol{X}'\boldsymbol{\beta})/n$. Substituting the maximum likelihood estimates into the log likelihood results in the log profile likelihood.

$$\ell_{profile}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{Z}\right) = -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\log\left(\hat{\sigma}^2\right) - \frac{n}{2}. \tag{2.27}$$

The issue of multimodality of the likelihood still exists in maximum likelihood with unknown mean. In general, maximum likelihood estimates for the spatial parameters are biased and can underestimate the variance (Mardia and Marshall, 1984).

REML Estimation

A third method for estimation of the spatial parameters when the mean is unknown is restricted maximum likelihood estimation (REML) estimation. Patterson and Thompson (1971) first introduced REML as a technique for estimating variance components in ANOVA models. REML estimation is motivated by the following situation (Smith. 2000). Suppose $Y_i, i = 1, \ldots, n$ are independent and identically distributed Gaussian random variables with unknown mean μ and unknown variance σ^2. Maximum likelihood estimates of μ and σ^2 are $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu})$. Calculating the expected value of the two estimators. $E[\hat{\mu}] = \mu$ but $E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$.

We can separate the likelihood into two parts

$$\begin{aligned} \mathcal{L}(\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{1}{2\sigma^2}(\boldsymbol{Y} - \hat{\mu})'(\boldsymbol{Y} - \hat{\mu})\right\} \\ &\quad * \exp\left\{\frac{1}{2\sigma^2}(\hat{\mu} - \mu)'(\hat{\mu} - \mu)\right\} \end{aligned} \tag{2.28}$$

where the first line is a function of the data. and the second line is a function of the data and the true mean μ. In fact. the distribution of the contrasts $(Y_1 - \hat{\mu}, \ldots, Y_{n-1} - \hat{\mu})$ does not depend on the true value μ. and the estimator of σ^2 which maximizes this function is $\tilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{\mu})$. an unbiased estimator of σ^2.

Expanding this example to include correlated observations along with a more complex regression function. REML estimation examines the likelihood function of the contrasts (Smith. 2000) in order to avoid possible underestimation of the spatial parameters and account for the estimation of $\boldsymbol{\beta}$. In general. define $\boldsymbol{W} = \boldsymbol{A}'\boldsymbol{Z}$ to be a vector of linearly independent contrasts where $\boldsymbol{A}$ is a n by n matrix. Then following from (1.1). $\boldsymbol{W}$ is distributed Gaussian with mean zero and variance matrix $\boldsymbol{A}'\boldsymbol{\Psi}\boldsymbol{A}$. Patterson and Thompson (1971) show that $\boldsymbol{A}$ can be chosen such that

$AA' = I - X(X'X)^{-1}X$ and $A'A = I$. The corresponding log likelihood of the contrasts is (Harville, 1974)

$$\ell_{REML}(\sigma^2, \theta) \propto -\frac{n-q}{2}\log(\sigma^2) - \frac{1}{2}\log|X'\Psi^{-1}X| - \frac{1}{2}\log|\Psi| + \frac{1}{2\sigma^2}\left(Z - X'\hat{\beta}\right)'\Psi^{-1}\left(Z - X'\hat{\beta}\right). \quad (2.29)$$

where $\hat{\beta}$ is the generalized least squares estimate of β. Continuing with the profile likelihood techniques, we can use the maximum likelihood estimate of σ^2 which is $\tilde{\sigma}^2 = \left(Z - X'\hat{\beta}\right)'\Psi^{-1}\left(Z - X'\hat{\beta}\right)$ to further simplify the log likelihood (2.29) to be

$$\ell_{REML}(\theta; \tilde{\sigma}^2) \propto -\frac{n-q}{2}\log(\tilde{\sigma}^2) - \frac{1}{2}\log|X'\Psi^{-1}X| - \frac{1}{2}\log|\Psi|. \quad (2.30)$$

Equation (2.30) is then maximized for the parameters of the correlation function θ.

2.3.4 Comparison between ML and REML Techniques

Cressie (1993, p.92-93) supports the use of REML over ML as a method of estimation when the number of parameters in the mean function $(p + 1)$ is large. The differences between the two estimation techniques (2.27 and 2.29) include the coefficient of $\log(\sigma^2)$ as well as an additional term $-\frac{1}{2}|X'\Psi^{-1}X|$ in REML estimation. The estimate of σ^2 is biased for the ML estimator and unbiased for REML estimator. The differences in estimation between the REML and ML estimators of θ, the spatial correlation parameters, are more difficult to determine.

Zimmerman and Zimmerman (1991) used simulation methods to compare ML, REML and other estimation procedures to predict a random field. Data were simulated on a grid design with 16 or 36 total observations using two different correlation functions and a flat mean. These authors found that when simulating data using an exponential autocorrelation function, ML estimates of σ^2 have a negative bias while REML estimates have a positive bias, although the bias is larger for the REML estimates of σ^2.

Note that the difference in the denominator for the estimates of σ^2 is n for ML and $n-1$ for REML. When estimating the range parameter, θ_1, of the exponential autocorrelation function, the two estimators are similar for smaller values of θ_1. When θ_1 is larger, corresponding to stronger spatial dependence, the REML estimator has less bias. The Zimmerman and Zimmerman article examined fitting correlation functions when the mean function was constant, and so the differences between REML and ML should be few due to the few differences in the likelihoods. In Section 5.5 I investigate the behavior of REML and ML estimation when fitting a mean function with three parameters.

2.4 Model Selection

As described in Section 1.1, the modeling decisions in spatial data consist of selecting a set of explanatory variables for use in modeling the large scale variation, choosing a family of autocorrelation functions for modeling the spatial dependence, and finally estimation of both sets of parameters. Little research has been done on techniques or guidelines for making decisions regarding the modeling spatial data such as explanatory variable selection. Cressie (1993, p.104-498) notes that spatial model selection criteria is an area for future research.

2.4.1 Background and Traditional Methods

Standard techniques for modeling spatial data arise from traditional model selection techniques for linear models. In the first step, the data are assumed to be independent. Under this assumption, the explanatory variables for modeling the large scale variation are chosen using a model selection technique such as Akaike's Information Corrected Criteria (AICC) (Hurvich and Tsai, 1989) described in Section 2.4.1. Second, the residuals from the model are examined for spatial correlation. A correlation function family is chosen based on the residuals from the model as

well as information about the process (Section 2.1). Finally, the parameters are estimated using one of the techniques discussed in Section 2.3.

Explanatory Variables verses Autocorrelation Structure

The problems with the approach described above are many but are primarily concerned with the connection between explanatory variables and autocorrelation function. The trade off between inclusion or exclusion of explanatory variables and the choice of autocorrelation function can result in the fitting of two very distinct models, yet both can model the data reasonably well.

In other words, information about the correlation structure of the process may be carried through the explanatory variables. The inclusion of the explanatory variable may remove or reduce the correlation structure of the residuals from the model. For example, Ver Hoef et al. (1999) demonstrate the similarities between a model with independent errors and a linearly decreasing mean and a model with correlated errors and a constant mean.

Alternatively, ignoring the autocorrelation structure of the error process may mask explanatory variables which are very important in modeling the mean function. The additional noise in the data can overwhelm the information in the data, resulting in the identification of fewer important explanatory variables. Examples of this behavior will be explored in Section 5.2.

Also, ignoring the autocorrelation structure of the error process may lead to the identification of explanatory variables that are not related to the true model. Ver Hoef (1999) found that when analyzing the lizard data set in Section 6.2 that the model with independent errors selected a model with six explanatory variables using stepwise selection. When the model included autocorrelated errors, the number of significant explanatory variables dropped to two.

The AICC Statistic

One model selection technique used in this dissertation is AICC. This allows for simple comparison between models and the identification of a "best" model. Akaike's Information Criteria (AIC) (1973) was developed as an estimator of the Kullback-Liebler Information (KLI). If M_T is the true model, and M_a is a candidate model, AIC approximates the loss in information when M_a was used to model M_T.

Burnham and Anderson (1998) outline the development of the AIC statistic which follows from the identification of f_T, the true model or true behavior, and f_a, a candidate model based on the data $\boldsymbol{x}$ and parameters ϑ. The KLI is defined to be

$$I(f_T, f_a) = \int f_T(\boldsymbol{x}) \log\left(\frac{f_T(\boldsymbol{x})}{f_a(\boldsymbol{x}\,|\,\vartheta)}\right) d\boldsymbol{x}. \tag{2.31}$$

The goal is to select a model which minimizes the estimated KLI. The AIC statistic estimates (2.31) where

$$AIC = -2\log\left(L\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\,|\,\boldsymbol{Z}\right)\right) + 2k. \tag{2.32}$$

Here $L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\,|\,\boldsymbol{Z})$ is the likelihood of the parameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ given the data $\boldsymbol{Z}$ and k is the number of estimated parameters.

The corrected AIC statistic (AICC) (Hurvich and Tsai, 1989) includes an additional penalty term for when the ratio of the number of parameters to number of observations is large and performs better for small sample sizes. (Burnham and Anderson, 1998). The AICC statistic is

$$AICC = -2\log\left(L\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\,|\,\boldsymbol{Z}\right)\right) + 2k + \frac{2k(k+1)}{n-k-1}. \tag{2.33}$$

Other methods that could be used include R^2, Schwart's BIC criteria, and Mallows's C_p statistic (Neter *et al.*, 1996).

Due to it's simplicity, the AICC statistic can be used both when assuming the data has independent error structure, as well as when modeling the autocorrelation

structure of the data. In the next section I will propose a method for model selection for spatial data based on the AICC statistic which reduces some of the challenges in modeling decisions described in Section 2.4.1.

2.4.2 Spatial AICC

Model selection techniques for spatial models need to include the correlation structure in determining the best set of predictors. In order to address some of the challenges described in Section 2.4.1, I propose a technique which fits the mean and autocorrelation functions simultaneously. This method removes the ambiguity in determining which set of explanatory variables best fits the data when the process is correlated. This method combines ML estimation and the AICC statistic in order to arrive at a statistic which measures the ability of the model to fit the data.

Let M_i, $i = 1, \ldots, K$ be a set of explanatory variables along with an autocorrelation family. For each model M_i, use the profile likelihood (2.27)

$$\ell_{profile}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{Z}\right) = -\frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\log\left(\hat{\sigma}^2\right) - \frac{n}{2}.$$

to estimate the regression coefficients $\boldsymbol{\beta}$ and the autocorrelation parameters $\boldsymbol{\theta}$. Plug the estimated parameters into the AICC statistic (2.33).

$$AICC = -2\log\left(L\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\right)|\boldsymbol{Z}\right) + 2k + \frac{2k(k+1)}{n-k-1}.$$

where k is the number of parameters in $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ^2. Finally, use each model's AICC statistic as a tool for comparing models in terms of sets of explanatory variables as well as autocorrelation function families.

This technique requires parameter estimation of all of the models. By computing AICC for all possible sets of explanatory variables and autocorrelation functions, one can find a single "best" model or a set of models which fit the data well. Since this method simultaneously examines the explanatory variables while fitting

an autocorrelation structure to the data, it can overcome the correlated noise in the process and identify explanatory variables which are important in modeling the mean function. In Section 5.2 I describe a simulation study comparing spatial AICC model selection to traditional methods.

Chapter 3

BAYESIAN KRIGING WITH PROPER PRIORS

In this chapter I describe Bayesian kriging, the selection of prior distributions and hyperparameters, and the derivation of the posterior predictive distributions. In addition, I detail a fully Bayesian kriging setup, with proper prior distributions on both the regression parameters $\boldsymbol{\beta}$ and the spatial parameters $\boldsymbol{\theta}$. This setup results in a posterior distribution which incorporates the uncertainty in parameter estimation.

3.1 Prior Distributions

Recall equation (1.1) which models the random field as

$$Z(s) = \boldsymbol{X}(s)\boldsymbol{\beta} + \delta(s),$$

where $\delta(s)$ is assumed to be a stationary, isotropic Gaussian random field with mean zero, variance σ^2, and autocorrelation parameterized by $\boldsymbol{\theta}$. Standard Bayesian methods incorporate uncertainty of the parameters of the model into the analysis. The parameters are seen as random variables. In order to accomplish this, prior distributions will be placed on the parameters. If $f(\boldsymbol{Z}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ is the likelihood equation and $f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ is the prior distribution on the parameters, then using Bayes rule the posterior distribution of the parameters given the data $\boldsymbol{Z}$ is

$$f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|\boldsymbol{Z}) = \frac{f(\boldsymbol{Z}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})}{\int \cdots \int f(\boldsymbol{Z}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) d\boldsymbol{\theta} d\sigma^2 d\boldsymbol{\beta}}. \quad (3.1)$$

The prior distribution represents knowledge and uncertainty about the parameters, while the posterior distribution represents the updated distribution of the parameters after observing the data.

Let Δ be some quantity of interest, for example the random field at an unobserved location. The posterior distribution of Δ given the observed data is calculated as

$$f(\Delta|Z) = \int \cdots \int f(\Delta|Z, \beta, \sigma^2, \theta) f(\beta, \sigma^2, \theta|Z) \, d\theta d\sigma^2 d\beta . \qquad (3.2)$$

where $f(\Delta|Z, \beta, \sigma^2, \theta)$ is the conditional distribution of Δ given the data Z and the parameters. $f(\beta, \sigma^2, \theta|Z)$ is the posterior distribution of the parameters described in equation (3.1).

3.1.1 Types of Prior Distributions

Prior distributions are typically classified according to several characteristics. The term "proper" prior refers to a prior distribution that integrates to 1. An improper prior distribution integrates to infinity.

Prior distributions can be further classified as informative or noninformative. An noninformative prior contains no information about the parameters and therefore does not favor any value over any other. For example, $f(\beta) = c > 0$, for $-\infty < \beta < \infty$, is both improper and noninformative because $\int_{-\infty}^{\infty} f(\beta) d\beta = c \int_{-\infty}^{\infty} d\beta = \infty$, and $f(b_i) = f(b_j)$ for all $b_i, b_j \in \mathbb{R}$. A frequently used noninformative prior for scale parameter σ is $f(\sigma) = \sigma^{-1}$. In general, noninformative priors can be chosen via Jeffreys method (Berger, 1985, p.87-88) where

$$f(\theta) = [I(\theta)]^{\frac{1}{2}} , \qquad (3.3)$$

where $I(\theta)$ is the expected Fisher information matrix.

Noninformative and improper prior distributions are acceptable prior distributions as long as the resulting posterior distribution is proper. The issue of a proper posterior will be addressed in Section 3.1.3.

Calculation of the posterior distribution can be simplified by the use of conjugate priors. Certain likelihood functions and prior distributions result in closed form posterior distributions. When using conjugate priors, it is not necessary to calculate the denominator in equation (3.1) because the posterior distribution can be identified from the form of the numerator. For example, if the likelihood function of a linear model is Gaussian, then the conjugate prior distribution on (β, σ^2) is Gaussian and inverse gamma distributions, respectively, with β and σ^2 independent. The resulting posterior distribution is also Gaussian. Conjugate priors for a Gaussian random field consist of Gaussian for β, and inverse gamma for σ^2. The autocorrelation functions described in this thesis do not have conjugate priors.

3.1.2 Connection between Bayesian Kriging and Universal Kriging

Omre and Halvorsen (1989) describe Bayesian kriging as a bridge which connects simple kriging and universal kriging. Simple kriging refers to kriging when the mean function is assumed known, while universal kriging refers to kriging with a general mean function that is unknown (Section 1.2). Simple kriging can be seen as Bayesian kriging using prior distributions on β which place all the mass at the estimated location such that $f(\beta_i) = \hat{\beta}_i I_{(\beta_i = \hat{\beta}_i)}$, where $I_{(\beta_i = \hat{\beta}_i)} = \begin{cases} 1 & \beta_i = \hat{\beta}_i \\ 0 & \beta_i \neq \hat{\beta}_i \end{cases}$ for $i = 0, 1, \ldots, p$ and $\hat{\beta}$ is the known value of β. This corresponds to no uncertainty in the estimated mean function but it does not address the uncertainty in the estimated autocorrelation function.

On the other hand, universal kriging incorporates the uncertainty in the estimated mean function. Kitanidis (1986) first noted the relationship between Bayesian

kriging and universal kriging. He used noninformative priors on β, assumed that both σ^2 and θ were known and carried out the integration in equation (3.1). The resulting posterior distribution is identical to the universal kriging equations.

Assuming Z is a Gaussian random field with mean $X\beta$ and variance $\sigma^2\Psi$, the standard conditional distribution of Z_0 given Z

$$Z_0|\beta, \sigma^2, \theta, Z \sim N\left(\psi'\Psi^{-1}Z + (X_0 - X'\Psi^{-1}\psi)'\beta, \sigma^2(1 - \psi'\Psi^{-1}\psi)\right). \quad (3.4)$$

The posterior distribution of β given θ, σ^2 and Z is

$$\begin{aligned} f(\beta|\theta, \sigma^2, Z) &\propto f(Z|\beta, \sigma^2, \theta) f(\beta) \\ &\sim N\left(X\hat{\beta}, \sigma^2(X'\Psi^{-1}X)^{-1}\right) \end{aligned} \quad (3.5)$$

where $\hat{\beta}$ is the generalized least squares estimate of β and $f(\beta)$ is a noninformative prior on β. Therefore the posterior predictive distribution of $Z_0|\sigma^2, \theta, Z$ is

$$\begin{aligned} f(Z_0|\sigma^2, \theta, Z) &\propto f(Z_0|\beta, \sigma^2, \theta, Z) f(\beta|\theta, \sigma^2, Z) \\ &\sim N\left(\psi'\Psi^{-1}Z + (X_0 - X'\Psi^{-1}\psi)'\hat{\beta}, \sigma^2 G\right). \end{aligned} \quad (3.6)$$

where

$$G = \sigma^2\left(1 - \psi'\Psi^{-1}\psi + (X_0 - X'\Psi^{-1}\psi)'(X'\Psi^{-1}X)^{-1}(X_0 - X'\Psi^{-1}\psi)\right).$$

This distribution is identical to the universal kriging distribution in equation (1.4).

3.1.3 History of Bayesian Kriging

The use of Bayesian methodology applied to kriging first appeared in Kitanidis (1986) where uncertainty due to the regression parameters was incorporated into a Gaussian spatial process. Kitanidis examined the effect of using a noninformative prior $f(\beta) = c$ as well as the standard Gaussian conjugate prior distribution on β.

Additionally, the priors on σ^2 included a noninformative prior $f(\sigma^2) = 1/\sigma^2$ and the conjugate inverse gamma prior. Kitanidis addresses the integration of the spatial parameters $\boldsymbol{\theta}$ in general terms as well as estimating the autocorrelation parameters using the mode of the product of the marginal posterior distribution of the data given $\boldsymbol{\theta}$ times the prior distribution of $\boldsymbol{\theta}$. Specifics are not given here.

Handcock and Stein (1993) extend Bayesian kriging to include parameterization of the autocorrelation function and estimation of the spatial parameters. In their article, the autocorrelation function is modeled by the Matern autocorrelation function without a nugget term. Assuming independence of the mean and variance, Handcock and Stein use a prior of the form

$$f(\sigma^2, \boldsymbol{\beta}, \boldsymbol{\theta}) \propto \frac{g(\boldsymbol{\theta})}{(\sigma^2)^a} \tag{3.7}$$

where $a = 1$, $g(\boldsymbol{\theta})$ is the prior distribution on the spatial parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ in equation (2.18), the Matern autocorrelation function. Therefore, they have placed noninformative priors on $\boldsymbol{\beta}$ and σ^2. They use a uniform prior on both θ_1, the smoothness, and θ_2, the range. They also discuss an informative prior on θ_1 which would limit the support to the interval $(0.5, 2)$. The rationale being that they expect the process to be continuous ($\theta_2 > .5$) and less than twice differentiable (Handcock and Stein, 1993).

The issue of proper posteriors is a concern in the selection of prior distributions for the spatial parameters $\boldsymbol{\theta}$. Stein(1999, p.224) addressed the improper prior distribution $f(\theta_1, \theta_2, \sigma^2) = 1$ on $(0, \infty)^3$. Stein states that for any fixed values of σ^2 and θ_1, the posterior distribution of θ_1 is not proper.

Following the prior in equation (3.7), Berger, De Oliveira and Sanso (2000) gave necessary and sufficient conditions on the hyperparameter a when the prior distribution for the spatial parameters is flat, $f(\boldsymbol{\theta}) = 1$, in order to have proper posterior. When an intercept term is included in $\boldsymbol{\beta}$, then $a > 1 + (2\min\{1, \theta_2^{-1}\})^{-1}$, otherwise

$a > 1/2 + (2\min\{1, \theta_2^{-1}\})^{-1}$. Therefore, they note that the prior $f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) \propto 1$ as well as the priors where $a = 1$ in equation (3.7) and $g(\theta) \propto 1$ or $g(\theta) \propto \frac{1}{\theta}$ for all $\boldsymbol{\theta}$ result in improper posterior distributions.

Berger, De Oliverira and Sanso (2000) develop a reference prior for the parameters that result in proper posterior distributions. The reference prior is developed from the integrated likelihood

$$L^I(\sigma^2, \boldsymbol{\theta} | \boldsymbol{Z}) = \int_{R^{p+1}} f(\boldsymbol{Z} | \sigma^2, \boldsymbol{\theta}, \boldsymbol{\beta}) f(\boldsymbol{\beta} | \sigma^2, \boldsymbol{\theta}) d\boldsymbol{\beta} \quad (3.8)$$

$$\propto (\sigma^2)^{-\frac{n-(p+1)}{2}} |\boldsymbol{\Psi}|^{-1/2} |\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}|^{-1/2} \exp\left\{-\frac{S^2}{2\sigma^2}\right\}. \quad (3.9)$$

where $S^2 = \left(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \boldsymbol{\Psi}^{-1} \left(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$. $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimate of $\boldsymbol{\beta}$ assuming $\boldsymbol{\theta}$ is known, and $\boldsymbol{\Psi}$ is the known autocorrelation matrix . The reference prior is then

$$f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) \propto \left\{tr[\boldsymbol{W}^2] - \frac{1}{n-p-1}(tr[\boldsymbol{W}])^2\right\}^{\frac{1}{2}}. \quad (3.10)$$

where

$$\boldsymbol{W} = \left(\frac{\delta}{\delta\boldsymbol{\theta}}\boldsymbol{\Psi}\right)\boldsymbol{\Psi}^{-1}\boldsymbol{P}. \quad (3.11)$$

$$\boldsymbol{P} = I - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Psi}^{-1}. \quad (3.12)$$

and $\left(\frac{\delta}{\delta\boldsymbol{\theta}}\boldsymbol{\Psi}\right)$ is the matrix of derivatives with respect to the spatial parameters $\boldsymbol{\theta}$.

Other prior distributions include proper distributions for θ_2 such as the inverse gamma (Ecker and Gelfand, 1997) as well as truncating the range of the parameters so that the prior is proper. One possible truncation for θ_1 would be to limit the possible values to the observed distances. Therefore, the maximum distance between observed locations would be the maximum value for θ_1.

3.1.4 Selection of Prior Distributions

In this thesis I use proper priors for β, σ^2 and θ. Inclusion of prior distributions in Bayesian kriging offers improvements over universal kriging due to the inclusion of prior information and expert opinion about the parameters of the model. In this section I develop prior distributions and discuss sensitivity for the spatial model following results in Hoeting (1994) for linear models. In some cases, the prior distributions can have a direct effect on the corresponding parameter estimates.

I use the standard conjugate priors for β and σ^2, following the normal-gamma family (Berger, 1985, p.130). As discussed in the previous section, the spatial parameters θ do not have a conjugate family. A variety of prior distributions on θ will be examined, including uniform over a finite range and inverse gamma. I assume the following prior distributions on the regression parameters and variance:

$$\beta \mid \sigma^2, \theta \sim N(\mu, \sigma^2 V). \tag{3.13}$$

$$\frac{1}{\sigma^2}\Big| \theta \sim \frac{1}{\nu\lambda}\chi^2_\nu. \tag{3.14}$$

The selection of hyperparameters ν, λ, μ, and V will be discussed.

Hyperparameters for β

The design matrix X is standardized to simplify the determination of the hyperparameters μ and V. The design matrix consists of a column of ones corresponding to the overall mean β_0, as well as p columns corresponding to additional explanatory variables $\beta_1, \ldots, \beta_p$. In order to simplify prior specification, columns 2 through $p+1$ are scaled so that each column has zero mean and variance one. This allows each β value to reflect the strength of that explanatory variable in modeling the response Z. The distributions of β_1 through β_p are centered at zero. The prior distribution on the mean β_0 is centered at the sample mean of the response $\bar{Z}$.

In choosing hyperparameter V, the regression parameters are assumed independent a priori. This assumption simplifies the hyperparameter selection and reduces the problem to specifying the diagonal elements of V. The prior variance of each β_i, $i = 1, \ldots, p$ is $\sigma^2\upsilon^2$, and the prior variance of β_0 is the σ^2 times the variance of the response. Therefore,

$$\begin{aligned} \mu &= (\bar{Z}, 0, \ldots, 0)' \\ V &= \begin{pmatrix} s_Z^2 & 0 & \ldots & 0 \\ 0 & \upsilon^2 & 0 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \upsilon^2 \end{pmatrix} \end{aligned}$$

The hyperparameter υ is chosen so that the distribution of each β is reasonably flat over $(-1, 1)$. I will let $\upsilon = 2.85$ which follows from Hoeting (1994).

Hyperparameters for σ^2

The hyperparameters ν and λ are chosen so that the expected value of σ^2 is approximately equal to the variance of the response, s_Z^2, along with specifying that the distribution gives weight to reasonable values of σ^2. The relationship between the moments of the distribution of σ^2 and the hyperparameters yields the following equalities:

$$\begin{aligned} E[\sigma^2] &= \frac{\nu\lambda}{\nu - 2} \\ \text{Var}[\sigma^2] &= \frac{(\nu\lambda)^2}{(\nu-2)^2\left(\frac{\nu}{2} - 2\right)}. \end{aligned}$$

I estimate $E[\sigma^2]$ and $\text{Var}[\sigma^2]$, and therefore ν and λ from either the observed data, true values known from simulation, or knowledge about reasonable values of variance. $\text{Var}[\sigma^2]$ may be estimated so that a 95% confidence interval includes a

majority of reasonable values for σ^2. Using the method of moments, I equate the hyperparameters to the estimated mean and variance of σ^2, resulting in

$$\begin{aligned} \nu &= 2\left(\frac{E[\sigma^2]^2}{\text{Var}[\sigma^2]} + 2\right) \\ \lambda &= E[\sigma^2]\left(1 - \left(\frac{E[\sigma^2]^2}{\text{Var}[\sigma^2]} + 2\right)^{-1}\right). \end{aligned}$$

The goal of this prior is to give weight to reasonable values of σ^2. Small changes in the hyperparameters that retain the mean and variance of the prior distribution on σ^2 do not have a large effect. Large changes in the hyperparameters can have dramatic impact on the results. Proper prior distributions that are relatively flat can lead to improper posterior distributions (Berger *et al.*, 2000) and convergence problems. Therefore, the prior on σ^2 must be chosen carefully.

Prior Distribution for θ

There are a wide variety of autocorrelation functions and spatial parameters θ as described in Section 2.2. For purposes of this dissertation, the Matern autocorrelation function (equation (2.18)) will be used for all Bayesian kriging models. The spatial parameters of interest, θ, include the range, θ_1, smoothness, θ_2, and percentage of nugget, θ_3 in equation (2.11). The nugget is not included in all models ($\theta_3 = 0$). As described in Section 3.1.3, the choice of prior distribution for the autocorrelation parameters is very important and can lead to improper posterior distributions. One solution that leads to proper posterior distributions is to use conjugate priors for σ^2 as well as proper priors on θ.

The prior distribution on the range parameter, θ_1, is uniform on the interval $(0, d_{max})$, where d_{max} is the maximum distance between observed locations. I use two different prior distributions on the smoothness parameter, θ_2. First I use a uniform prior limiting θ_2 to be in $(0, c)$. The upper limit c is set so that a wide variety of

behaviors are possible. As can be seen in Figure 2.5, there is very little difference in the autocorrelation function for θ_2 values 4 and 8. In fact, for large values of θ_2, it is very difficult to distinguish between autocorrelation functions and therefore limiting the magnitude of θ_2 does not greatly influence the behavior near the origin. Another prior for θ_2 is an inverse gamma distribution. This prior distribution down weights large values of θ_2 more naturally than the uniform prior with an arbitrary upper limit.

Although large values of θ_2 are acceptable in the Matern class, they can cause problems in convergence and evaluation of parameters. It may be necessary to limit the domain of θ_1 or θ_2 or use a prior distribution which would down weight large values to avoid difficulties in singularity of the autocorrelation matrix. Some possibilities for the maximum of the domain for θ_1 include the maximum distance between points, or some function of the maximum distance between points. This is an area of future research when using automatic methods to fit covariance functions. In particular, Stein (1999, p173) states that the Matern class may not "possess a maximum in the interior of the parameter space."

When θ_3 is included in the analysis, one prior is a uniform distribution on the interval (0, 1) which gives equal weight to all possible values. When $\theta_3 = 0$, there is no additional nugget variability in the model, and when $\theta_3 = 1$, the spatial process does not have any detectable spatial autocorrelation in the error. The beta distribution can also be used as a prior distribution for θ_3. This could be used to include expert opinion via increasing the weight for 'reasonable' values of θ_3 while down weighting large values.

There is an interaction between the inclusion of θ_3 and the prior distribution on θ_2. Including θ_3 in the analysis can cloud information about the smoothness of the process. Therefore, a prior distribution which places more weight on smaller values of θ_2 is required to prevent convergence problems. Simulations which include measurement error will use an inverse gamma distribution.

3.2 Derivation of Distributions

Once the prior distributions have been selected in Section 3.1.4, the posterior predictive density of Z_0 given the data $\boldsymbol{Z}$ and a set of explanatory variables, the optimal predictor $E[\,Z_0|\boldsymbol{Z}]$ is found via

$$f(\,Z_0|\boldsymbol{Z}) = \int\cdots\int f(\,Z_0|\boldsymbol{\beta}\,,\sigma^2,\boldsymbol{\theta},\boldsymbol{Z})\,f(\boldsymbol{\beta}\,,\sigma^2|\boldsymbol{\theta},\boldsymbol{Z})\,f(\boldsymbol{\theta}|\boldsymbol{Z})\,d\boldsymbol{\theta}\,d\boldsymbol{\beta}\,d\sigma^2. \quad (3.15)$$

In this section I derive a closed form expression for (3.15). The derivations are similar to Hoeting (1994). The following sections will derive the required distributions.

3.2.1 Posterior Prior Distributions

I need to derive the posterior prior distribution for $\boldsymbol{\beta}$ and σ^2, $f(\boldsymbol{\beta}\,,\sigma^2|\boldsymbol{\theta},\boldsymbol{Z})$ from equation (3.15).

$$\begin{aligned}
f(\boldsymbol{\beta}\,,\sigma^2|\boldsymbol{\theta},\boldsymbol{Z}) &= f(\boldsymbol{Z}|\boldsymbol{\beta}\,,\sigma^2,\boldsymbol{\theta})\,f(\boldsymbol{\beta}\,|\sigma^2,\boldsymbol{\theta})\,f(\sigma^2|\boldsymbol{\theta}) \\
&= |\sigma^2\boldsymbol{\Psi}|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\,)'\boldsymbol{\Psi}^{-1}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\,)\right\}|\sigma^2\boldsymbol{V}|^{-1/2} \\
&\quad \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}\,-\boldsymbol{\mu}\,)'\boldsymbol{V}^{-1}(\boldsymbol{\beta}\,-\boldsymbol{\mu}\,)\right\}\left(\frac{\nu\lambda}{\sigma^2}\right)^{\frac{\nu}{2}-1}\exp\left\{-\frac{\nu\lambda}{2\sigma^2}\right\} \\
&= (\sigma^2)^{-(n+p)/2}|\boldsymbol{\Psi}|^{-1/2}|\boldsymbol{V}|^{-1/2}\left(\frac{\nu\lambda}{\sigma^2}\right)^{\frac{\nu}{2}-1} \qquad (3.16) \\
&\quad \exp\left\{-\frac{1}{2\sigma^2}\left[\nu\lambda+(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\,)'\boldsymbol{\Psi}^{-1}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta}\,)\right]\right\} \\
&\quad \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta}\,-\boldsymbol{\mu}\,)'\boldsymbol{V}^{-1}(\boldsymbol{\beta}\,-\boldsymbol{\mu}\,)\right]\right\}.
\end{aligned}$$

Note that the following holds:

$$
\begin{aligned}
(Z-X\beta)'\ \Psi^{-1}\ (Z-X\beta)+(\beta-\mu)'V^{-1}(\beta-\mu)+\nu\lambda & \\
= \ & Z'\Psi^{-1}Z+\mu'V^{-1}\mu+\beta'(X'\Psi^{-1}X+V^{-1})\beta+\nu\lambda \\
& -(X'\Psi^{-1}Z+V^{-1}\mu)'\beta-\beta'(X'\Psi^{-1}Z+V^{-1}\mu) \\
& +(X'\Psi^{-1}Z+V^{-1}\mu)'(X'\Psi^{-1}X+V^{-1})^{-1}(X'\Psi^{-1}Z+V^{-1}\mu) \\
& -(X'\Psi^{-1}Z+V^{-1}\mu)'(X'\Psi^{-1}X+V^{-1})^{-1}(X'\Psi^{-1}Z+V^{-1}\mu) \\
= \ & Z'\Psi^{-1}Z+\mu'V^{-1}\mu+\nu\lambda \\
& -(X'\Psi^{-1}Z+V^{-1}\mu)'(X'\Psi^{-1}X+V^{-1})^{-1}(X'\Psi^{-1}Z+V^{-1}\mu) \\
& +(\beta-\mu^*)'(X'\Psi^{-1}X+V^{-1})(\beta-\mu^*) \\
= \ & \nu^*\lambda^*+(\beta-\mu^*)'V^{*-1}(\beta-\mu^*).
\end{aligned}
$$

where

$$\mu^* = (X'\Psi^{-1}X+V^{-1})^{-1}(X'\Psi^{-1}Z+V^{-1}\mu) \tag{3.17}$$

$$V^* = (X'\Psi^{-1}X+V^{-1})^{-1} \tag{3.18}$$

$$\nu^* = n+\nu \tag{3.19}$$

$$\lambda^* = \frac{1}{n+\nu}\left(Z'\Psi^{-1}Z+\mu'V^{-1}\mu-(X'\Psi^{-1}Z+V^{-1}\mu)'\mu^*+\nu\lambda\right) \tag{3.20}$$

So from equation (3.16)

$$
\begin{aligned}
f(\beta,\sigma^2,|\theta,Z) \ \propto \ & (\sigma^2)^{-(n+p)/2}\left(\frac{\nu\lambda}{\sigma^2}\right)^{\frac{\nu}{2}-1}|\Psi|^{-1/2}|V|^{-1/2} \\
& \exp\left\{-\frac{1}{2\sigma^2}\left[\nu^*\lambda^*(\beta-\mu^*)'V^{*-1}(\beta-\mu^*)\right]\right\} \\
= \ & f(\beta\,|\sigma^2,\theta,Z)\,f(\sigma^2|\theta,Z).
\end{aligned} \tag{3.21}
$$

where $(\beta\,|\sigma^2,\theta,Z)\sim N(\mu^*,\sigma^2V^*)$ and $(\frac{\nu^*\lambda^*}{\sigma^2}|\,\theta,Z)\sim\chi^2_{\nu^*}$.

The posterior distribution for θ, $f(\theta|Z)$ from equation (3.15) can not be calculated in closed form. Approximations for $f(\theta\,|Z)$ will be suggested in Section 4.4.3.

3.2.2 Posterior Predictive Distribution

Next, I need to calculate $f(\ Z_0|\theta, Z)$, the posterior predictive distribution conditional on θ used in equation (3.15).

$$f(\ Z_0|\theta, Z) \ = \ \int\int f(\ Z_0|\beta\ , \sigma^2, \theta, Z)\, f(\beta\ , \sigma^2|\theta, Z)\, d\beta\ d\sigma^2 \qquad (3.22)$$

To begin the derivations, I know that

$$Z_0|\beta\ , \sigma^2, Z, \theta \ \sim \ N\left(\psi'\Psi^{-1}Z + (X_0 - X'\Psi^{-1}\psi)'\beta\ , \sigma^2(1-\psi'\Psi^{-1}\psi)\right)$$

$$\beta\ |\sigma^2, \theta, Z \ \sim \ N(\mu^*, V^*) \qquad (3.23)$$

$$\left.\frac{\nu^*\lambda^*}{\sigma^2}\right| \theta, Z \ \sim \ \chi^2_{\nu^*}. \qquad (3.24)$$

where ν^*, λ^*, μ^* and V^* are defined in equations (3.19), (3.20), (3.17) and (3.18). In integrating out β, I can resort to consideration of a standard regression model. If Z_0 and Z are jointly distributed Gaussian random variables, then the distribution of Z_0 conditional on Z and the parameters β, σ^2 and θ equals $\psi'\Psi^{-1}Z + (X_0 - X'\Psi^{-1}\psi)'\beta + \epsilon$, where ϵ and β are independent, and distributed $\epsilon \sim$ Normal $(0, \sigma^2(1-\psi'\Psi^{-1}\psi))$ and $\beta \sim$ Normal $(\mu^*, \sigma^2 V^*)$. Then

$$E[\ Z_0\,|Z] = \psi'\Psi^{-1}Z + (X_0 - X'\Psi^{-1}\psi)'\mu^* \qquad (3.25)$$

and

$$\text{Var}[Z_0\,|Z] = \sigma^2\left[(X_0 - X'\Psi^{-1}\psi)'V^*(X_0 - X'\Psi^{-1}\psi) + (1-\psi'\Psi^{-1}\psi)\right]. \quad (3.26)$$

Note that since β and ϵ are both normally distributed, then Z_0 is normally distributed with mean $\left(\psi'\Psi^{-1}Z + (X_0 - X'\Psi^{-1}\psi)'\mu^*\right)$, and variance

$$\sigma^2\left((X_0 - X'\Psi^{-1}\psi)'V^*(X_0 - X'\Psi^{-1}\psi) + (1-\psi'\Psi^{-1}\psi)\right).$$

Then, I need to integrate out σ^2.

$$
\begin{aligned}
f(\,Z_0|\boldsymbol{Z},\boldsymbol{\theta}) &= \int f(\,Z_0|\sigma^2,\boldsymbol{Z},\boldsymbol{\theta})\, f(\sigma^2|\boldsymbol{\theta},\boldsymbol{Z})\, d(\sigma^2) \\
&= \int (2\pi\sigma^2)^{-1/2} \\
&\quad \left|(\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\, \boldsymbol{V}^*\, (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}) + (1 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right|^{-1/2} \\
&\quad \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\,Z_0 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} - (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\,\boldsymbol{\mu}^*\right)'\right.\right. \\
&\quad \left((\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\, \boldsymbol{V}^*\, (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}) + (1 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right)^{-1} \\
&\quad \left.\left.\left(\,Z_0 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} - (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\,\boldsymbol{\mu}^*\right)\right]\right\} \\
&\quad \frac{\left(\frac{\nu^*\lambda^*}{2}\right)^{\nu^*/2}}{\Psi\left(\frac{\nu^*}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu^*}{2}-1} \exp\left\{-\frac{\nu^*\lambda^*}{2\sigma^2}\right\} d\left(\frac{1}{\sigma^2}\right) \\
&= \int \left|\lambda^*\left[(\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\, \boldsymbol{V}^*\, (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}) + (1 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right]\right|^{-1/2} \\
&\quad (2\pi)^{-1/2}\, \frac{\left(\frac{\nu^*}{2}\right)^{\nu^*/2}}{\Psi\left(\frac{\nu^*}{2}\right)} \left(\frac{\lambda^*}{\sigma^2}\right)^{\frac{\nu^*}{2}-1} \left(\frac{\lambda^*}{\sigma^2}\right)^{1/2} \\
&\quad \exp\left\{-\frac{\lambda^*}{\sigma^2}\frac{1}{2\lambda^*}\left[\left(\,Z_0 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} - (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\,\boldsymbol{\mu}^*\right)'\right.\right. \\
&\quad \left((\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\, \boldsymbol{V}^*\, (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}) + (1 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right)^{-1} \\
&\quad \left.\left.\left(\,Z_0 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} - (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\,\boldsymbol{\mu}^*\right) + \nu^*\lambda^*\right]\right\} d\left(\frac{\lambda^*}{\sigma^2}\right).
\end{aligned}
$$

Let

$$
\begin{aligned}
s &= \frac{\lambda^*}{\sigma^2}\frac{1}{2\lambda^*}\left\{\left[\,Z_0 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} - (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\,\boldsymbol{\mu}^*\right]'\right. \\
&\quad \left[(\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\, \boldsymbol{V}^*\, (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}) + (1 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right]^{-1} \\
&\quad \left.\left[\,Z_0 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} - (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\,\boldsymbol{\mu}^*\right] + \nu^*\lambda^*\right\} \\
\text{and } ds &= \frac{1}{2\lambda^*}\left\{\left[\,Z_0 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} - (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\,\boldsymbol{\mu}^*\right]'\right. \\
&\quad \left[(\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\, \boldsymbol{V}^*\, (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}) + (1 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right]^{-1} \\
&\quad \left.\left[\,Z_0 - \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z} - (\boldsymbol{X}_0 - \boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\,\boldsymbol{\mu}^*\right] + \nu^*\lambda^*\right\} d\left(\frac{\lambda^*}{\sigma^2}\right).
\end{aligned}
$$

Then,

$$
\begin{aligned}
f(\,Z_0|\boldsymbol{\theta}.\boldsymbol{Z}) \;=\; & \int (2\pi)^{-1/2}(\lambda^*)^{-1/2}\frac{\left(\frac{\nu^*}{2}\right)^{\frac{\nu^*}{2}}}{\Gamma\left(\frac{\nu^*}{2}\right)}\frac{\Gamma\left(\frac{\nu^*}{2}\right)}{\Gamma\left(\frac{\nu^*+1}{2}\right)} \\
& \left|(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{V}^*(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})+(1-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right|^{-1/2} \\
& \left\{\frac{1}{2\lambda^*}\left[\left(\,Z_0-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}-(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{\mu}^*\right]'\right.\right. \\
& \left[(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{V}^*(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})+(1-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right]^{-1} \\
& \left.\left[\,Z_0-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}-(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{\mu}^*\right]+\nu^*\lambda^*\right]\right\}^{-(\nu^*+1)/2} \\
& s^{(1+\nu^*)/2-1}\exp^{-s}\,ds \qquad (3.27) \\
= \; & \left|\lambda^*\left[(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{V}^*(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})+(1-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right]\right|^{-1/2} \\
& \Gamma\left(\frac{\nu^*}{2}\right)^{-1}\Gamma\left(\frac{1+\nu^*}{2}\right)(\pi)^{-1/2}(\nu^*)^{\nu^*/2} \\
& \left\{\left[\,Z_0-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}-(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{\mu}^*\right]'\right. \\
& \left[(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{V}^*(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})+(1-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right]^{-1} \\
& \left.\left[\,Z_0-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}-(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{\mu}^*\right]+\nu^*\lambda^*\right\}^{-(1+\nu^*)/2}. \qquad (3.28)
\end{aligned}
$$

This is a non-central t distribution with ν^* degrees of freedom and

$$
\begin{aligned}
\mathrm{E}[\,Z_0|\boldsymbol{\theta}.\boldsymbol{Z}] \;&=\; \boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{Z}+(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{\mu}^* \qquad (3.29) \\
\mathrm{Var}[\,Z_0|\boldsymbol{\theta}.\boldsymbol{Z}] \;&=\; \frac{\nu^*\lambda^*}{\nu^*-2} \qquad (3.30) \\
& \quad \left[(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})'\boldsymbol{V}^*(\boldsymbol{X}_0-\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})+(1-\boldsymbol{\psi}'\boldsymbol{\Psi}^{-1}\boldsymbol{\psi})\right].
\end{aligned}
$$

3.2.3 MLE Approximations

Now we have an expression for $f(Z_0|\boldsymbol{\theta}.\boldsymbol{Z})$ and the final step in evaluating (3.15) is to find $f(\,Z_0|\boldsymbol{Z}) = \int f(\,Z_0|\boldsymbol{\theta}.\boldsymbol{Z})\,f(\boldsymbol{\theta}|\boldsymbol{Z})\,d\boldsymbol{\theta}$. There are two major difficulties with this integral. First, a closed form expression for $f(\boldsymbol{\theta}|\boldsymbol{Z})$, the posterior distribution of $\boldsymbol{\theta}$ is unavailable. I could approximate this distribution using either the Laplace or BIC approximation which will be described in Section 4.4.3, but I

would still need to approximate $\int f(\ Z_0|\boldsymbol{\theta}, \boldsymbol{Z})\, f(\boldsymbol{\theta}|\boldsymbol{Z})\, d\boldsymbol{\theta}$. which is then an approximation of an approximation.

A more direct solution is to use the maximum likelihood estimation approximation. where

$$f(\ Z_0|\boldsymbol{Z}) \ \approx \ f\Big(\ Z_0|\hat{\boldsymbol{\theta}}, \boldsymbol{Z}\Big). \tag{3.31}$$

where $\hat{\boldsymbol{\theta}}$ is the posterior mode of $f(\boldsymbol{\theta}|\ \boldsymbol{Z}) \ \propto \ f(\ Z|\boldsymbol{\theta})\, f(\boldsymbol{\theta})$. A similar approach was used in the context of Bayesian model averaging for censored survival models (Volinsky. 1997).

Chapter 4

BAYESIAN MODEL AVERAGING FOR SPATIAL MODELS

4.1 Introduction

This chapter builds on the previous chapters and considers the important problem of accounting for uncertainty in explanatory variable selection. For the purposes of this thesis and the interpretation of Bayesian model averaging (BMA), I define a model to have a mean function specified by a set of explanatory variables and an autocorrelation function modeled using the Matern autocorrelation function. Traditionally, modeling spatial data consists of the following steps: first a "best" set of explanatory variables is chosen using the techniques described in Section 2.4. Second, the parameters are estimated via methods in Section 2.3.1. Third, kriging technology is used to predict the process at unobserved locations. When using the Matern autocorrelation function and placing prior distributions on the parameters, the uncertainty in the parameter estimates of both the regression function and autocorrelation function are accounted for, but the uncertainty in the model is ignored. Failing to account for model uncertainty can lead to underestimation of the prediction error.

In many applications of statistics, there may be a number of models that fit equally well according to model selection criteria. The choice of one "best" model may not be clear. At times, a researcher may propose two or more models which fit the data equally well, but have different interpretations and provide different

predictions. How can the results from different models be incorporated into one set of predictions?

One approach to account for model uncertainty is Bayesian model averaging (BMA), a technique that allows for the incorporation of model uncertainty into prediction by averaging the results from a number of different models. This chapter is organized as follows: the history of combining models for improving prediction is first described. Then the general BMA framework and some associated issues are provided. Finally, BMA for spatial models together with the required derivations of distributions and approximations necessary for analysis is developed.

4.2 History of Combining Models

Volinsky (1997) is a good source on the history of combining models. The origins of combining models can be found in Laplace's Deuxieme Supplement a la Theorie Analytique des Probabilities. Stigler (1973) provides the translation of Laplace (1818) showing how "two estimators could be combined to provide a new estimator which would be better than either." This is not model averaging, but formed a basis for future work. Cochran (1937) established the foundation for incorporating estimators from multiple samples. In the quality control literature, Bates and Granger (1969) were the first to combine results from different models fit to the same data. Articles on combining predictions from different models soon followed. Bates and Granger (1969) demonstrated that the the weighted average of two model forecasts is better than forecasts from the individual models as long as each model contains some "independent information."

The philosophy of BMA and model uncertainty can be traced to two articles from the statistical community. Roberts (1965) proposed a distribution which combines the results of two models, foreshadowing the development of BMA for multiple models. Volinsky credits Leamer (1978) with addressing the idea that "averaged

estimators incorporate the ambiguity about the correct model into the posterior distribution." verifying the philosophy that averaging over models can account for model uncertainty.

A variety of articles have reviewed methodology for BMA and have addressed some of the problems with selecting one 'true model.' Draper (1995) focuses on Bayesian hierarchical models, building an expanded modeling structure where the highest level of structure corresponds to the probability of a specific model. Chatfield (1995) challenges the statistical community to account for model uncertainty, stating that failure to do so may be "more serious than other sources of uncertainty." Kass and Raftery (1995) focus on Bayes factors, the ratio of posterior odds of a model to its prior odds, as a tool for comparing and combining models. Hoeting et al. (1999) summarize BMA methodology.

Burnham and Anderson (1998) detail a frequentist model averaging framework which uses the AICC statistic to estimate the Kullback-Liebler Information (KLI) to provide model weights, detailed in Section 2.4.1. This framework does not average over all models, instead it focuses on 'good' models and incorporating expert opinion.

In the last decade, a variety of methods, both Bayesian and frequentist, have been proposed that average the results from multiple models. Implementation of BMA for spatial data will be described in the following sections.

4.3 Implementing Bayesian Model Averaging

Following the model development in Section 1.2, the underlying random field on $\mathbb{R}^d$ is assumed to be stationary and Gaussian. The goals remain the same: from data collected at n sites, predict the spatial process at unobserved locations and accurately assess the uncertainty of the estimate. As discussed in Section 4.2, combining the results from different models can improve the calibration of the models and account for model uncertainty.

4.3.1 General Bayesian Model Averaging Framework

A first step in accounting for model uncertainty is to define a set of models $\mathcal{M}$ for prediction. Let $\mathcal{M}$ be the "model space": the set of models under consideration, where $\mathcal{M} = (M_1, M_2, \ldots, M_K)$. For spatial data, the models M_i in this set $\mathcal{M}$ may differ by including different sets of explanatory variables, autocorrelation functions, or transformations on the response Z.

The set of models is determined by a variety of methods, including all possible models given the explanatory variables and models deemed important by previous studies or other scientific considerations. In this thesis, the model space will include all possible subsets of explanatory variables. If there are five explanatory variables, all fit using the same family of autocorrelation functions, then the total number of models is $2^5 = 32$.

The second step is to compute the BMA posterior predictive distribution. Define Δ to be a quantity of interest, such as a spatial process at an unobserved location. Note that Δ must have the same meaning for all models.

The posterior distribution of Δ given data $\boldsymbol{Z}$ is

$$f(\Delta|\boldsymbol{Z}) = \sum_{k=1}^{K} f(\Delta|M_k, \boldsymbol{Z}) f(M_k|\boldsymbol{Z}), \tag{4.1}$$

where $f(M_k|\boldsymbol{Z})$ is the posterior model probability (PMP) for model M_k and $f(\Delta|M_k, \boldsymbol{Z})$ is the predictive distribution of the quantity of interest given the data under a specific model M_k. This BMA posterior distribution is a weighted average of the posterior predictive distributions for each model, where the weights are specified by the PMP.

The PMP is determined using Bayes' Rule from the product of the prior model probability $f(M_k)$ and the integrated likelihood,

$$f(M_k|\boldsymbol{Z}) = \frac{f(\boldsymbol{Z}|M_k) f(M_k)}{\sum_{l=1}^{K} f(\boldsymbol{Z}|M_l) f(M_l)}. \tag{4.2}$$

The selection of the prior distribution on the model space, $f(M_k)$, will be discussed in Section 4.3.3. The integrated likelihood for each model is given by

$$f(\boldsymbol{Z}\,|M_k) = \int f(\boldsymbol{Z}\,|\Omega_k, M_k)\, f(\Omega_k\,|M_k)\, d\Omega_k \tag{4.3}$$

where $f(\boldsymbol{Z}\,|\Omega_k, M_k)$ is the distribution of the data given a model M_k and the parameters of that model Ω_k, and $f(\Omega_k\,|M_k)$ is the prior distribution on the parameters. In this spatial modeling framework, Ω_k represents the regression parameters and the autocorrelation function parameters, i.e., $\Omega_k = (\boldsymbol{\beta}_k, \sigma^2, \theta)$. Here $\boldsymbol{\beta}_k$ corresponds to the explanatory variables specified by M_k.

The BMA framework accounts for model uncertainty, conditional on the set of models being considered. In the next section, I address some of the issues associated with the model space, $\mathcal{M}$. This includes methods for exploring the model space when the number of models is large. A large number of models can quickly become an issue when the data set includes a large number of explanatory variables.

4.3.2 Approaches for Large Model Space

Although this thesis focuses on averaging over all possible models, there are situations where the number of models in the model space $\mathcal{M}$ is so large that averaging over all models is not viable. Methods for either reducing the model space or approximating (4.1) have been developed. Some of the methods include Occam's window (Madigan and Raftery, 1994; Hoeting, 1994), leaps and bounds (Volinsky *et al.*, 1997; Volinsky, 1997), Markov Chain Monte Carlo model composition (MC^3) (Madigan and York, 1995; Hoeting, 1994) and stochastic search variable selection (SSVS) (George and McCulloch, 1993). In what follows I focus on the model space defined as the set of all possible explanatory variables.

Occam's window is built on two ideas. First, a typical situation is that most of the models in the model space $\mathcal{M}$ do not predict as well as smaller subset of

models in $\mathcal{M}$. These poor-fitting models can be removed from the model space. Therefore, the posterior predictive distribution does not average over any models who's fit is inadequate. The second idea is Occam's Razor: if the probability of two models given the data are equal, then the simpler model is better. These two principles can reduce the model space significantly depending on the user's definition of poor-fitting and equivalent models.

Leaps and bounds (Furnival and Wilson, 1974) is an algorithm for linear regression that reduces the model space by 'trimming' sections of model space that do not fit well. Leaps and bounds is used by Volinsky et al. (1997) to quickly identify a subset of models to be used in (4.1).

On the other hand, some methods approximate (4.1) via a stochastic search algorithm. For example, MC^3 builds a stochastic process that travels through model space. In this approach, the proportion of time spent at each model should approximate the posterior model probability. Furthermore, an approximation of the average of the posterior predictive distribution is the average posterior prediction for each step in the stochastic process.

Another method which uses a stochastic search methodology is SSVS. SSVS embeds the regression set-up into a hierarchical Bayes normal mixture model. In this framework, a set of latent variables specifies the inclusion of explanatory variables. As the chain progresses, the frequency of specific sets of latent variables is useful in identifying promising models. Sets of explanatory variables that appear often fit the data better.

A more complete description of methods to use when the model space is large can be found in Hoeting et al. (1999). In summary, both the Occam's window and leaps and bounds approaches average over a smaller subset of the models which are upheld from the data. When the number of explanatory variables is large, i.e., $K > 30$, these methods "are too expensive computationally or do not explore a large enough region of the model space" (Clyde, 1999).

The last two methods require Markov chain Monte Carlo or importance sampling techniques to travel through the model space. Clyde (1999) points out that these methods "can be viewed as special cases of reversible jump MCMC algorithms." In these algorithms, the chain travels through model space as well as parameter space of potentially different dimensions.

4.3.3 Priors on the Models

The specification of a prior distribution on the model space is the last step in BMA. Like the prior distributions described in Section 3.1, both uninformative and informative priors can be used.

The inclusion of informative priors on the model space has been successful in improving predictive performance (Spiegelhalter *et al.*, 1993; Lauritzen *et al.*, 1994). When the models include different numbers of explanatory variables, it is possible to place a prior on the model space that will assign different probabilities on models with specific variables. For example, let

$$f(M_i) = \prod_{j=1}^{p} \pi_j^{\delta_{ij}} (1 - \pi_j)^{1-\delta_{ij}}. \tag{4.4}$$

where p is the total number of explanatory variables, $\pi_j \in [0, 1]$ is the prior probability that the coefficient for predictor j, β_j, does not equal 0, and δ_{ij} equals one if explanatory variable j is in model M_i and zero otherwise (Hoeting *et al.*, 1999). If $\pi_j = 0.5$ for all j, this prior corresponds to the uniform prior described above. If $\pi_j > 0$ for all j, models with a large number of explanatory variables have less weight. Expert opinion can also be included by using different values of π_j for different j, such as if $\pi_j = 1$, explanatory variable j is forced to be included in all models. This prior does assume that the inclusion of each explanatory variable is independent of other explanatory variables, although a more complex prior distribution could be created if necessary.

The prior distribution on the model space used in this thesis is the assumption that all models are equally likely, so that $p(M_i) = \frac{1}{K}$. Draper (1999) suggested a potential problem with this uniform prior for model space. If two explanatory variables are highly correlated, there may be some duplication in the models, a model is essentially counted twice by the prior. There are two possible situations in this instance (Hoeting *et al.*, 1999). First, the two explanatory variables could have significantly different interpretations and mechanisms and therefore describe the process from different points of view. In this instance, both models are useful and should be included. Second, if the two explanatory variables are measuring the same mechanism, then more thought is required in the determination of $\mathcal{M}$. In this instance, either one of the models should be removed to reduce the double counting or the prior distribution on the model space should take into account the correlation structure.

4.4 BMA for Spatial Models

In this section I describe the calculations necessary for BMA for spatial data. Building on the Bayesian spatial models in Section 3, the resulting formulas form the basis for spatial BMA.

4.4.1 The Posterior Predictive Distribution

The posterior predictive distribution for a new location Z_0 given the data $\boldsymbol{Z}$ averaged over all models is:

$$f(Z_0|\boldsymbol{Z}) = \sum_{k=1}^{K} f(Z_0|\boldsymbol{Z}, M_k) f(M_k|\boldsymbol{Z}). \quad (4.5)$$

where $f(Z_0|\boldsymbol{Z}, M_k)$ is the posterior predictive distribution for model k evaluated at Z_0 and $f(M_k|\boldsymbol{Z})$ is the PMP.

The posterior predictive distribution for model M_k is derived from the following integration:

$$f(Z_0|\boldsymbol{Z}, M_k) = \iiint f(Z_0|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{Z}, M_k) f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|\boldsymbol{Z}, M_k)\, d\boldsymbol{\beta}\, d\sigma^2 d\boldsymbol{\theta} \quad (4.6)$$

where $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ are the parameters of the model M_k, $f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|\boldsymbol{Z}, M_k)$ is the prior posterior density of $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ under model M_k, and $f(Z_0|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{Z}, M_k)$ is the predictive distribution of Z_0 given the parameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ under model M_k. Methods for evaluating the posterior predictive distribution in (4.6) are described in Section 3.2.3.

The PMP follows from Bayes' rule:

$$f(M_k|\boldsymbol{Z}) = \frac{f(\boldsymbol{Z}|M_k) f(M_k)}{\sum_{l=1}^{K} f(\boldsymbol{Z}|M_l) f(M_l)}. \quad (4.7)$$

where the integrated likelihood is

$$f(\boldsymbol{Z}|M_k) = \int\int\int f(\boldsymbol{Z}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, M_k) f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|M_k)\, d\boldsymbol{\beta}\, d\sigma^2 d\boldsymbol{\theta}. \quad (4.8)$$

Here $f(\boldsymbol{Z}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, M_k)$ is the likelihood, $f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|M_k)$ is the prior distribution of the parameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ under model M_k, $f(\boldsymbol{Z}|M_k)$ is the marginal likelihood of model M_k, and $f(M_k)$ is the prior probability that M_k is the true model. Note that all probabilities are conditional on the set of models being considered.

The computation of the PMPs plays a key role in calculating the posterior predictive distribution. The PMPs are found by first integrating out the parameters, and then using Bayes' rule to find the probability of a model given the data. For simplicity, the following derivations are assumed to be for a model M_k and notation specifying the model M_k has been dropped.

4.4.2 Posterior Model Probability

To estimate the PMP I first compute the integrated likelihood (4.8). Under a given model, M_k, the marginal likelihood is given by

$$f(\boldsymbol{Z}) = \int\int\int f(\boldsymbol{Z}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})\, f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})\, d\boldsymbol{\beta}\, d\sigma^2 d\boldsymbol{\theta},$$

where, for notational simplicity, $f(\boldsymbol{Z}) = f(\boldsymbol{Z}|M_k)$. This computation will be done in several steps, evaluating one integral at a time.

First, the calculation of

$$f(\boldsymbol{Z}|\sigma^2, \boldsymbol{\theta}) = \int f(\boldsymbol{Z}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})\, f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\theta})\, d\boldsymbol{\beta}$$

is considered. Write $\boldsymbol{Z} = \boldsymbol{X\beta} + \epsilon$, where ϵ and $\boldsymbol{\beta}$ are independent with $\epsilon \sim N(0, \sigma^2\boldsymbol{\Psi})$, $\boldsymbol{\beta}|\sigma^2, \boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{V})$, and $\boldsymbol{\Psi}, \boldsymbol{\mu}$, and $\boldsymbol{V}$ are as defined in Section 3.1.4. Then $E[\boldsymbol{Z}] = \boldsymbol{X\mu}$ and $Var[\boldsymbol{Z}] = \sigma^2(\boldsymbol{XVX'} + \boldsymbol{\Psi})$. Note that since $\boldsymbol{\beta}$ and ϵ are both normally distributed, $\boldsymbol{Z}|\sigma^2, \boldsymbol{\theta}\ (\boldsymbol{X\mu}, \sigma^2(\boldsymbol{XVX'} + \boldsymbol{\Psi}))$.

Next, integrating out σ^2 gives

$$f(\boldsymbol{Z}|\boldsymbol{\theta}) = \int f(\boldsymbol{Z}|\sigma^2, \boldsymbol{\theta})\, f(\sigma^2|\boldsymbol{\theta})\, d\sigma^2 \tag{4.9}$$

$$= \int (2\pi)^{-n/2} u^{n/2} |\boldsymbol{XVX'} + \boldsymbol{\Psi}|^{-1/2} \tag{4.10}$$
$$\exp\left\{-\frac{u}{2}\left\{(\boldsymbol{Z} - \boldsymbol{X\mu})'(\boldsymbol{XVX'} + \boldsymbol{\Psi})^{-1}(\boldsymbol{Z} - \boldsymbol{X\mu})\right\}\right\}$$
$$\frac{\left(\frac{\nu\lambda}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}(u)^{\frac{\nu}{2}-1}\exp\left\{-\frac{\nu u}{2}\right\} du.$$

Therefore,

$$
\begin{aligned}
f(\boldsymbol{Z}|\boldsymbol{\theta}) &= \left(\frac{1}{2\lambda}\left\{(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\mu})'(\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}'+\boldsymbol{\Psi})^{-1}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\mu})+\nu\lambda\right\}\right)^{-\frac{n+\nu+1}{2}} \\
&\quad (2\pi)^{-n/2}(\lambda)^{\nu/2}|\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}'+\boldsymbol{\Psi}|^{-1/2}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\int \exp\{-s\}s^{\left(\frac{n+\nu}{2}-1\right)}ds \\
&= (\pi)^{-n/2}(\nu\lambda)^{\nu/2}|\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}'+\boldsymbol{\Psi}|^{-1/2}\Gamma\left(\frac{\nu}{2}\right)^{-1}\Gamma\left(\frac{n+\nu}{2}\right) \qquad (4.11) \\
&\quad \left(\left\{(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\mu})'(\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}'+\boldsymbol{\Psi})^{-1}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\mu})+\nu\lambda\right\}\right)^{-(n+\nu)/2}.
\end{aligned}
$$

This is a multivariate non-central Student's t distribution with mean $\boldsymbol{X}\boldsymbol{\mu}$, variance $\frac{\nu\lambda}{\nu-2}(\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}'+\boldsymbol{\Psi})$, and ν degrees of freedom.

4.4.3 Laplace and BIC Approximations

The final step in the calculation of the PMP is to compute the marginal distribution for $\boldsymbol{Z}$ by performing the last integration in equation (4.8)

$$f(\boldsymbol{Z}) = \int f(\boldsymbol{Z}|\boldsymbol{\theta})f(\boldsymbol{\theta})\,d\boldsymbol{\theta}. \qquad (4.12)$$

The integration is difficult to compute in closed form. Instead, two approximations, (4.12), a Laplace approximation and the "BIC approximation" of the Laplace approximation, are used. The Laplace approximation is defined by

$$\int \exp\{L(\boldsymbol{\theta})\}\,d\boldsymbol{\theta} \approx (2\pi)^{b/2}|\boldsymbol{\Sigma}|^{1/2}\exp\left\{L\left(\hat{\boldsymbol{\theta}}\right)\right\} \qquad (4.13)$$

where $L(\boldsymbol{\theta}) = \log f(\boldsymbol{Z}|\boldsymbol{\theta}, M_k) + \log f(\boldsymbol{\theta}|M_k)$ is the product of (4.12) and the prior distribution for $\boldsymbol{\theta}$ given the model, b is the dimension of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$, and $\boldsymbol{\Sigma}$ is minus the inverse Hessian of $L(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}$. This approach was used for generalized linear models by Raftery (1996), and for censored survival models by Volinsky (1997).

The Bayesian information criteria (BIC) approximation (Raftery, 1996) is an approximation to the Laplace approximation, incorporating $\log|\boldsymbol{\Sigma}| \approx b\log n$. Both

the BIC approximation and the Laplace approximation improve as the distribution of the vector $\boldsymbol{\theta}$ approaches normality, and the approximations are exact when $\boldsymbol{\theta}$ is normally distributed.

For our application, the approximations are

$$\begin{aligned} f(\boldsymbol{Z}) &\approx \left|\hat{\boldsymbol{\Sigma}}\right|^{1/2} f\left(\boldsymbol{Z}|\hat{\boldsymbol{\theta}}\right) f\left(\hat{\boldsymbol{\theta}}\right). \quad \text{Laplace approximation} \\ f(\boldsymbol{Z}) &\approx \left(\exp^{b\log n}\right)^{1/2} f\left(\boldsymbol{Z}|\hat{\boldsymbol{\theta}}\right) f\left(\hat{\boldsymbol{\theta}}\right). \quad \text{BIC approximation} \end{aligned} \tag{4.14}$$

Note that the BIC approximation is not the same as the BIC model selection (Schwartz, 1978) criteria which is

$$\text{BIC} = -2\log f(\boldsymbol{Z}|\hat{\alpha}) + K\log n \tag{4.15}$$

where K is the dimension of the estimated parameters, and $\hat{\alpha}$ are the maximum likelihood estimates for the parameters.

In the simulations in Chapter 5 and in the examples in Chapter 6 of this thesis, I use the BIC approach to approximate (4.2). This simplifies the calculations by not requiring estimation of $\boldsymbol{\Sigma}$, the inverse Hessian described above.

4.5 Bayesian Model Selection

Bayesian model selection (BMS) is a procedure that uses the posterior model probabilities derived in the last section to select a best model. This method does not incorporate uncertainty in model selection, instead it serves as a technique for selecting a set of explanatory variables for Bayesian kriging (Section 3). After the set of explanatory variables has been selected, the model is fit following the procedures in Chapter 3 using the Matern autocorrelation function.

This model selection technique is similar to the methodology of Bayes Factors (Kass and Raftery, 1995; Raftery, 1995). A Bayes factor is defined to be

$$B_{10} = \frac{pr(\boldsymbol{Z}|M_1)}{pr(\boldsymbol{Z}|M_0)}. \tag{4.16}$$

Table 4.1: Bayes Factor Decision Table. The null hypothesis is that model 0 is the true model. The alternative hypothesis is that model 1 fits better.

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 12	Positive
12 to 150	Strong
> 150	Decisive

where $pr(\boldsymbol{Z}|M_k)$ is the integrated likelihood for model M_k, equation (4.3). Bayes factors provide a method for model comparison. Kass and Raftery provide a framework for evaluation of the evidence for model 1 fitting the data better than model 0 (Table 4.5). These criteria help to quantify decisions made about explanatory variables. Bayes factors have been used for model selection in a large variety of statistical models, including linear models, stochastic processes, survival analysis and multivariate analysis.

For BMS, the model with the largest PMP is selected for further examination. This "best" model for BMS is similar to the methodology for spatial AICC in that all of the candidate models in the model space $\mathcal{M}$ must be fit. BMS forms a middle ground between spatial AICC (Section 2.4.2) and BMA that incorporates parameter uncertainty into the selection of the model and parameter estimation, but does not take model uncertainty into account. In Section 5.2 I compare the model selection capabilities of BMS, spatial AICC, and AICC with independent errors.

Chapter 5

SIMULATION STUDIES

This chapter focuses on the results from four simulation studies investigating the estimation of spatial parameters and spatial prediction methods described in Chapters 2, 3, and 4. First, the model selection techniques of Sections 2.4, 2.4.2 and 4.5 are examined. Second, spatial BMA as discussed in Chapter 4 is compared to traditional spatial modeling techniques discussed in Chapter 2 and Section 1.2. Third, the necessity of including a nugget in spatial BMA compared to not including a nugget in modeling the spatial data with measurement error is evaluated. Fourth, REML and ML estimation methods for spatial models which was discussed in Section 2.3.3, are compared.

5.1 Simulation Set-Up

In this section I describe the basic set-up of all four simulation studies. Each simulated data set consists of two parts, sampling locations and explanatory variables. The sampling locations follow a variety of different spatial patterns described in Section 5.1.1. The explanatory variables and the models used for simulation are described in Section 5.1.2. In Section 5.1.3, the statistics used for assessing the quality of prediction are formulated.

5.1.1 Sampling Patterns

Spatial design is a growing area of research (Cressie, 1993, Section 5.6). When sampling a spatial process, the locations of the observations can impact parameter

estimation. In general, the sampling pattern should give good coverage over the entire area of interest, plus some observations at small distances so that it is possible to estimate both the large and small scale variability (Stein, 1999, p.187). Instead of investigating optimal sampling designs, I consider realistic sampling designs and investigate estimation and prediction for different sampling patterns.

Decisions about the sampling locations correspond to trade-offs between the two scales of variability. On one extreme, the set of sample locations should have good coverage in terms of observations throughout the entire sample space in order to estimate the large scale variability. However, this results in larger distances between locations, and thus less information is available to model small scale variation. At the other extreme, the set of locations should have observations at small distances, which provides information about the small scale variability of the fitted model. The small scale variability of the model is directly related to smoothness.

In this thesis I use five different sampling patterns for all simulations. The sampling patterns used in the simulations are classified into four groups: grid patterns, regular patterns, random designs, and clustered designs (Figure 5.1). My goal was to select sampling patterns that reflect a wide range of possible sampling patterns. These groups are described in the following sections.

Cluster Process

The first two spatial designs in Figure 5.1 are clustered or aggregated. In my simulations I use these two cluster patterns, a lightly clustered pattern and a highly clustered pattern. This process allows sample points to be close together and therefore facilitates modeling of the small scale variation. This pattern may be lacking information to model the large scale variation because the locations are grouped together and therefore there may be areas of D that do not have any sample points.

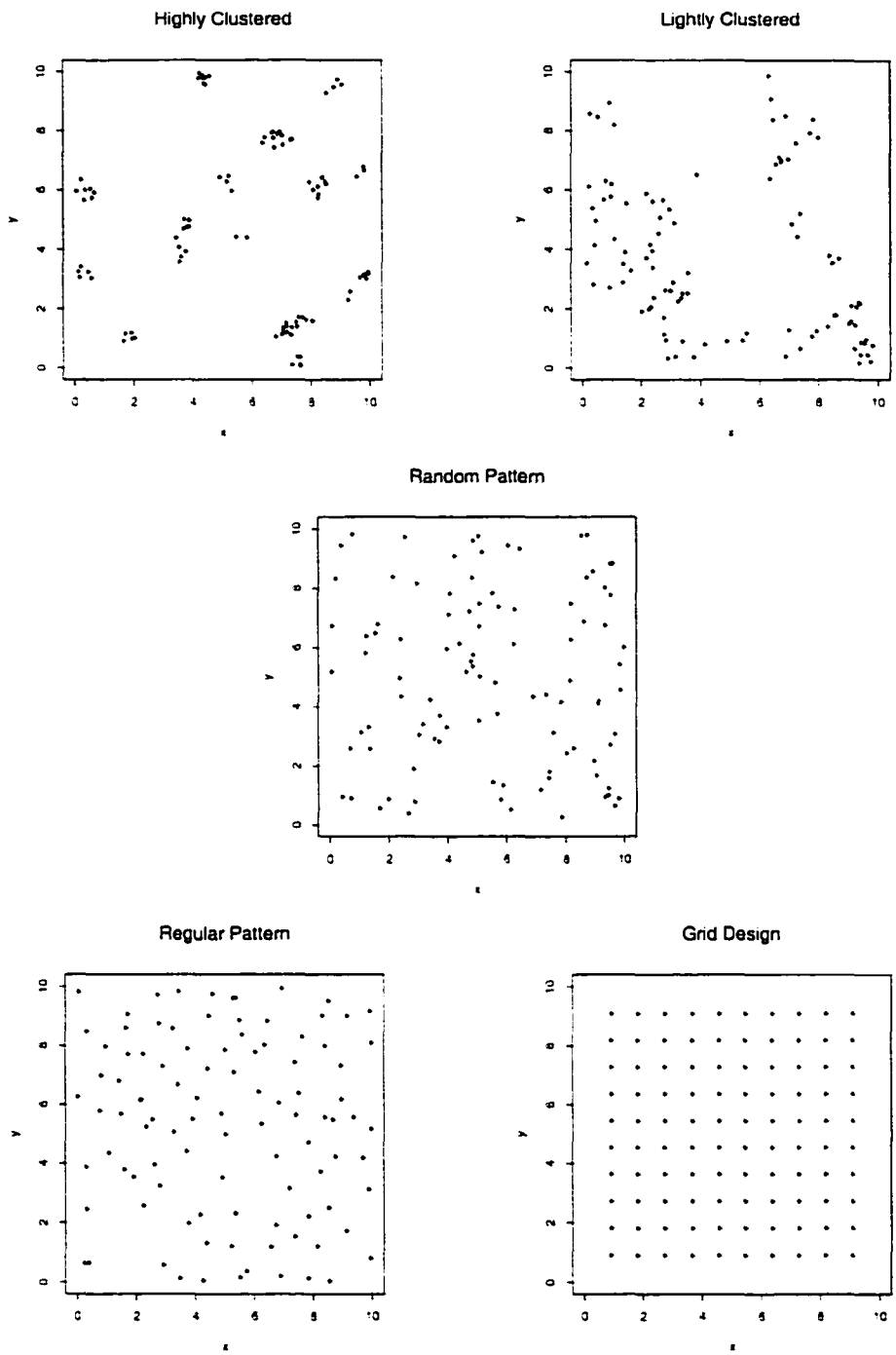

Figure 5.1: Spatial Point Patterns for the Estimation Set.

The aggregation process is generated in two stages (Ripley, 1981, Section 8.4). First, the number of clusters $m, m < n$ is simulated from a Poisson process with mean c. The sample parent locations $x_1, \ldots, x_m$ are uniformly distributed over the entire area D. The remaining sample points $x_{m+1}, \ldots, x_n$ are randomly assigned to a parent location x_i and are located uniformly on a disk with radius r about the x_i. In my simulations I use two different cluster patterns, a lightly clustered pattern with $c = 20$ and $r = 0.8$ and a highly clustered pattern with $c = 20$ and $r = 0.4$.

Poisson Process

The third spatial design in Figure 5.1 is a Poisson process. Conditioning on n, the total number of observations, this process is a random point pattern. This pattern is less structured than the regular process, but does not encourage clustering.

The Poisson process is generated by conditioning on n. The locations $x_1, \ldots, x_n$ are uniformly distributed over the entire area, D (Cressie, 1993, p.620). The Poisson process forms a process which neither limits the number of locations at small distances as the Strauss process (described below), nor encourages clustering of locations as the previous spatial pattern, the cluster process.

Regular Process

The fourth spatial design in Figure 5.1 is an inhibition process (Strauss process) (Cressie, 1993, Section 8.5.5) which not only allows for good coverage of the sampling space but also allows for two locations to be close together with a small probability. This pattern includes more information about the smoothness compared to the grid pattern because some observations are close together, although less information than the next two spatial patterns.

The Strauss process is simulated sequentially. The first location x_1 is generated uniformly in D. Next, generate a location U uniformly on D. Accept $x_2 = U$ with

probability 1 if $|U - x_1| > \delta$ and with probability c if $|U - x_1| \leq \delta$. The simulation continues with a random location U on D, accept $x_n = U$ with probability 1 if $|U - x_j| > \delta$ for $j = 1, \ldots, n-1$ and with probability c^r, where r is the number of locations x_i that are within δ of U. Therefore, as the number of points within a disk of size δ increases, the probability of accepting the new location within the disk decreases. In my simulations I use a regular process with $c = 0.1$ and $r = 0.7$.

Grid Patterns

The final spatial design is a grid (Figure 5.1). If I define D to be the sampling space, the n sampling locations are equally spaced throughout D. This pattern results in complete coverage and tends to model the large scale variability well, but can provide poor estimates of the small scale variability. For example, if I use a grid with locations at the integers, this pattern has pairs of locations at distances of $1, \sqrt{2}, 2, \sqrt{5}, \ldots$. Therefore, there is very little information about the smoothness of the process, or the autocorrelation function at distances less than one unit.

From these four spatial point patterns, six sets of 100 locations are created in a 10 by 10 grid, one set is held out for prediction (random location). The remaining five sets used for estimation include a) grid design, b) regular pattern, c) random pattern, d) lightly clustered, and e) highly clustered and can be seen in Figure 5.1.

The simulations in Sections 5.2, 5.3, and 5.4 fix the locations. For the investigation of the behavior of REML estimation compared to ML estimation in Section 5.5, every run of the simulation includes a new set of locations. For more information, see the corresponding sections.

5.1.2 Explanatory Variables and Models

In these simulations, the data is divided into two sets, an estimation set and a prediction set. This allows for assessment of model selection techniques, estimation

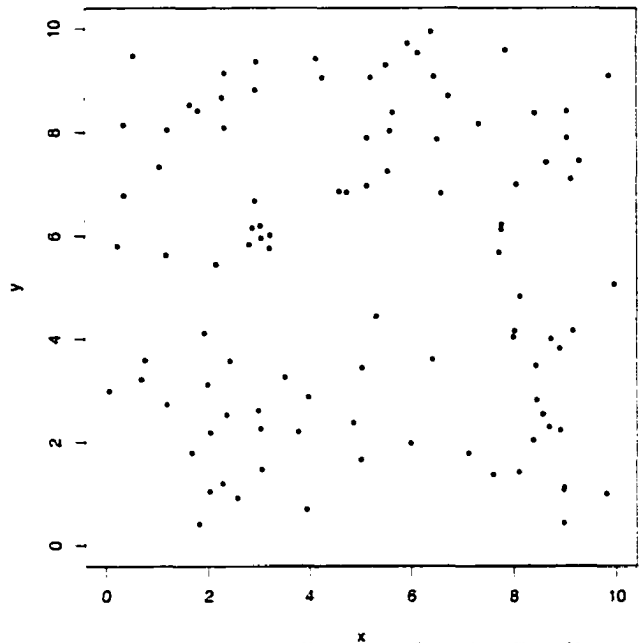

Figure 5.2: Spatial Point Pattern for the Prediction Set.

of parameters using the estimation set, and prediction of the unobserved locations in the prediction set. Each set includes locations and explanatory variables. The locations for the estimation set are described in Section 5.1.1. The locations for the prediction set are simulated from a random process (Figure 5.2).

The explanatory variables are identical and fixed for all simulations. I create five explanatory variables for 200 locations, the first 100 used for estimation, the second 100 used for prediction. Each explanatory variable is generated from a standardized Student's t distribution with 12 degrees of freedom. $\boldsymbol{X}_i \sim \sqrt{\frac{12}{10}} t_{df=12}$ for $i = 1, \ldots, 5$. In the simulations I wanted the explanatory variables to have slightly heavier tails compared to a normal distribution.

The estimation and prediction sets are simulated jointly from the model

$$\boldsymbol{Z} \sim \text{Normal}\left(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Psi}\right) \tag{5.1}$$

Table 5.1: Models for Simulations

Model	Explanatory Variables	Model	Explanatory Variables
M_1	none	M_{17}	X_1, X_2, X_3
M_2	X_1	M_{18}	X_1, X_2, X_4
M_3	X_2	M_{19}	X_1, X_2, X_5
M_4	X_3	M_{20}	X_1, X_3, X_4
M_5	X_4	M_{21}	X_1, X_3, X_5
M_6	X_5	M_{22}	X_1, X_4, X_5
M_7	X_1, X_2	M_{23}	X_2, X_3, X_4
M_8	X_1, X_3	M_{24}	X_2, X_3, X_5
M_9	X_1, X_4	M_{25}	X_2, X_4, X_5
M_{10}	X_1, X_5	M_{26}	X_3, X_4, X_5
M_{11}	X_2, X_3	M_{27}	X_1, X_2, X_3, X_4
M_{12}	X_2, X_4	M_{28}	X_1, X_2, X_3, X_5
M_{13}	X_2, X_5	M_{29}	X_1, X_2, X_4, X_5
M_{14}	X_3, X_4	M_{30}	X_1, X_3, X_4, X_5
M_{15}	X_3, X_5	M_{31}	X_2, X_3, X_4, X_5
M_{16}	X_4, X_5	M_{32}	X_1, X_2, X_3, X_4, X_5

where $\mu = 2 + .75X_1 + .5X_2 + .25X_3$, $\sigma^2 = 50$, and Ψ is the Matern covariance function with parameters $\theta = (4.0, 1.0)$. After simulating the data, the explanatory variables, except for the constant term, are standardized within the prediction and estimation set to simplify prior selection, resulting in a mean of zero, and a variance of one.

With five explanatory variables, there are $2^5 = 32$ possible models. All possible models are listed in Table 5.1. I constructed this model for simulation so that there is some model uncertainty. This is accomplished by progressively reducing the size of the coefficients. For example, the difference between model M_{17} to model M_7 is explanatory variable X_3 which has a coefficient of 0.25. This difference may be slight if the variability of the process is large.

The value of σ^2 varies by simulation. Some simulations include a nugget term and therefore require less variability in order to be able to observe the signal from the noise of the process. It is also important that the signal from the data is

not too strong, which would result in the correct model being chosen for every replication. This was designed to mimic a realistic situation with model uncertainty. The simulations which do not include the nugget term need a larger variability to provide model uncertainty. In the cases with a nugget term, I used $\sigma^2 = 10$. Cases without a nugget term, I used $\sigma^2 = 50$.

5.1.3 Model Assessment

The results from different modeling techniques will be compared and contrasted within each simulation. In Section 5.5, the modeling methodologies are compared by examining the bias and standard errors of the simulations. In Section 5.2 the concern is model selection, so AICC independent models are compared to the spatial AICC selection procedure as well as Bayesian model selection (BMS). Sections 5.3 and 5.4 evaluate the prediction intervals for unobserved locations for a variety of techniques. This section describes three statistics for comparing prediction results.

The first statistic for comparison is mean square prediction error (MSPE). Let Z_j identify the simulated value of the spatial process at location j in the prediction set, where $j = 1, \ldots, 100$ (equation 5.1). For each modeling technique, let $\hat{Z}_j$ denote the predicted value at the j^{th} prediction location. Then, the average MSPE is

$$\text{MSPE} = \frac{1}{100}\sum_{j=1}^{100}\left(Z_j - \hat{Z}_j\right)^2. \tag{5.2}$$

Relative MSPE (rMSPE) is MSPE divided by the corresponding mean square prediction error based on the true model. That is,

$$\text{rMSPE} = \frac{\sum_{j=1}^{100}\left(Z_j - \hat{Z}_j\right)^2}{\sum_{j=1}^{100}\left(Z_j - \tilde{Z}_j\right)^2}. \tag{5.3}$$

where $\tilde{Z}_j$ is the universal kriging predictor for the j^{th} prediction location (1.4) using the true values of $\boldsymbol{\beta}$ and the true autocorrelation matrix $\boldsymbol{\Psi}(\boldsymbol{\theta})$ defined in (5.1). For both measures of MSPE, smaller values indicate more accurate prediction.

The second statistic is predictive coverage (PC). Let $\left(\hat{Z}_j^L, \hat{Z}_j^U\right)$ represent the endpoints of the estimated 95% prediction interval based on each modeling technique. Then, for one replicate meanPC is

$$\text{meanPC} = \frac{1}{100}\sum_{j=1}^{100} Pr\left(Z_j \in \left(\hat{Z}_j^L, \hat{Z}_j^U\right) | \text{true model}\right) \tag{5.4}$$

where $Pr(\cdot|\text{the true model})$ is the probability the true prediction value is in the estimated prediction interval. When the data are simulated, the true model is known. For example in this thesis the data are simulated from (5.1), the conditional distribution of a prediction location given the estimation set is also known and is used to calculate the predictive coverage for all members of the prediction set.

In examining this statistic for a simulation, there are two goals. First, meanPC for all replicates should be centered at 95% indicating that on average the prediction intervals are the correct size. Second, the variability of meanPC over all replicates should be small. For each set of 100 prediction locations, I also consider the minimum, maximum, and standard deviation of predictive coverage, where

$$\text{minPC} = \min_j Pr\left(Z_j \in \left(\hat{Z}_j^L, \hat{Z}_j^U\right) | \text{true model}\right) \tag{5.5}$$

$$\text{maxPC} = \max_j Pr\left(Z_j \in \left(\hat{Z}_j^L, \hat{Z}_j^U\right) | \text{true model}\right) \tag{5.6}$$

$$\text{stdPC} = \left(\frac{1}{100}\sum_{j=1}^{100}\left(Pr\left(Z_j \in \left(\hat{Z}_j^L, \hat{Z}_j^U\right) | \text{true model}\right) - \text{meanPC}\right)^2\right)^{\frac{1}{2}} \tag{5.7}$$

These statistics can be calculated when the true model is known, which is true of simulation studies. When the true model is not known, empirical predictive coverage (EPC) can be used as an approximation. Let $I_{EPC}(\cdot)$ denote an indicator function for the prediction interval, where $I_{\text{EPC}}(Z_j) = 1$ if $Z_j \in \left(\hat{Z}_j^L, \hat{Z}_j^U\right)$ and zero otherwise. Then EPC is

$$\text{EPC} = \frac{1}{100}\sum_{j=1}^{100} I_{EPC}(Z_j). \tag{5.8}$$

Together, these statistics allow for the comparison of different techniques based on their predictive performance.

5.2 Model Selection

The goal of this simulation is to investigate the performance of the model selection techniques discussed in this thesis. Comparisons will be made between the traditional methods (Section 2.4.1), spatial AICC (Section 2.4.2) and Bayesian model selection (BMS) (Section 3). Traditional methods for spatial model selection first select the best set of explanatory variables assuming the observations are independent and then estimate the autocorrelation function parameters. This approach is called AICC-Independent below. In spatial AICC model selection the AICC statistic is based on the model with autocorrelated errors. The AICC statistic is computed for all possible subsets of explanatory variables. Bayesian model selection places prior distributions on the parameters and chooses the set of explanatory variables by selecting the model with the highest posterior model probability.

5.2.1 Details

As described in Section 5.1, the data are simulated from the model $\boldsymbol{Z} = 2 + 0.75\boldsymbol{X}_1 + 0.50\boldsymbol{X}_2 + 0.25\boldsymbol{X}_3 + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a Gaussian random field with mean zero, $\sigma^2 = 50$, and autocorrelation Matern with parameters $\theta_1 = 4$ and $\theta_2 = 1$. The sampling patterns are held constant for all simulations and are shown in Figure 5.1.

For the traditional method, the AICC statistic is calculated for each of the 32 models described in Section 5.1.2. For the spatial AICC model selection, the profile likelihood equation (2.27)

$$\ell_{profile}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{Z}\right) = -\frac{1}{2}\log|\Psi| - \frac{n}{2}\log\left(\hat{\sigma}^2\right) - \frac{n}{2} \tag{5.9}$$

is maximized for (θ_1, θ_2). Splus was used to carry out the calculations for all 32 models. The model that maximizes the profile likelihood is "best".

For the BMS method, the prior distribution is as specified in Section 3.1.4: β is Gaussian with mean $\mu = (\bar{z}, 0, \ldots, 0)'$ and variance

$$V = \begin{pmatrix} s_Z^2 & 0 & \ldots & 0 \\ 0 & \vartheta^2 & 0 & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \vartheta^2 \end{pmatrix}.$$

where s_Z^2 is the variance of the observed data, $\bar{z}$ is the mean of the observed data, and $\vartheta^2 = 2.85$. This choice of ϑ^2 follows from research in Hoeting (1994) investigating hyperparameter selection when the explanatory variables are standardized. The prior distribution on σ^2 is inverse gamma with parameters $\nu = 10$ and $\lambda = 50$. This corresponds to $E[\sigma^2] = 62.5$ and $\mathrm{Var}[\sigma^2] = 1302$. The prior distribution on θ_1 is uniform over the interval $(0, d_{max})$, where d_{max} is the maximum distance between locations (Figure 5.3). The prior on θ_2 is also uniform over the interval $(0, c)$ where $c = 10$.

The posterior distribution of (θ_1, θ_2), equation (4.11), is maximized. The resulting estimates of the spatial parameters are the posterior modes. The PMP are estimated using the BIC approximation. Finally, the PMP are calculated using equation (4.2). For the BMS, the model with the highest PMP is chosen

5.2.2 Results

Results for the Random Pattern

Focusing on the random pattern (Table 5.2), the drawbacks with the standard model selection approach are very clear. When independence is assumed, the AICC statistic selects the true model (M_{17}) for only 3 out of 1000 simulations (0.30%) while the model with no explanatory variables (M_1) is selected 370 times. The AICC independence approach selects the model with explanatory variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ 5.50% of the time. In total, the first explanatory variable is in 31.3% of the

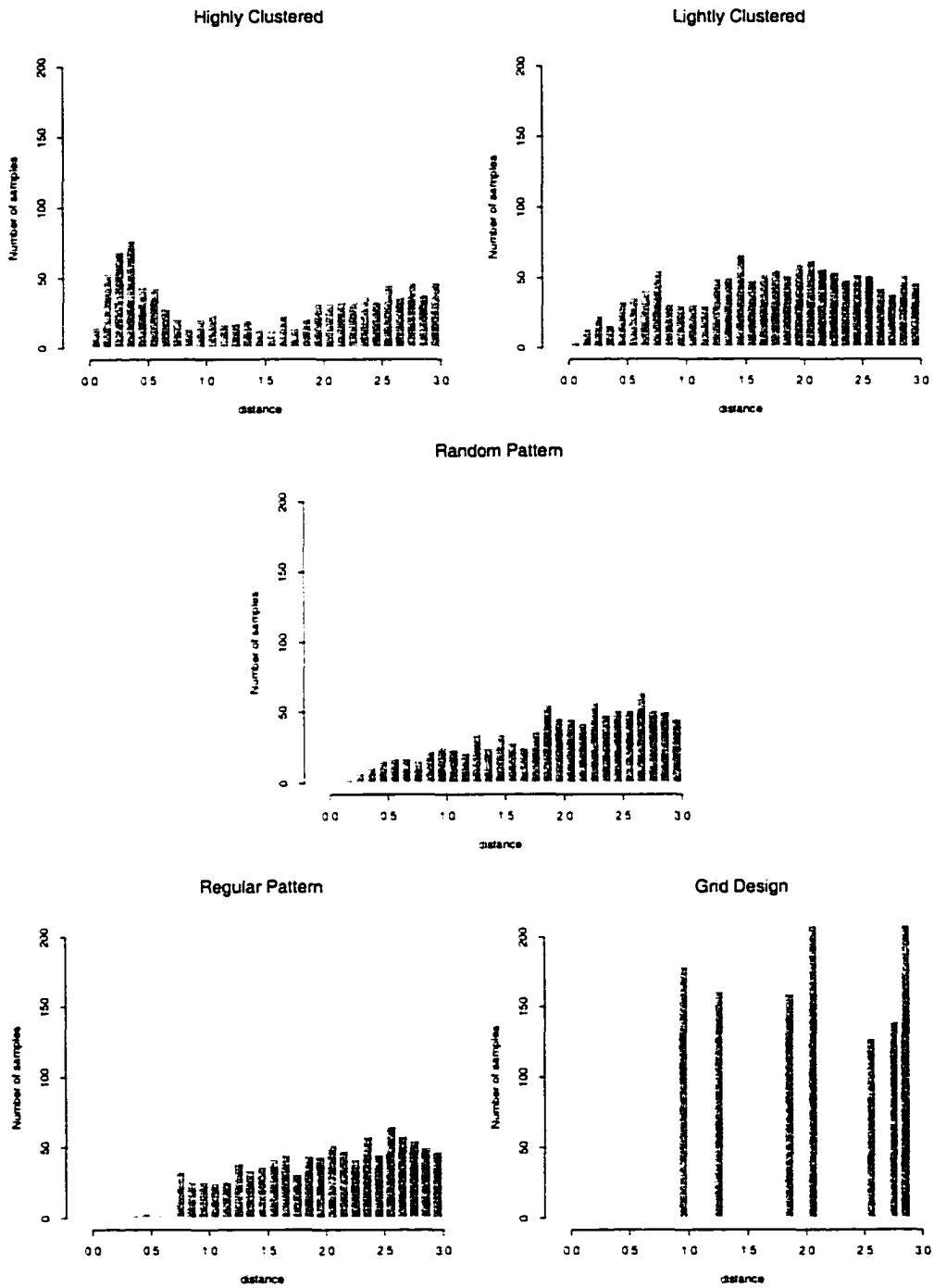

Figure 5.3: Distance Between Locations for the Spatial Point Patterns.

Table 5.2: Model Selection Results for the Random Pattern. AICC Independent, Spatial AICC, and BMS report the percentage of times each model was selected. The number of simulations that converged is denoted by n. 1000 replicates were simulated. The true model is M_{17}.

Model		Independent AICC ($n = 1000$)	Spatial AICC ($n = 999$)	BMS ($n = 1000$)
M_1		37.70	0.00	0.30
M_2	X_1	17.50	0.50	23.50
M_3	X_2	11.90	0.00	0.20
M_4	X_3	2.90	0.00	0.00
M_5	X_4	9.10	0.00	0.00
M_6	X_5	1.40	0.00	0.00
M_7	X_1, X_2	5.50	17.92	61.10
M_8	X_1, X_3	1.20	1.40	2.00
M_9	X_1, X_4	4.00	0.30	0.00
M_{10}	X_1, X_5	1.20	0.10	0.00
M_{11}	X_2, X_3	1.10	0.00	0.00
M_{12}	X_2, X_4	3.30	0.00	0.10
M_{13}	X_2, X_5	0.20	0.00	0.00
M_{14}	X_3, X_4	0.20	0.00	0.00
M_{15}	X_3, X_5	0.60	0.00	0.00
M_{16}	X_4, X_5	0.30	0.00	0.00
M_{17}	X_1, X_2, X_3	0.30	45.85	11.80
M_{18}	X_1, X_2, X_4	0.70	5.51	0.50
M_{19}	X_1, X_2, X_5	0.00	6.21	0.20
M_{20}	X_1, X_3, X_4	0.30	0.20	0.00
M_{21}	X_1, X_3, X_5	0.20	0.30	0.00
M_{22}	X_1, X_4, X_4	0.30	0.00	0.00
M_{23}	$X_2, X_3, X_4,$	0.00	0.00	0.00
M_{24}	X_2, X_3, X_5	0.00	0.00	0.00
M_{25}	X_2, X_4, X_5	0.00	0.00	0.00
M_{26}	X_3, X_4, X_5	0.00	0.00	0.00
M_{27}	$X_1, X_2, X_3, X_4,$	0.00	7.91	0.20
M_{28}	X_1, X_2, X_3, X_5	0.00	11.01	0.10
M_{29}	X_1, X_2, X_4, X_5	0.00	1.30	0.00
M_{30}	X_1, X_3, X_4, X_5	0.10	0.20	0.00
M_{31}	X_2, X_3, X_4, X_5	0.00	0.00	0.00
M_{32}	X_1, X_2, X_3, X_4, X_5	0.00	1.30	0.00

selected models, and the second explanatory variable is included in 23% of the models.

Spatial AICC performs much better than the independence method. The true model is selected 45.85% of the time, the highest of all four methods for this sampling pattern. This method tends to overestimate the number of parameters in the model, selecting models with one or two extra variables 21.52% percent of the time.

The BMS method does not overestimate the number of significant explanatory variables. Instead, the true model is selected 11.8% of the time, but 61.1% of the models included the first two explanatory variables. The first explanatory variable is included in 99.4% of the models, while the second explanatory variable is in 74.2% of the models, and the third explanatory variable is in 14.1% of the models. These values indicate that BMS tends to select the appropriate explanatory variables with the explanatory with the largest signal ($\boldsymbol{X}_1$) in a majority of models. As the coefficient decreases, the percent of models that select that explanatory variable also decreases.

In conclusion, the AICC independence method tends to select models with very few explanatory variables, and does not do well in selecting models that contain the true parameters. The spatial AICC method does very well in selecting the true model for this simulation, although it frequently selects models with extra predictors. The BMS method does well in selecting models with the correct explanatory variables.

Results for the other Sampling Patterns

The sampling pattern has a large effect on the choice of explanatory variables for model selection techniques that account for the autocorrelation. The independence technique does not show much dependence on the sampling pattern. The highly clustered pattern has the most dramatic results (Table 5.3). Both methods

Table 5.3: Model Selection results for the Highly Clustered Pattern. AICC Independent, Spatial AICC, and BMS report the percentage of times each model was selected. n is the number of simulations that converged. 500 replicates were simulated. The true model is M_{17}.

Model		Independent AICC ($n = 500$)	Spatial AICC ($n = 469$)	BMS ($n = 499$)
M_1		30.80	0.00	0.00
M_2	X_1	14.40	0.00	0.00
M_3	X_2	10.80	0.00	0.00
M_4	X_3	2.80	0.00	0.00
M_5	X_4	10.40	0.00	0.00
M_6	X_5	0.40	0.00	0.00
M_7	X_1, X_2	9.20	0.00	5.81
M_8	X_1, X_3	6.00	0.00	0.00
M_9	X_1, X_4	4.00	0.00	0.00
M_{10}	X_1, X_5	0.40	0.00	0.00
M_{11}	X_2, X_3	3.00	0.00	0.00
M_{12}	X_2, X_4	4.00	0.00	0.00
M_{13}	X_2, X_5	0.40	0.00	0.00
M_{14}	X_3, X_4	1.00	0.00	0.00
M_{15}	X_3, X_5	0.00	0.00	0.00
M_{16}	X_4, X_5	0.00	0.00	0.00
M_{17}	X_1, X_2, X_3	0.80	73.35	93.99
M_{18}	X_1, X_2, X_4	1.00	0.00	0.00
M_{19}	X_1, X_2, X_5	0.00	0.00	0.00
M_{20}	X_1, X_3, X_4	0.20	0.00	0.00
M_{21}	X_1, X_3, X_5	0.00	0.00	0.00
M_{22}	X_1, X_4, X_4	0.20	0.00	0.00
M_{23}	$X_2, X_3, X_4,$	0.20	0.00	0.00
M_{24}	X_2, X_3, X_5	0.00	0.00	0.00
M_{25}	X_2, X_4, X_5	0.00	0.00	0.00
M_{26}	X_3, X_4, X_5	0.00	0.00	0.00
M_{27}	$X_1, X_2, X_3, X_4,$	0.00	11.94	0.20
M_{28}	X_1, X_2, X_3, X_5	0.00	10.23	0.00
M_{29}	X_1, X_2, X_4, X_5	0.00	0.00	0.00
M_{30}	X_1, X_3, X_4, X_5	0.00	0.00	0.00
M_{31}	X_2, X_3, X_4, X_5	0.00	0.00	0.00
M_{32}	X_1, X_2, X_3, X_4, X_5	0.00	4.48	0.00

Table 5.4: Model Selection results for the Lightly Clustered Pattern. AICC Independent. Spatial AICC, and BMS report the percentage of times each model was selected. n is the number of simulations that converged. 500 replicates were simulated. The true model is M_{17}.

Model		Independent AICC ($n = 500$)	Spatial AICC ($n = 497$)	BMS ($n = 500$)
M_1		42.00	0.00	0.00
M_2	X_1	32.00	0.00	0.20
M_3	X_2	15.80	0.00	0.00
M_4	X_3	1.20	0.00	0.00
M_5	X_4	0.60	0.00	0.00
M_6	X_5	1.40	0.00	0.00
M_7	X_1, X_2	1.00	2.21	55.40
M_8	X_1, X_3	3.60	0.00	2.20
M_9	X_1, X_4	0.20	0.00	0.00
M_{10}	X_1, X_5	0.60	0.00	0.00
M_{11}	X_2, X_3	0.00	0.00	0.00
M_{12}	X_2, X_4	0.40	0.00	0.00
M_{13}	X_2, X_5	0.80	0.00	0.00
M_{14}	X_3, X_4	0.00	0.00	0.00
M_{15}	X_3, X_5	0.00	0.00	0.00
M_{16}	X_4, X_5	0.00	0.00	0.00
M_{17}	X_1, X_2, X_3	0.00	65.39	42.20
M_{18}	X_1, X_2, X_4	0.20	1.81	0.00
M_{19}	X_1, X_2, X_5	0.00	0.20	0.00
M_{20}	X_1, X_3, X_4	0.00	0.00	0.00
M_{21}	X_1, X_3, X_5	0.20	0.00	0.00
M_{22}	X_1, X_4, X_4	0.00	0.00	0.00
M_{23}	$X_2, X_3, X_4,$	0.00	0.00	0.00
M_{24}	X_2, X_3, X_5	0.00	0.00	0.00
M_{25}	X_2, X_4, X_5	0.00	0.00	0.00
M_{26}	X_3, X_4, X_5	0.00	0.00	0.00
M_{27}	$X_1, X_2, X_3, X_4,$	0.00	13.48	0.00
M_{28}	X_1, X_2, X_3, X_5	0.00	13.08	0.00
M_{29}	X_1, X_2, X_4, X_5	0.00	0.20	0.00
M_{30}	X_1, X_3, X_4, X_5	0.00	0.00	0.00
M_{31}	X_2, X_3, X_4, X_5	0.00	0.00	0.00
M_{32}	X_1, X_2, X_3, X_4, X_5	0.00	3.62	0.00

Table 5.5: Model Selection Results for the Regular Pattern. AICC Independent, Spatial AICC, and BMS report the percentage of times each model was selected. n is the number of simulations that converged. 500 replicates were simulated. The true model is M_{17}.

Model		Independent AICC ($n = 500$)	Spatial AICC ($n = 499$)	BMS ($n = 500$)
M_1		43.20	0.00	1.80
M_2	X_1	23.20	0.80	37.00
M_3	X_2	12.00	0.00	0.00
M_4	X_3	3.20	0.00	0.00
M_5	X_4	2.00	0.00	0.00
M_6	X_5	4.40	0.00	0.00
M_7	X_1, X_2	4.60	21.44	48.60
M_8	X_1, X_3	1.80	1.20	1.60
M_9	X_1, X_4	1.00	1.00	0.20
M_{10}	X_1, X_5	1.80	0.60	0.00
M_{11}	X_2, X_3	0.40	0.00	0.00
M_{12}	X_2, X_4	0.40	0.00	0.00
M_{13}	X_2, X_5	0.80	0.00	0.00
M_{14}	X_3, X_4	0.00	0.00	0.00
M_{15}	X_3, X_5	0.20	0.00	0.00
M_{16}	X_4, X_5	0.00	0.00	0.00
M_{17}	X_1, X_2, X_3	0.20	43.49	9.00
M_{18}	X_1, X_2, X_4	0.40	5.21	1.00
M_{19}	X_1, X_2, X_5	0.40	6.81	0.60
M_{20}	X_1, X_3, X_4	0.00	0.80	0.00
M_{21}	X_1, X_3, X_5	0.00	0.40	0.00
M_{22}	X_1, X_4, X_4	0.00	0.00	0.00
M_{23}	$X_2, X_3, X_4,$	0.00	0.00	0.00
M_{24}	X_2, X_3, X_5	0.00	0.00	0.00
M_{25}	X_2, X_4, X_5	0.00	0.00	0.00
M_{26}	X_3, X_4, X_5	0.00	0.00	0.00
M_{27}	$X_1, X_2, X_3, X_4,$	0.00	8.02	0.20
M_{28}	X_1, X_2, X_3, X_5	0.00	7.21	0.00
M_{29}	X_1, X_2, X_4, X_5	0.00	1.20	0.00
M_{30}	X_1, X_3, X_4, X_5	0.00	0.20	0.00
M_{31}	X_2, X_3, X_4, X_5	0.00	0.00	0.00
M_{32}	X_1, X_2, X_3, X_4, X_5	0.00	1.60	0.00

Table 5.6: Model Selection Results for the Grid Design. AICC Independent, Spatial AICC, and BMS report the percentage of times each model was selected. n is the number of simulations that converged. 500 replicates were simulated. The true model is M_{17}.

Model		Independent AICC ($n = 500$)	Spatial AICC ($n = 500$)	BMS ($n = 500$)
M_1		26.60	0.00	8.60
M_2	X_1	21.20	7.20	48.60
M_3	X_2	12.60	1.60	15.80
M_4	X_3	5.80	0.00	0.00
M_5	X_4	4.60	0.00	0.00
M_6	X_5	2.00	0.00	0.00
M_7	X_1, X_2	3.80	34.80	19.80
M_8	X_1, X_3	6.40	8.20	4.60
M_9	X_1, X_4	3.60	1.40	0.20
M_{10}	X_1, X_5	1.20	1.40	0.80
M_{11}	X_2, X_3	4.40	0.00	0.00
M_{12}	X_2, X_4	2.60	0.40	0.00
M_{13}	X_2, X_5	0.60	0.00	0.00
M_{14}	X_3, X_4	0.00	0.00	0.00
M_{15}	X_3, X_5	0.00	0.00	0.00
M_{16}	X_4, X_5	0.40	0.00	0.00
M_{17}	X_1, X_2, X_3	0.80	15.80	1.00
M_{18}	X_1, X_2, X_4	0.40	5.60	0.20
M_{19}	X_1, X_2, X_5	0.20	8.20	0.40
M_{20}	X_1, X_3, X_4	0.60	1.20	0.00
M_{21}	X_1, X_3, X_5	1.00	2.20	0.00
M_{22}	X_1, X_4, X_4	0.40	0.00	0.00
M_{23}	$X_2, X_3, X_4,$	0.40	0.00	0.00
M_{24}	X_2, X_3, X_5	0.40	0.20	0.00
M_{25}	X_2, X_4, X_5	0.00	0.00	0.00
M_{26}	X_3, X_4, X_5	0.00	0.00	0.00
M_{27}	$X_1, X_2, X_3, X_4,$	0.00	3.20	0.00
M_{28}	X_1, X_2, X_3, X_5	0.00	6.60	0.00
M_{29}	X_1, X_2, X_4, X_5	0.00	1.20	0.00
M_{30}	X_1, X_3, X_4, X_5	0.00	0.40	0.00
M_{31}	X_2, X_3, X_4, X_5	0.00	0.00	0.00
M_{32}	X_1, X_2, X_3, X_4, X_5	0.00	0.40	0.00

that incorporate the spatial autocorrelation (spatial AICC and BMS) select the true model over 70% of the time. BMS the true model 94%, the spatial AICC method selects the true model 73.35%. Again, the spatial AICC method tends to overestimate the number of explanatory variables compared to BMS.

The highly clustered pattern has a large number of observations at small distances (Figure 5.3). These observations at small distances when the simulation does not include nugget ($\theta_3 = 0$) means that there is a lot of information about the autocorrelation function near the origin, improving the ability to model the small scale variability. Removing the estimated error process from the observations results in $\boldsymbol{X\beta}$, from which simulation results show it is straightforward to identify the set of explanatory variables.

The lightly clustered pattern (Table 5.4) forms a middle step between the highly clustered pattern and the random pattern. The percentages for the true model fall between the former two patterns. Continuing the progression, the models selected for the regular pattern (Table 5.5) are very similar to those chosen for the random pattern for all three methods. The regular pattern has more weight on the model with only $\boldsymbol{X}_1$ compared to the random pattern for spatial AICC. Correspondingly, the model with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ has less weight compared to the random pattern for both spatial AICC and BMS.

Finally, the grid pattern (Table 5.6) has the least amount of information about the autocorrelation function at small distances, and therefore has the most difficulty selecting the true model. The spatial AICC selects the true model almost 16% of the time, and almost 35% selecting the model with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The BMS does not perform as well, with only 1% for the true model, and almost 20% for the model with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

Conclusion

First, independent AICC approach does not select the correct model. Second, the ability to select the correct model depends on the sampling pattern for both of the spatial methods. The spatial AICC does well in selecting the 'true' model, and BMS performs well and mimics the model uncertainty but it is not as successful at selecting the true model as spatial AICC.

Dependence of model selection on the sampling pattern is very strong. The highly clustered pattern has the largest number of observations with distances that are close together. Therefore, when spatial autocorrelation is taken into account, it is possible to select the correct explanatory variables. On the other hand, the grid design has the no observations at small distances. Because of the difficulties in modeling the autocorrelation function, it is also challenging to pick out the 'true' model.

5.3 Simulation without Nugget

The primary goal of this simulation is to explore the behavior of BMA (Chapter 4) using the Matern class of covariance functions when the data are simulated with no nugget term ($\theta_3 = 0$). All spatial models in this simulation also have $\theta_3 = 0$. I compare predictions and estimates for seven different approaches: spatial BMA, spatial AICC, REML estimation of spatial models with exponential, spherical and Gaussian autocorrelation functions with the explanatory variables selected by independent AICC, Standard linear regression estimation, and BMA assuming independent errors.

5.3.1 Details

This simulation builds on the results of Section 5.2, comparing seven different methods for estimation and prediction. First, I use the prior distributions and

the model selection methods on the estimation set described in Section 5.2 to select explanatory variables. Second, I use a variety of methods to estimate the parameters of the models again using the estimation data set. Third, prediction intervals for the prediction set are calculated and evaluated using the statistics in Section 5.1.3.

Using AICC with independent errors to select the model, the standard linear model uses maximum likelihood estimates of the parameters. The corresponding prediction intervals follow standard theory.

To estimate parameters for BMA with independent errors, I use the same prior distributions on $\boldsymbol{\beta}$ and σ^2 as BMA Matern and integrate these parameters out. Therefore, the PMP for the independence model can be found in closed form. Following the same steps as BMA Matern (Section 4), I estimate prediction intervals. See Hoeting (1994) for more details.

The steps for spatial BMA consist of estimating parameters using the posterior distribution of $\boldsymbol{\theta}$ given the data $\boldsymbol{Z}$ (equation (4.11)). Next, following the derivations in Section 4.4.2, I calculate the PMP. Third, I approximate the quantiles of the posterior predictive distribution by using the MLE approximation (Section 3.2.3) as well as the estimate of the mean.

The spatial AICC method selects the explanatory variables following as described in Section 2.4.2. While selecting a model, the spatial AICC method also estimates the parameters from the profile likelihood. Finally, the parameter estimates are substituted into the kriging equations to obtain prediction estimates and prediction intervals.

Building on the AICC independence model selection method, I fit the spherical, exponential, and Gaussian autocorrelation functions that were described in Section 2.2 using REML estimation for $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and σ^2 (equation (2.30)). Next, the parameter estimates are substituted into the kriging equations (1.4) and (1.5) obtaining prediction estimates and prediction intervals.

Five spatial patterns are used for the simulations. The spatial patterns are fixed and can be seen in Figure 5.2. One thousand replications are run for the random pattern, while five hundred replications are run for the other five spatial patterns. The details for this simulation follow the setup in Section 5.1.

5.3.2 Results

MSPE

As described in Section 5.1.3, mean squared prediction error (MSPE) measures the departures of the predicted value from the observed value where predicted values are generated from one of the model fitting techniques: BMA Matern autocorrelation, spatial AICC Matern autocorrelation, exponential autocorrelation, spherical autocorrelation, Gaussian autocorrelation, independent BMA, and standard independence prediction. Figure 5.4 shows boxplots of MSPE when the locations for estimation follow a random pattern. This figure demonstrates the difficulties encountered when ignoring spatial autocorrelation. For this simulation set-up with autocorrelated observations the model fit when assuming independence and independent BMA have a much larger MSPE than those fit assuming non-zero spatial autocorrelation.

Gaussian Autocorrelation Function

Similarly, the Gaussian autocorrelation model also has increased MSPE compared to the other spatial autocorrelation methods. At the extreme, when the design is highly clustered, the MSPE for the Gaussian autocorrelation function is larger than the MSPE for both independent and BMA independent approaches (Table 5.7). The Gaussian autocorrelation function has smaller MSPE for grid spatial pattern with MSPE of 4.66, approximately 1.4 units larger than the MSPE for the spherical

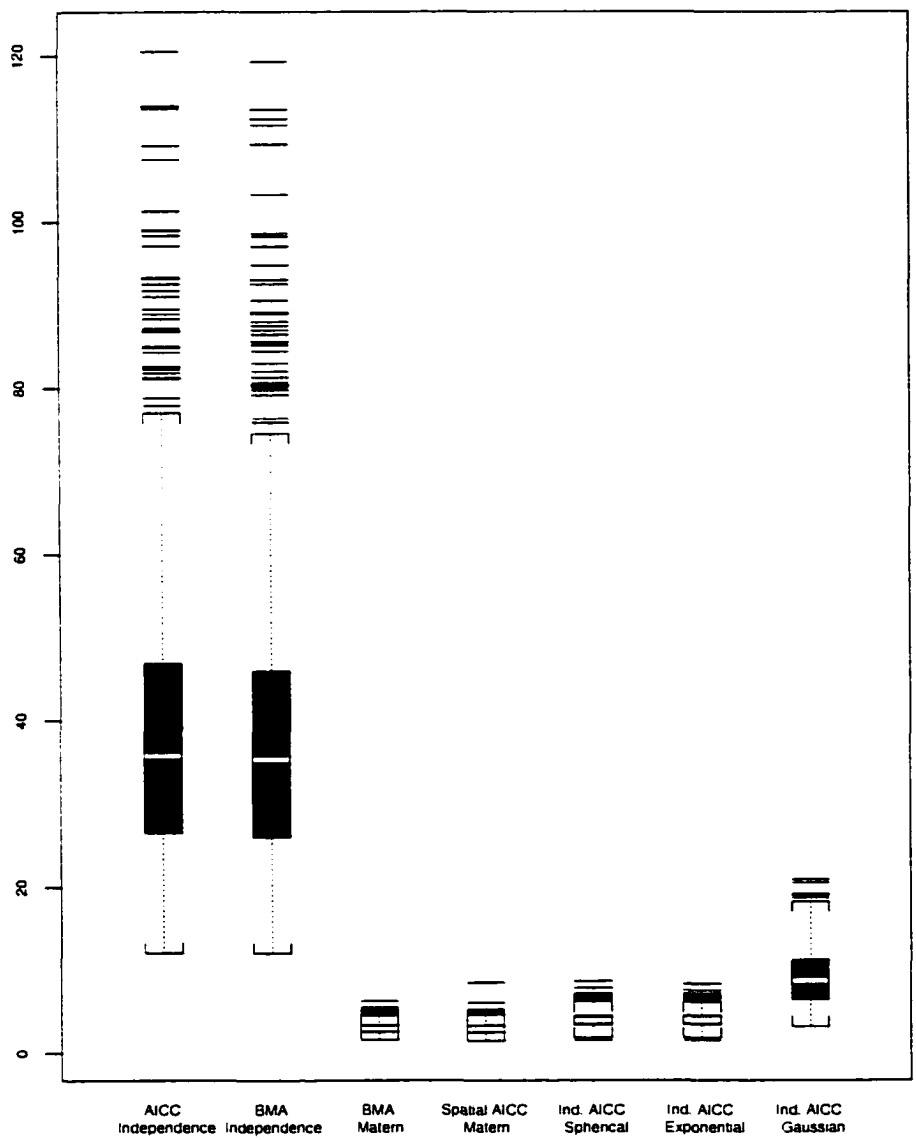

Figure 5.4: MSPE for the random pattern when the data does not include a nugget. Each observation represents the MSPE for 100 prediction locations. This figure represents 1000 replicates.

and exponential autocorrelation functions, and approximately 2 units larger than the Matern autocorrelation function and the BMA Matern (Table 5.11).

The behavior of the Gaussian autocorrelation is connected to the distance between points and the characteristics of the Gaussian autocorrelation function described in Section 2.2.1. Along with the smooth behavior of this autocorrelation function, there is very little flexibility in this function. This can result in convergence problems as well as problems with nonnegative definite correlation matrices.

For the Gaussian autocorrelation, the function may be fit according to the behavior near the origin, which leads to incorrect conclusions. As discussed in Section 5.2, when the choice of explanatory variables is made without accounting for spatial autocorrelation, the model may not include any of the important explanatory variables. For example, 42% of the simulations for the random spatial pattern result in the model with no explanatory variables.

In conclusion, the Gaussian autocorrelation function should not be fit without a nugget term because the behavior near the origin may determine the behavior of the curve, which can have dramatic results for the highly clustered design. There is a lack of flexibility in the Gaussian autocorrelation function that can cause difficulties in convergence. Since the distance between observations increases as the designs change from the highly clustered pattern, through the random pattern, regular pattern, and grid pattern, the difficulty of the Gaussian autocorrelation function decreases. For the grid spatial pattern, there is very little information near the origin, so the Gaussian autocorrelation function is free to fit the autocorrelation function to larger distances. Overall, the Gaussian autocorrelation function performs worse than the exponential and spherical autocorrelation functions. Therefore, the remainder of this simulation will concentrate on models fit using either of the Matern autocorrelation methods, exponential autocorrelation or spherical autocorrelation functions.

Comparison of Methods

Four methods, the independence, spherical, exponential and Gaussian, use AICC independence for selecting the explanatory variables. Figure 5.4 shows the dramatic improvement in MSPE by using an autocorrelation function to model the error structure.

In order to examine the results in more detail, Figure 5.5 shows MSPE for the spatial AICC model, spatial BMA, along with the spherical and exponential autocorrelation functions fit to the best AICC independent model for all five spatial designs. In general, MSPE is smaller for both the Matern autocorrelation function and the BMA using the Matern autocorrelation function when compared to the traditional methods fitting spherical and exponential autocorrelation.

The differences between the Matern models and the traditional models (spherical and exponential) can be attributed to the additional flexibility in the Matern class due to the additional smoothness parameter. Also, choosing the mean function while assuming the data are uncorrelated allows for under fitting the mean function, and missing important predictors in the analysis. This behavior can be seen in all of the spatial designs.

Comparison of Sampling Patterns

In more concrete terms, Tables 5.7 through 5.11 demonstrate that the improvement in MSPE by fitting spatial autocorrelation continues for the other spatial designs. When assuming independence of the locations, the MSPE is approximately 40 for all five sampling patterns. In comparison, the MSPE for the spatial AICC Matern autocorrelation function ranges from 10.45 for the highly clustered pattern (Table 5.7) to 2.61 for the regular pattern (Table 5.10). It is clear that in terms of MSPE, fitting a spatial autocorrelation has a dramatic improvement on the accuracy of prediction when the data are autocorrelated.

Focusing on the sampling patterns, the highly clustered design has a large number of points with small distances between points, which results in a good deal of information about the behavior near the origin. In fact, when comparing the number of pairs of locations with distances less than 3 units apart over all the spatial designs, (Figure 5.3) dramatic differences appear. At one extreme, the grid spatial pattern has pairs at discrete distances, and therefore very little information that is useful to describe the continuous behavior of the spatial autocorrelation. At the other extreme, the highly clustered spatial pattern has pairs at a variety of small distances which can assist in describing the continuous spatial autocorrelation.

The highly clustered process has the largest MSPE when compared to the other spatial designs, because it has the least coverage in terms of points used for fitting the model over the entire space. Therefore, this results in areas where there is very little information about the mean structure, and no points that are highly correlated to assist in improving the prediction. The highly clustered process also provides the most information about small scale autocorrelation, due to the large number of observations at small distances. The differences between the spatial BMA and spatial AICC models are slight for MSPE for this design, both use the Matern autocorrelation function. Similarly the differences between the spherical and exponential are slight.

For a given model, the random spatial pattern (Table 5.9) has smaller MSPE than the two clustered patterns. The reduction in MSPE continues as the designs form a more regular pattern over the domain $(0, 10)^2$. As mentioned above, The highly clustered process has a large gaps in the coverage over the space $(0, 10)^2$, and therefore is lacking in information about the mean structure over areas of the domain. In comparison, the grid design has no gaps in its coverage, but it also has very little information about the autocorrelation structure, and very few point pairs with large autocorrelations that can assist in prediction. Good predictions need

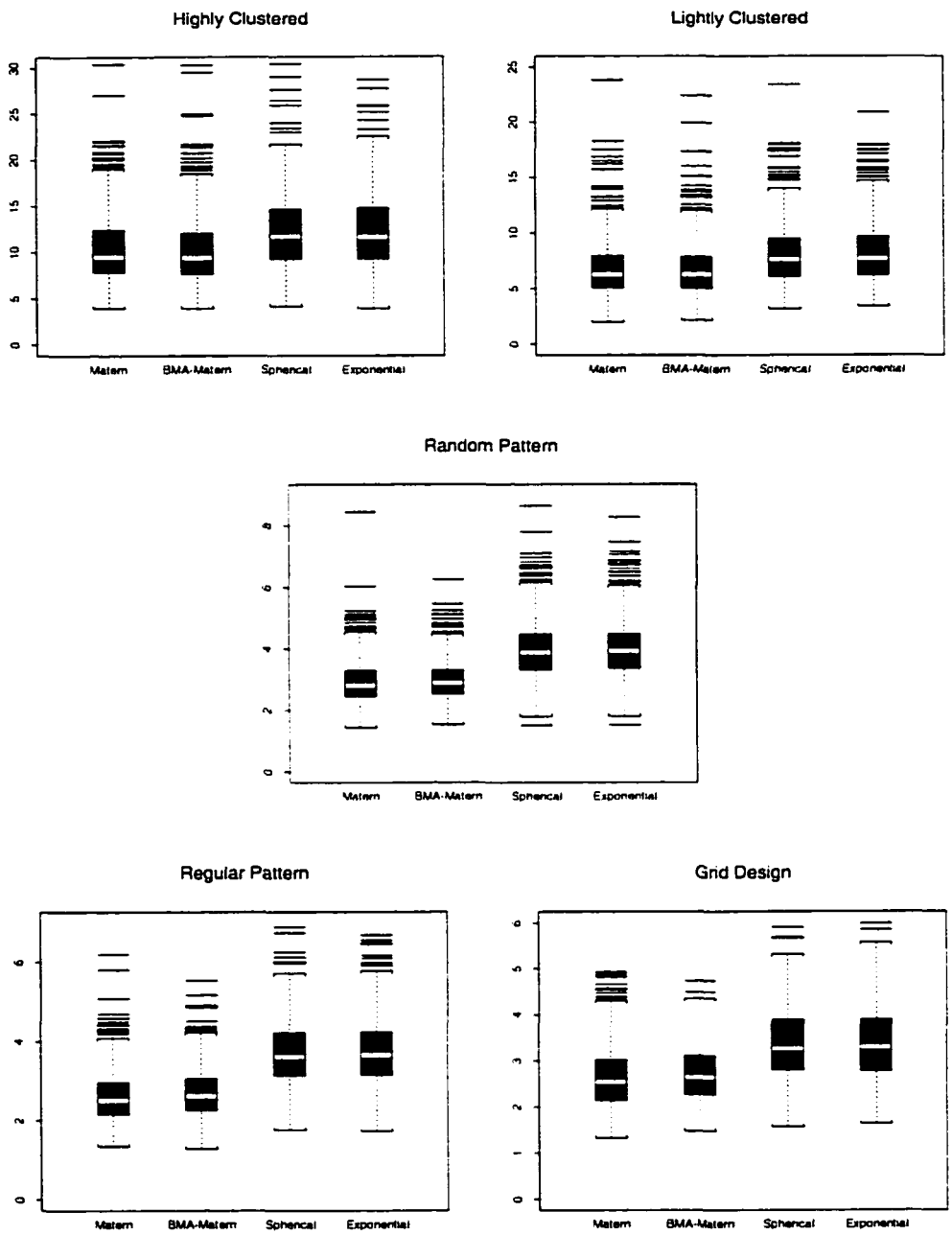

Figure 5.5: MSPE for Different Spatial Designs. This includes BMA Matern, spatial AICC Matern, Spherical and Gaussian models. Note that the y-axis scale is different for each plot.

Table 5.7: Summary Statistics for Highly Clustered Pattern without Nugget.

	Ind	BMA-Ind	BMA-Matern	Matern	Spherical	Exponential	Gaussian
MSPE	40.48	39.51	10.26	10.45	12.36	12.40	45.51
EPC	0.93	0.95	0.94	0.94	0.92	0.92	0.92
minPC	0.00	0.00	0.49	0.17	0.44	0.37	0.00
meanPC	0.93	0.95	0.95	0.94	0.92	0.92	0.92
maxPC	1.00	1.00	1.00	1.00	1.00	1.00	1.00
stdPC	0.26	0.29	0.07	0.13	0.08	0.09	0.22

Table 5.8: Summary Statistics for Lightly Clustered Pattern without Nugget.

	Ind	BMA-Ind	BMA-Matern	Matern	Spherical	Exponential	Gaussian
MSPE	42.52	42.07	6.72	6.80	8.13	8.18	27.83
EPC	0.93	0.95	0.95	0.93	0.94	0.94	0.91
minPC	0.00	0.00	0.02	0.03	0.00	0.00	0.00
meanPC	0.93	0.95	0.95	0.93	0.94	0.94	0.91
maxPC	1.00	1.00	1.00	1.00	1.00	1.00	1.00
stdPC	0.24	0.28	0.20	0.21	0.36	0.37	0.10

Table 5.9: Summary Statistics for Random Sampling Pattern without Nugget.

	Ind	BMA-Ind	BMA-Matern	Matern	Spherical	Exponential	Gaussian
MSPE	38.85	38.17	2.96	2.92	3.94	3.95	8.99
EPC	0.95	0.96	0.94	0.92	0.96	0.96	0.86
minPC	0.00	0.00	0.00	0.00	0.00	0.00	0.00
meanPC	0.95	0.96	0.94	0.92	0.96	0.96	0.86
maxPC	1.00	1.00	1.00	1.00	1.00	1.00	1.00
stdPC	0.18	0.24	0.23	0.25	0.33	0.33	0.05

Table 5.10: Summary Statistics for Regular Pattern without Nugget.

	Ind	BMA-Ind	BMA-Matern	Matern	Spherical	Exponential	Gaussian
MSPE	39.14	38.48	2.72	2.61	3.75	3.73	6.06
EPC	0.95	0.96	0.94	0.89	0.97	0.97	0.84
minPC	0.00	0.00	0.01	0.00	0.17	0.17	0.00
meanPC	0.95	0.96	0.94	0.89	0.97	0.97	0.84
maxPC	1.00	1.00	1.00	1.00	1.00	1.00	1.00
stdPC	0.18	0.25	0.29	0.30	0.20	0.19	0.02

Table 5.11: Summary Statistics for Grid Sampling Design without Nugget.

	Ind	BMA-Ind	BMA-Matern	Matern	Spherical	Exponential	Gaussian
MSPE	40.56	39.52	2.72	2.63	3.37	3.36	4.66
EPC	0.94	0.96	0.93	0.89	0.97	0.97	0.74
minPC	0.00	0.00	0.00	0.00	0.04	0.05	0.00
meanPC	0.94	0.96	0.93	0.89	0.97	0.97	0.75
maxPC	1.00	1.00	1.00	1.00	1.00	1.00	1.00
stdPC	0.14	0.21	0.27	0.27	0.15	0.14	0.10

coverage over the entire area of study, plus enough observations so that information about the autocorrelation structure can improve prediction in terms of MSPE.

If the spatial design is fixed, for example due to the logistics of sampling the data, information about the relative improvements that Matern autocorrelation functions can add to prediction is very important. Figure 5.6 plots the relative MSPE, equation (5.3). Values closer to one are considered better because the prediction error is approximately equal to the simulated error. For all designs the Matern models have relative MSPE closer to one when compared to the exponential and Gaussian methods. The clustered patterns show that although the Matern autocorrelation functions have larger (Figure 5.5) MSPE. Even the Gaussian model, which does not fit the mean function well (as seen in Figure 5.4), come close to a rMSPE value of 1 when compared across spatial designs (Figure 5.6). As the distance between points increases, and the data becomes less clustered, relative MSPE is increased for all methods. Again, the reduction in information about the autocorrelation function and the small scale variability, reduce the ability of the models to predict at close distances, and therefore increases the relative MSPE.

The decision of whether or not to use the Matern class of autocorrelation functions needs to take into account the amount of precision necessary in the prediction or confidence interval. When slight decreases in MSPE are not important, standard methods for spatial prediction are acceptable. When small increases in MSPE are important, then using the Matern autocorrelation function yields improvements.

Empirical Predictive Coverage (EPC) and Predictive Coverage (PC)

Empirical predictive coverage, EPC, is the number of confidence intervals that contain the observed response. Plots of EPC for the five spatial designs without nugget can be seen in Figure 5.7. Predictive coverage, PC, calculates the probability under the true model that the estimate is in the prediction interval. Plots of PC for

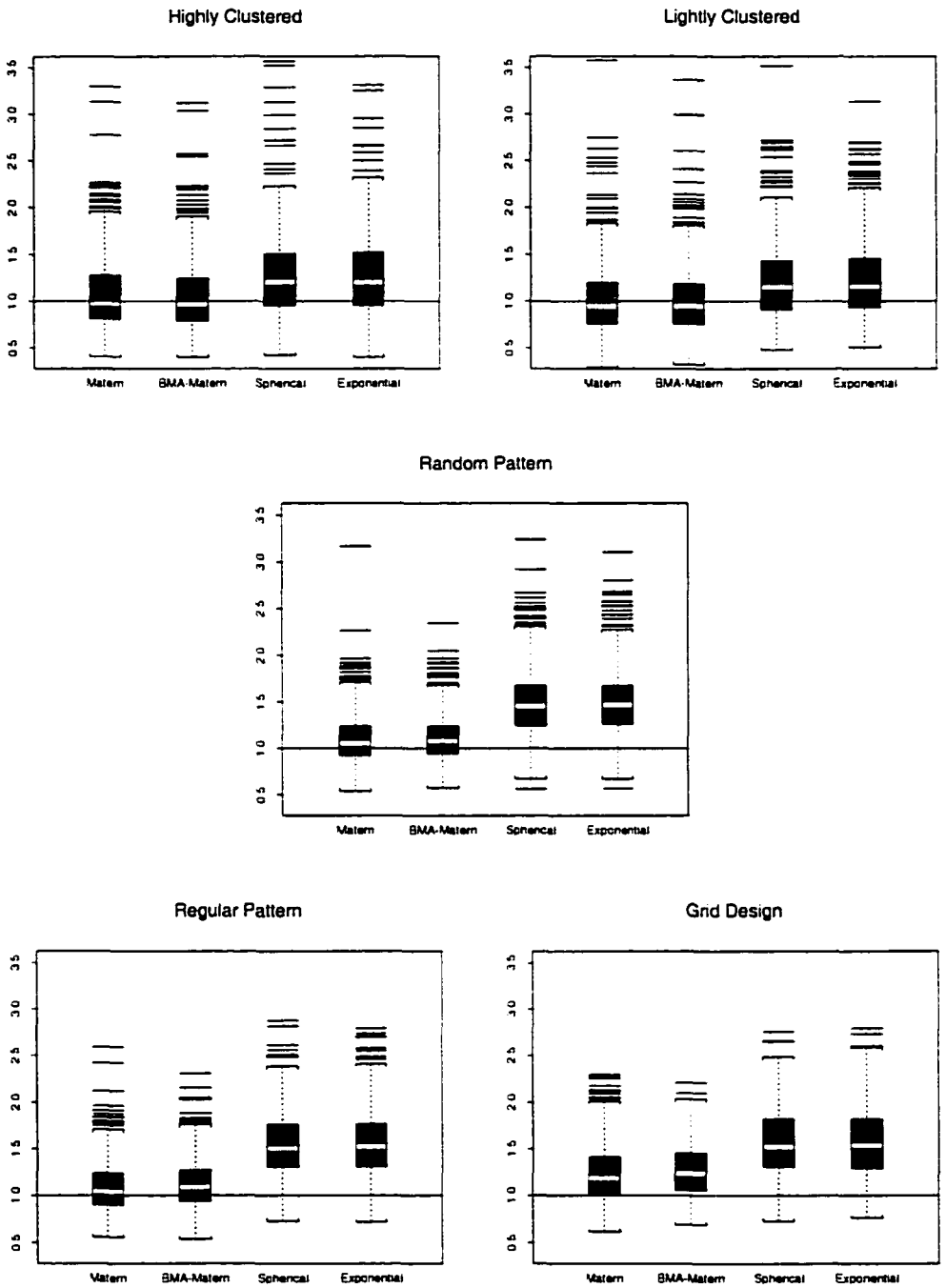

Figure 5.6: Relative MSPE for Different Spatial Designs. Note that the y-axis scale is the same for each plot.

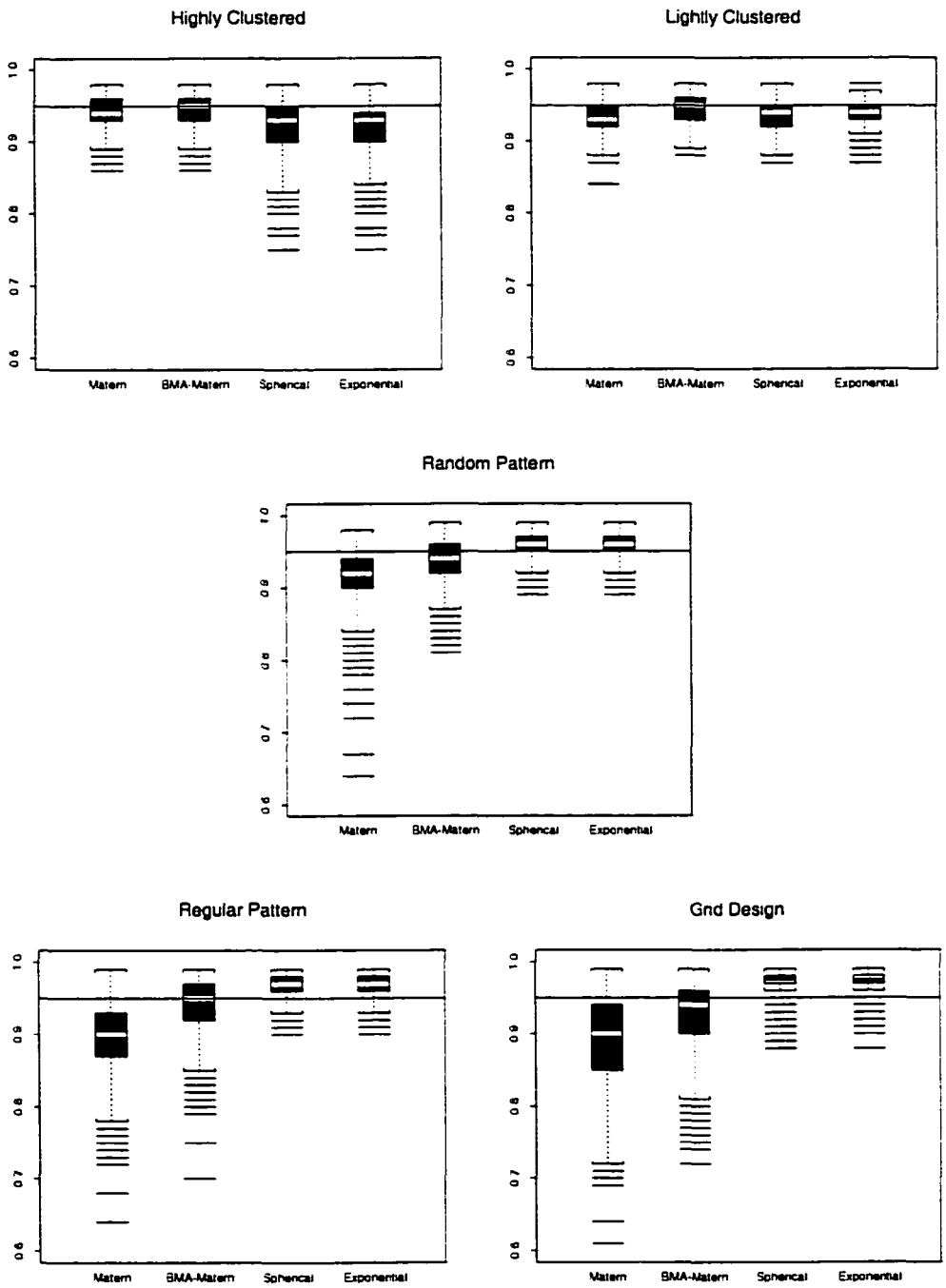

Figure 5.7: EPC for Different Spatial Designs.

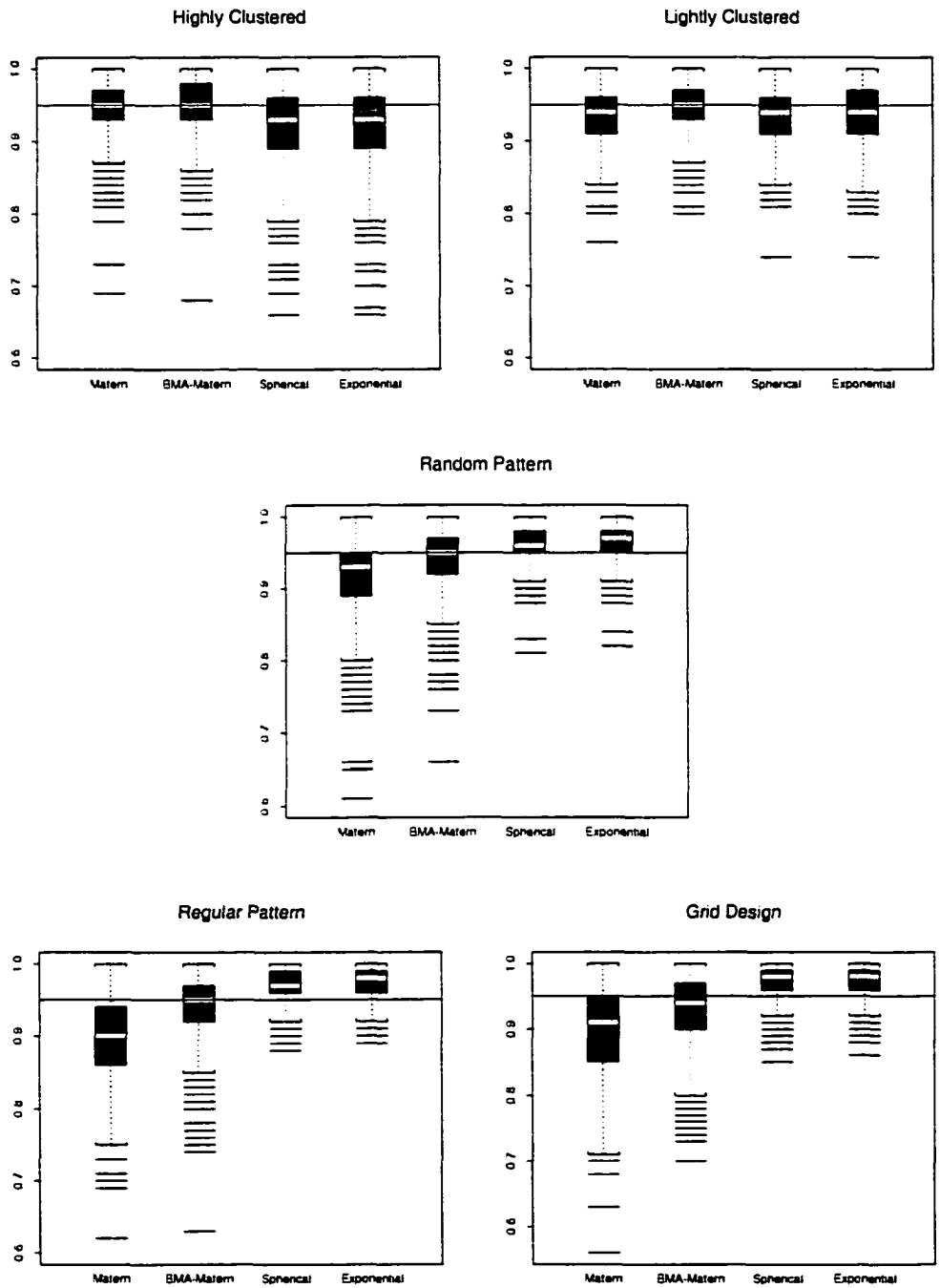

Figure 5.8: PC for Different Spatial Designs.

the five spatial designs without nugget can be seen in Figure 5.8. Comparing the two figures, PC and EPC are almost identical. This is reassuring for modeling real data, when it is not possible to calculate the PC statistics as described in Section 5.1.3.

Each prediction and confidence interval is calculated at the 95% level, therefore the results of EPC and PC should be 0.95. Over all designs, the BMA-Matern intervals are closest to the 95% level. This supports the contention that BMA-Matern predictive intervals more accurately describe the variability in prediction. The Matern-AICC model tends to be below the nominal predictive level for all five designs, although the amount of underestimation increases as the designs become less clustered.

The spherical and exponential autocorrelation models exhibit the most design dependent behavior. When the data are clustered, for example in the highly clustered pattern, the confidence levels are lower than 95%. This corresponds to underestimating the variance of the prediction, and therefore the confidence intervals are too small. On the other hand, the random spatial pattern, the regular pattern, and the grid pattern result in overestimates of the variance, and confidence intervals that are too large and therefore the confidence levels are above the 95% level. Similar behavior can be seen in Figure 5.8, which is only available when the true distribution of the observations is know.

In examining PC, it is the maximum, minimum and standard deviation that are of interest. Focusing on the random pattern in Table 5.9, all six methods have identical minimum and maximum values. There is some variability in standard deviation, although the differences are not large and are not consistent for different sampling patterns.

Conclusion

Prediction interval accuracy varies across sampling pattern. When comparing the prediction results above to the model selection results in Section 5.2, sampling

pattern has a very different effect. Both of the spatial model selection techniques perform best in identifying the correct explanatory variables and also has the largest MSPE for the highly clustered pattern compared to the other sampling patterns. Conversely, the spatial model selection techniques does the worst for model selection but has the smallest MSPE for the grid pattern. For all five sampling patterns, the results for PC and EPC stay relatively consistent for the Matern autocorrelation functions although they vary for the spherical and exponential. In selecting a sampling pattern, it is very important to consider the goals of the analysis.

Examining MSPE, EPC and PC, the conclusions from this simulation are clear. Using the Matern autocorrelation function with either simultaneous model selection or BMA reduces MSPE and has better predictive coverage than other methods. BMA Matern does the best for all designs in terms of accurately assessing the variability of prediction.

5.4 Simulations with Nugget

This simulation examines the importance of allowing the nugget, $\theta_3 \geq 0$ when the data includes some amount of measurement error. Focusing on the Matern class of autocorrelation functions, I compare BMA with a nugget term to BMA without a nugget term.

5.4.1 Details

This simulation builds on the structure in Section 5.1 with different parameter values for σ^2 and θ_3. First, in order to investigate the impact of measurement error in the data, I will simulate from two different models. The first includes 20% measurement error or $\theta_3 = 0.20$. The second includes 50% measurement error or $\theta_3 = 0.50$.

In order to identify the signal from the noise, I reduced the variance of the data for this simulation, fixing $\sigma^2 = 10$. New hyperparameters for σ^2 were necessary to

retain a proper posterior distribution and allow for convergence of the estimation methods. I choose $(\nu, \lambda) = (12, 8)$ for this study.

The prior distributions on $\boldsymbol{\theta}$ are uniform on the interval $(0, d_{max})$ for θ_1, inverse gamma with parameters $(2, 1)$ for θ_2, and uniform on the interval $(0, 1)$ on θ_3. The prior on θ_2 has a mean of 1 and infinite variance, which arises from similar priors of (Ecker and Gelfand, 1997). Prior sensitivity for $\boldsymbol{\theta}$ is discussed in Section 5.4.3.

Two hundred simulations were done for each type of sampling design and at each level of θ_3, $\theta_3 \in (0.20, 0.50)$. The five sampling patterns are fixed, and generation of each replicate follows the steps in Section 5.1.

5.4.2 Results

Model Selection

For practical purposes, there are no differences between including or not including a nugget term. Tables 5.12 and 5.13 display the average PMP for all five sampling patterns. This simulation incorporates model uncertainty into the results. For example, two largest models for the random pattern with $\theta_3 = 0.20$ fit with $\theta_3 \geq 0$ accounts for 62.7% of the PMP for the random pattern. The top five models account for 82.9% of the PMP.

When measurement error is included in the data, it is more difficult to estimate parameters and select the correct explanatory variables. The PMP for the true model range from 6.6% to 10.0% for the simulation with $\theta_3 = 0.20$, and ranges from 2.7% to 5.5% when $\theta_3 = 0.50$. Similarly, for the model with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (M_7) the PMP ranges from 30.0% to 43.8% when $\theta_3 = 0.20$ and 14.2% to 19.9% when $\theta_3 = 0.50$. As a consequence, the PMP for the model with no explanatory variables is greater when $\theta_3 = 0.50$ compared to $\theta_3 = 0.20$.

In conclusion, sampling pattern and the inclusion of the parameter θ_3 does not alter the model selection results. Instead, it is the amount of nugget that effects the ability to model the data.

Table 5.12: Average PMP (%) for the Simulation with Nugget when $\theta_3 = 0.20$.

Model	Highly Clustered		Lightly Clustered		Random Pattern		Regular Pattern		Grid Design	
	$\theta_3 \geq 0$	$\theta_3 = 0$	$\theta_3 \geq 0$	$\theta_3 = 0$	$\theta_3 \geq 0$	$\theta_3 = 0$	$\theta_3 \geq 0$	$\theta_3 = 0$	$\theta_3 \geq 0$	$\theta_3 = 0$
M_1	2.4	3.6	0.6	0.6	1.4	1.5	1.9	2.1	1.3	1.2
M_2	26.4	27.2	18.5	18.1	24.3	23.7	27.8	27.7	28.0	26.5
M_3	4.5	4.1	6.6	7.1	4.6	4.3	4.5	4.3	5.4	5.2
M_4	0.3	0.4	0.0	0.0	0.1	0.1	0.3	0.3	0.1	0.1
M_5	0.1	0.2	0.0	0.0	0.2	0.2	0.2	0.2	0.1	0.1
M_6	0.1	0.2	0.1	0.1	0.3	0.2	0.1	0.1	0.1	0.2
M_7	35.6	34.0	43.8	42.2	38.4	37.0	31.9	30.0	36.9	36.9
M_8	8.3	7.5	5.8	6.2	6.4	6.4	9.1	10.1	7.2	7.0
M_9	1.8	2.0	1.0	1.1	1.5	1.6	1.9	2.4	2.1	2.2
M_{10}	1.4	1.6	2.0	2.1	2.3	2.6	2.5	2.7	2.4	2.5
M_{11}	0.4	0.3	0.5	0.6	0.4	0.4	0.4	0.4	0.4	0.4
M_{12}	0.3	0.2	0.8	0.8	0.6	0.6	0.4	0.5	0.4	0.4
M_{13}	0.9	0.8	0.5	0.7	0.3	0.3	0.3	0.3	0.5	0.5
M_{14}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_{15}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_{16}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_{17}	10.0	9.9	9.2	8.8	9.2	9.5	8.1	7.7	6.6	7.1
M_{18}	2.4	2.5	3.3	3.5	3.3	3.6	3.4	3.6	3.0	3.1
M_{19}	2.1	2.3	4.2	4.5	3.1	3.7	2.8	3.0	2.6	2.7
M_{20}	0.7	0.7	0.4	0.5	0.6	0.8	0.8	0.9	0.6	0.7
M_{21}	0.4	0.4	0.7	0.7	0.6	0.6	1.0	1.2	0.6	0.6
M_{22}	0.2	0.2	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2
M_{23}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_{24}	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
M_{25}	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
M_{26}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_{27}	0.9	0.9	0.6	0.7	0.7	0.9	0.7	0.7	0.7	0.7
M_{28}	0.5	0.6	0.8	0.9	1.0	1.1	1.2	1.2	0.4	0.5
M_{29}	0.1	0.2	0.3	0.3	0.2	0.3	0.3	0.3	0.2	0.2
M_{30}	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.1
M_{31}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_{32}	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.8

Table 5.13: Average PMP (%) for the Simulation with Nugget when $\theta_3 = 0.50$.

Model	Highly Clustered		Lightly Clustered		Random Pattern		Regular Pattern		Grid Design	
	$\theta_3 \geq 0$	$\theta_3 = 0$	$\theta_3 \geq 0$	$\theta_3 = 0$	$\theta_3 \geq 0$	$\theta_3 = 0$	$\theta_3 \geq 0$	$\theta_3 = 0$	$\theta_3 \geq 0$	$\theta_3 = 0$
M_1	11.6	11.5	7.1	7.3	10.0	10.2	9.6	9.3	11.9	11.6
M_2	30.3	28.9	27.8	27.6	28.3	27.7	32.2	30.9	31.8	31.0
M_3	11.8	12.1	14.0	13.5	13.9	13.3	8.6	8.6	9.3	9.2
M_4	0.9	1.1	0.6	0.8	0.8	0.9	1.5	1.5	1.3	1.3
M_5	0.9	1.0	0.7	0.8	1.7	1.7	1.2	1.3	1.0	1.0
M_6	1.2	1.6	0.7	0.8	1.1	1.2	1.6	1.6	1.0	1.0
M_7	15.1	15.4	19.4	19.9	17.2	17.1	14.2	14.8	15.9	16.0
M_8	7.0	6.8	8.2	8.5	4.4	4.6	7.3	7.5	8.6	8.7
M_9	3.2	3.2	2.1	2.2	3.6	3.8	4.4	4.4	2.9	3.0
M_{10}	3.3	3.2	2.4	2.5	4.0	4.0	4.9	5.1	2.6	2.6
M_{11}	0.9	0.9	1.7	1.3	1.5	1.6	2.4	2.1	1.7	1.8
M_{12}	1.6	1.6	1.3	1.0	1.7	1.7	0.8	0.8	0.8	0.8
M_{13}	1.3	1.4	1.5	1.6	1.3	1.4	0.9	1.0	0.8	0.8
M_{14}	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1
M_{15}	0.1	0.2	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1
M_{16}	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.1	0.1
M_{17}	2.7	2.8	5.5	5.1	3.2	3.0	3.2	3.4	3.6	3.9
M_{18}	2.5	2.5	1.4	1.6	1.6	1.8	1.5	1.5	1.6	1.7
M_{19}	1.5	1.6	2.2	2.3	2.1	2.2	1.3	1.4	1.5	1.6
M_{20}	1.1	0.9	0.7	0.7	0.6	0.6	1.0	1.1	0.8	0.8
M_{21}	0.6	0.8	0.6	0.6	0.6	0.7	0.8	0.8	0.9	0.8
M_{22}	0.3	0.4	0.2	0.2	0.4	0.4	0.6	0.6	0.3	0.3
M_{23}	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.2
M_{24}	0.1	0.1	0.2	0.1	0.1	0.1	0.2	0.2	0.1	0.1
M_{25}	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1
M_{26}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_{27}	0.6	0.6	0.3	0.4	0.4	0.4	0.3	0.4	0.5	0.5
M_{28}	0.3	0.3	0.5	0.5	0.3	0.4	0.3	0.3	0.4	0.4
M_{29}	0.2	0.3	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.2
M_{30}	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1
M_{31}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M_{32}	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Parameter Estimation

The question in parameter estimation is how does the estimation of the nugget term affect the estimation of the other spatial parameters. If the Matern class of functions is able to adequately describe the autocorrelation of the data while fixing $\theta_3 = 0$, then including a nugget term may not be an issue.

Focusing first on the simulation with 20% measurement error. Figures 5.9 and 5.10 show the corresponding parameter estimates for θ_1 and θ_2. When $\theta_3 \geq 0$, the sampling pattern does not alter the estimates and underestimate both θ_1 and θ_2. Conversely, when $\theta_3 = 0$ the behavior is sampling pattern dependent.

The random pattern estimates θ_1 well, but underestimates θ_2. There is a direct relationship between the estimation of the two parameters for the different sampling patterns. As an explanation, Figure 5.14 shows the average autocorrelation curves for the two estimation techniques. For the highly clustered pattern, θ_1 is overestimated and θ_2 underestimated in order to match the autocorrelation at small distances. The grid design does not have as many difficulties in estimation because it has very few observations at small distances and so the functions are able to focus on the overall curve.

When the patterns are clustered, estimation with $\theta_3 = 0$ tends to overestimate the tail of the correlation curve. In terms of the autocorrelation functions, incorporating θ_3 into the analysis is most important for these clustered sampling patterns. The differences between $\theta_3 = 0$ and $\theta_3 \geq 0$ for random, regular and grid patterns seem slight.

Focusing on the simulation with 50% measurement error, the results are less dramatic. Estimation of θ_1 does not seem to change based on the inclusion of θ_3 compared to $\theta_3 = 0$. Estimation of θ_2 does depend on sampling pattern when a nugget was not included, increasing as the number of observations decreases. Examining the autocorrelation curves in Figure 5.15 the behavior is similar to Figure 5.14 although

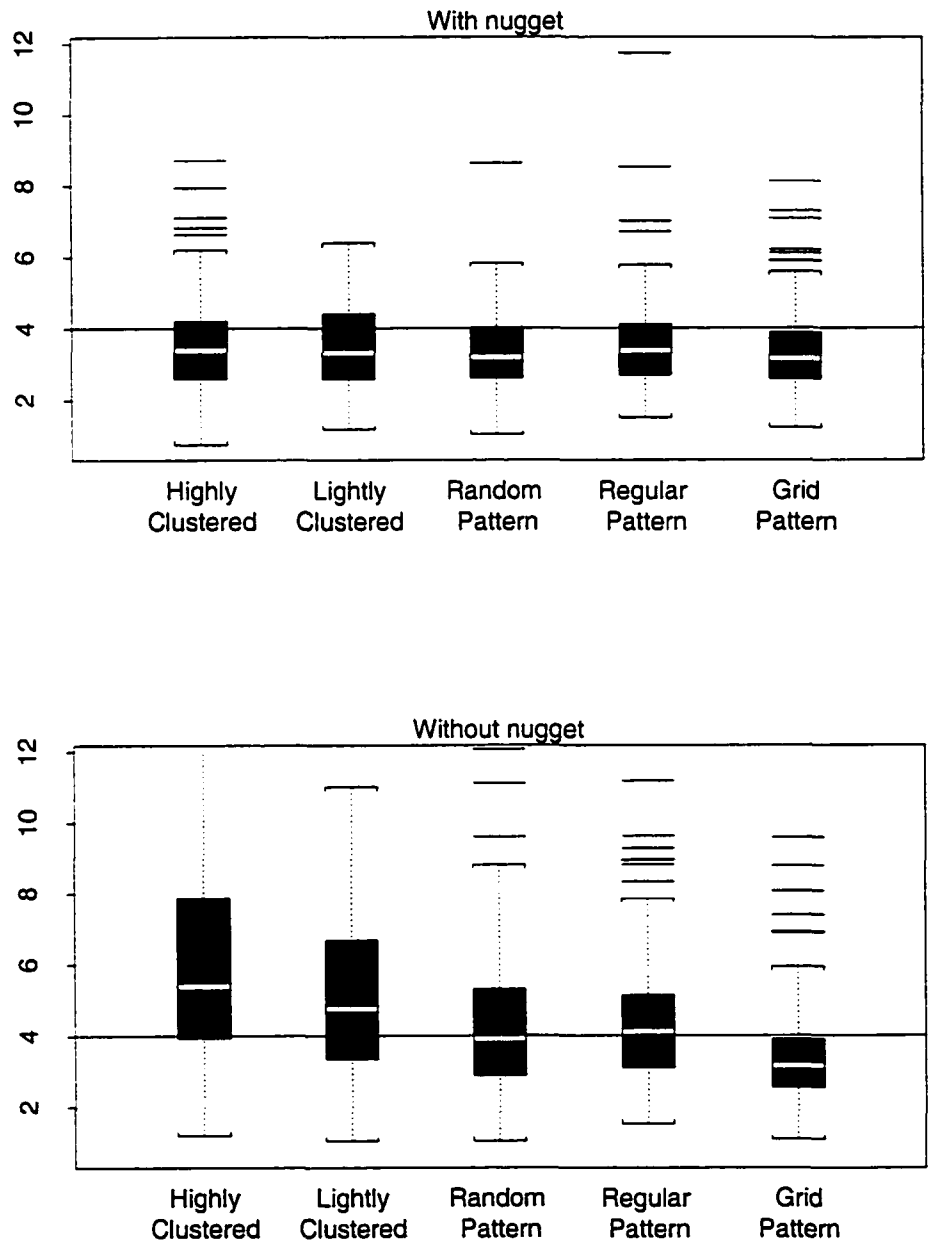

Figure 5.9: Comparison of Estimates of θ_1 for Matern Nugget and Matern no Nugget for all Spatial Patterns when the True Nugget is 0.20. The true value for θ_1 is 4.

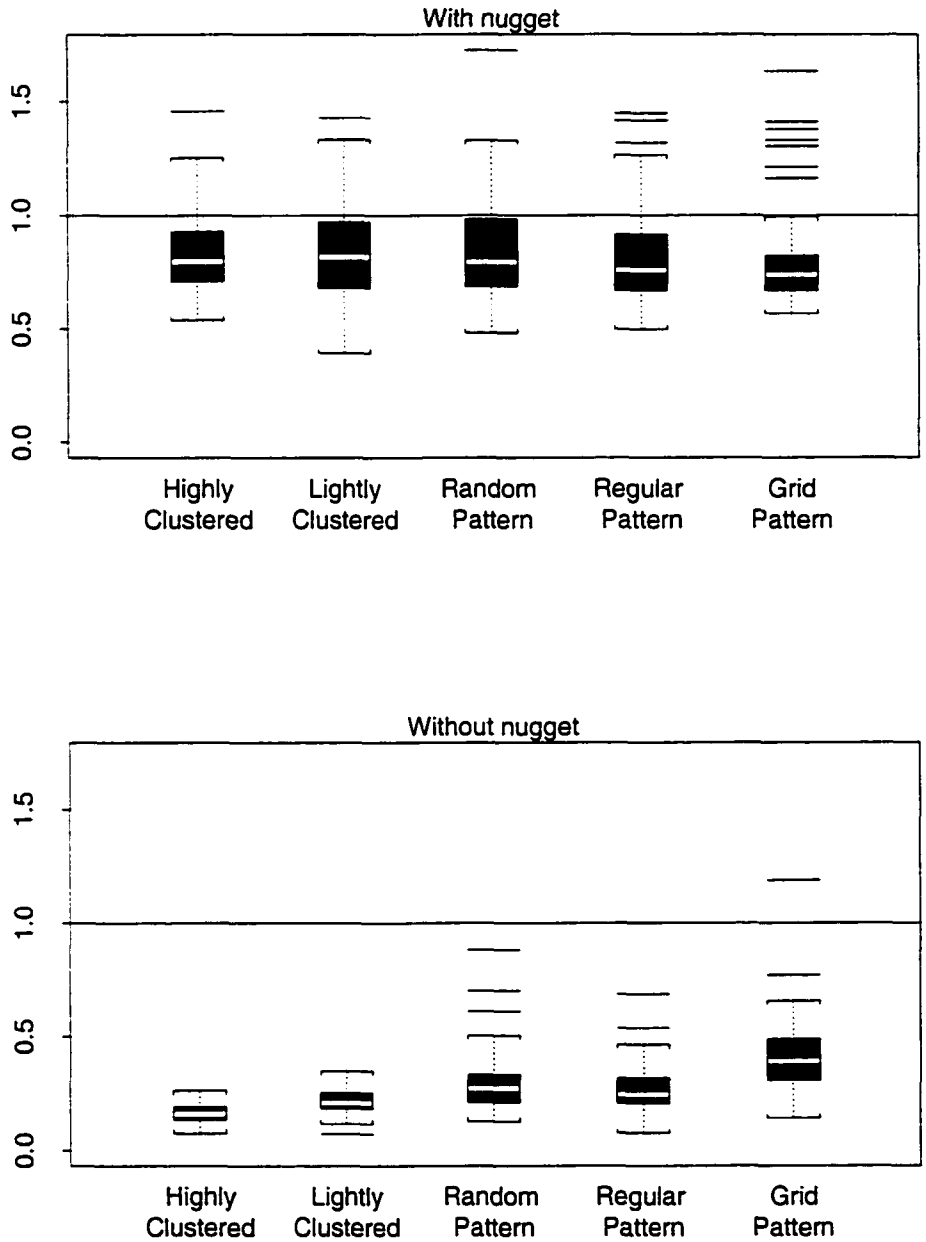

Figure 5.10: Comparison of Estimates of θ_2 for Matern Nugget and Matern no Nugget for all Spatial Patterns when the True Nugget is 0.20. The true value for θ_2 is 1.

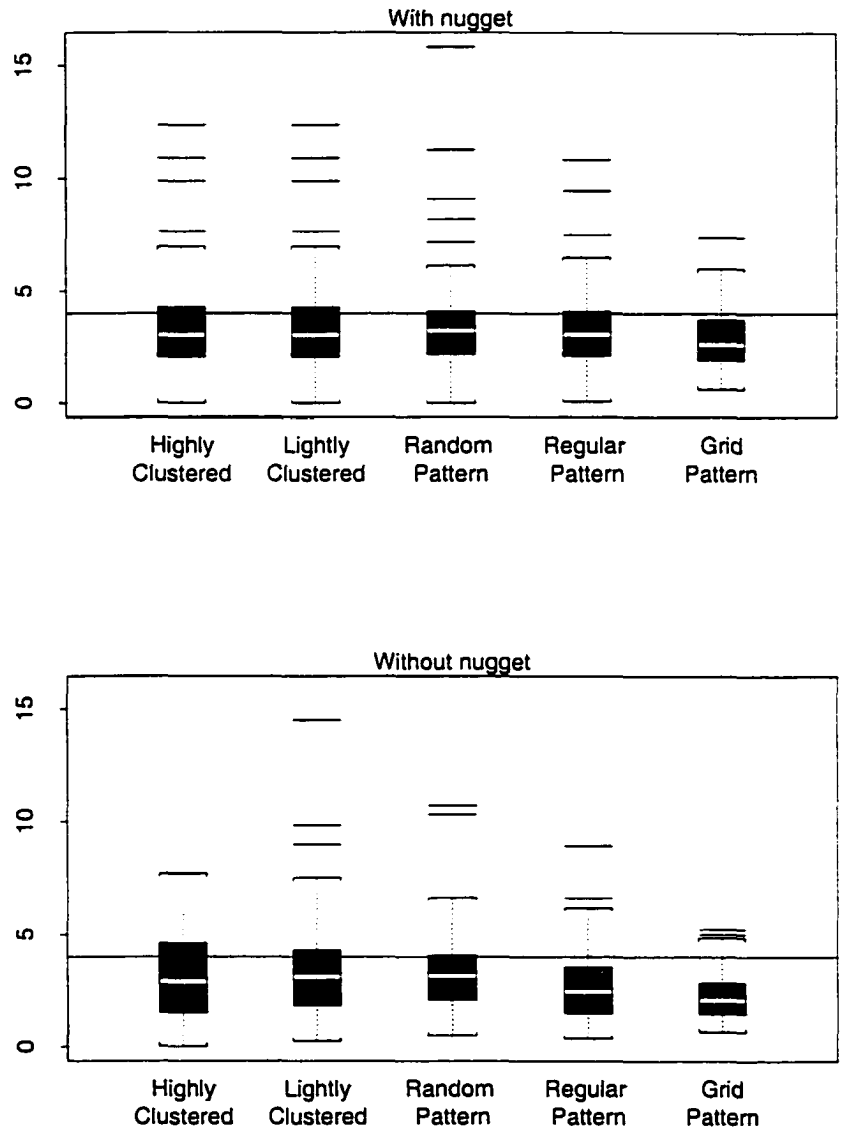

Figure 5.11: Comparison of Estimates of θ_1 for Matern Nugget and Matern no Nugget for all Spatial Patterns when the True Nugget is 0.50. The true value for θ_1 is 4.

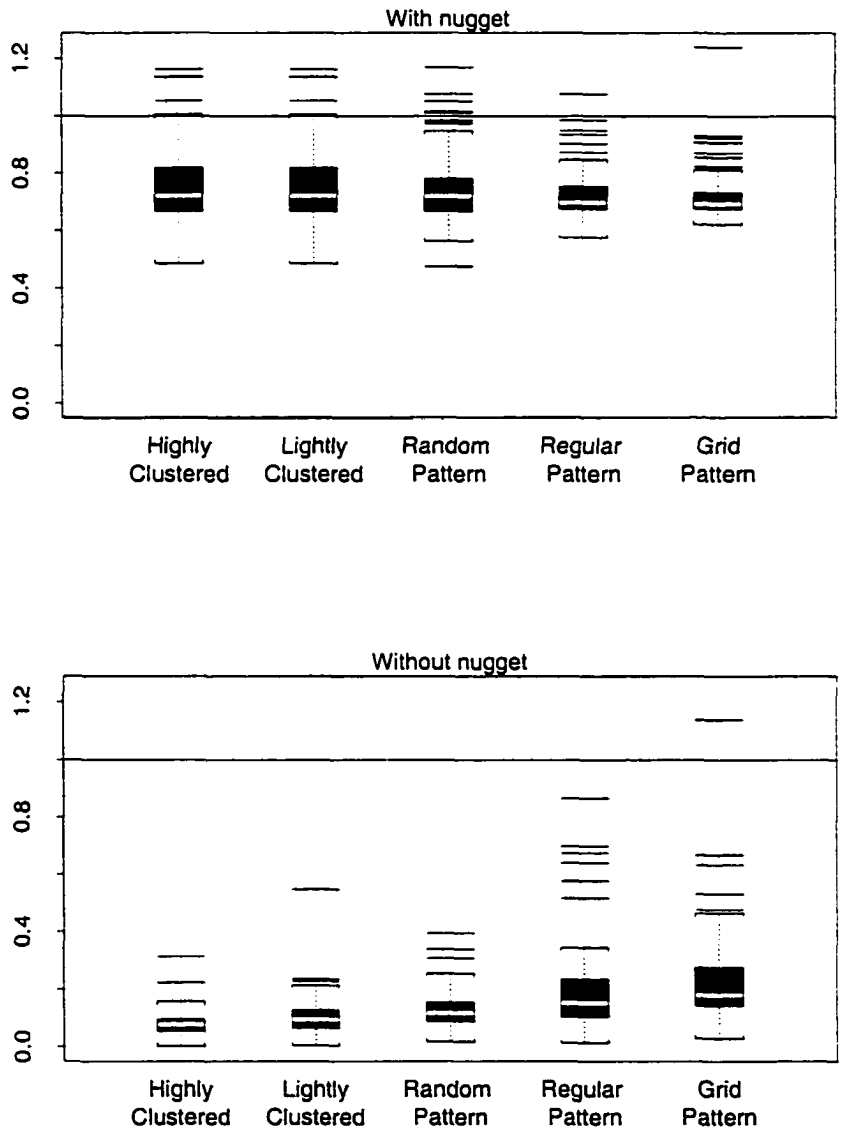

Figure 5.12: Comparison of Estimates of θ_2 for Matern Nugget and Matern no Nugget for all Spatial Patterns when the True Nugget is 0.50. The true value for θ_2 is 1.

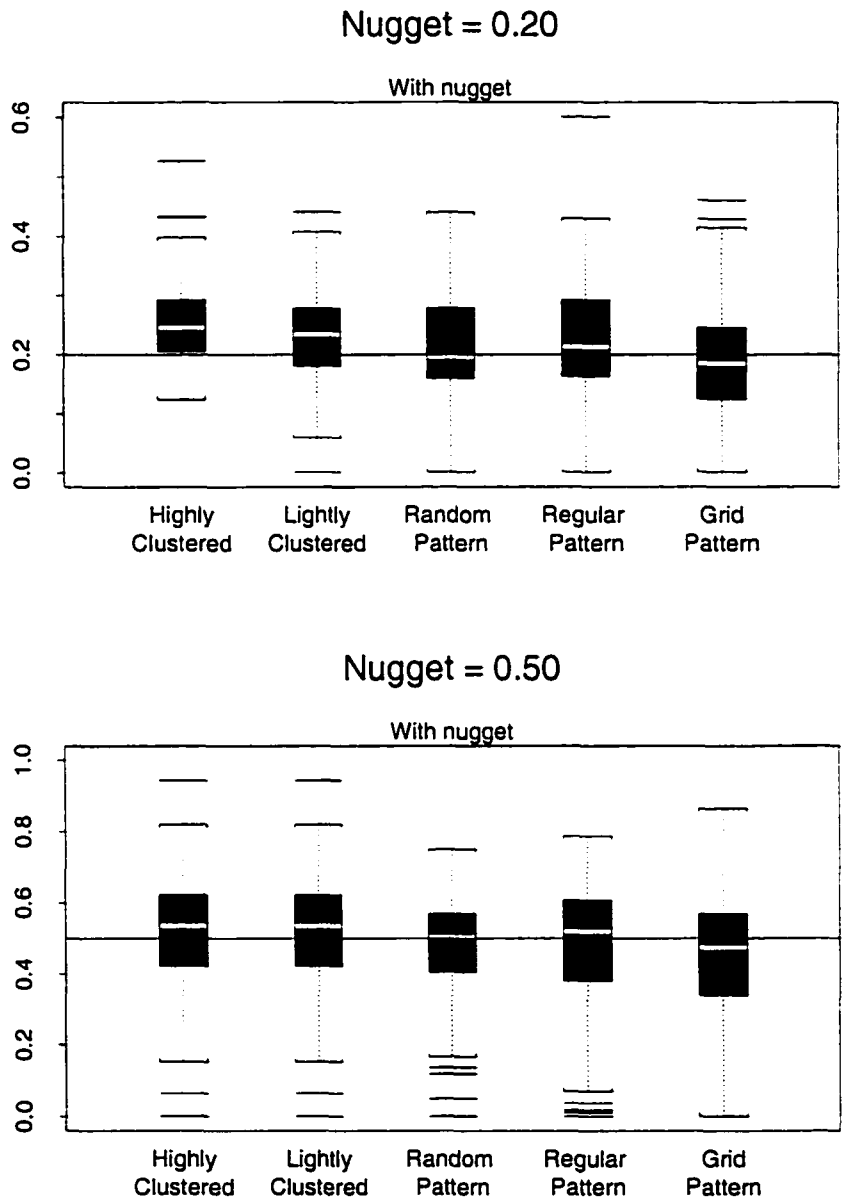

Figure 5.13: Comparison of Estimates of θ_3 for Matern with Nugget for all Spatial Patterns when the True Nugget is 0.20 and 0.50.

to a lesser extent. The highly clustered pattern has the most problems in estimating the autocorrelation function without the flexibility added by the inclusion of θ_3.

Estimation of θ_3 for both simulations does well. seen in Figure 5.13. There is a slight progression of overestimation with the highly clustered pattern to underestimation for the grid pattern. with the random pattern as most accurate.

In conclusion. estimation of the parameters is better when $\theta_3 \geq 0$ compared to $\theta_3 = 0$ as expected. But. the autocorrelation curves are able to compensate for the reduction in flexibility. The next sections focus on comparing the resulting prediction under different scenarios.

Predictive Statistics

There is no significant difference between $\theta_3 \geq 0$ and $\theta_3 = 0$ in terms of MSPE for either of the simulations (Figures 5.16 and 5.17). This indicates that the addition of a nugget term does not offer any improvement MSPE. The progression of MSPE values from highly clustered down to the grid pattern is the same behavior that was observed in Section 5.3. One difference between the simulations with 20% nugget and with 50% nugget is the larger the percentage of measurement error. the larger the MSPE.

The results for EPC and PC are identical for the two nugget percentages. therefore I will focus on EPC. In Figure 5.18. EPC is plotted for each of the five sampling patterns when the nugget is 20% and Figure 5.19 displays the plots when nugget is 50%. For the highly clustered pattern. not including a nugget results in an EPC slightly above the nominal 95^{th} percentile while including a nugget has a mean at the 95^{th} percentile for 20% nugget. The lightly clustered pattern has the opposite effect. where including a nugget has a slightly lower median EPC value compared to not including a nugget having the desired mean EPC for the same nugget. For both of these designs. the slight difference in variability is due to the simulation and

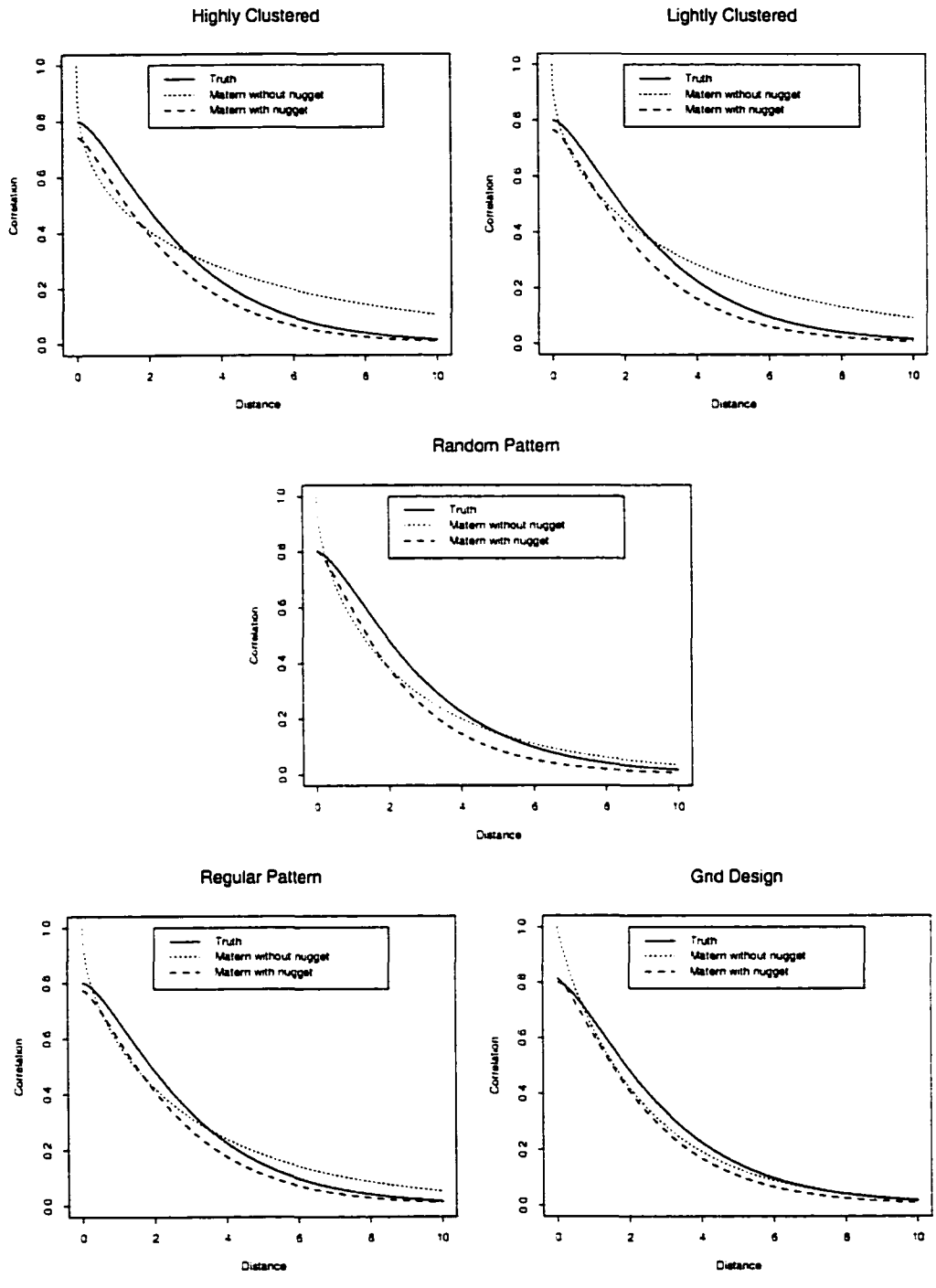

Figure 5.14: Comparison of Median Autocorrelation Functions for the Matern with Nugget and Matern without Nugget when $\theta_3 = 0.2$. The solid line is the true function, the dashed line is Matern with nugget, and the dotted line is Matern without nugget. The median autocorrelation function is based on the median values of θ_1, θ_2, and θ_3.

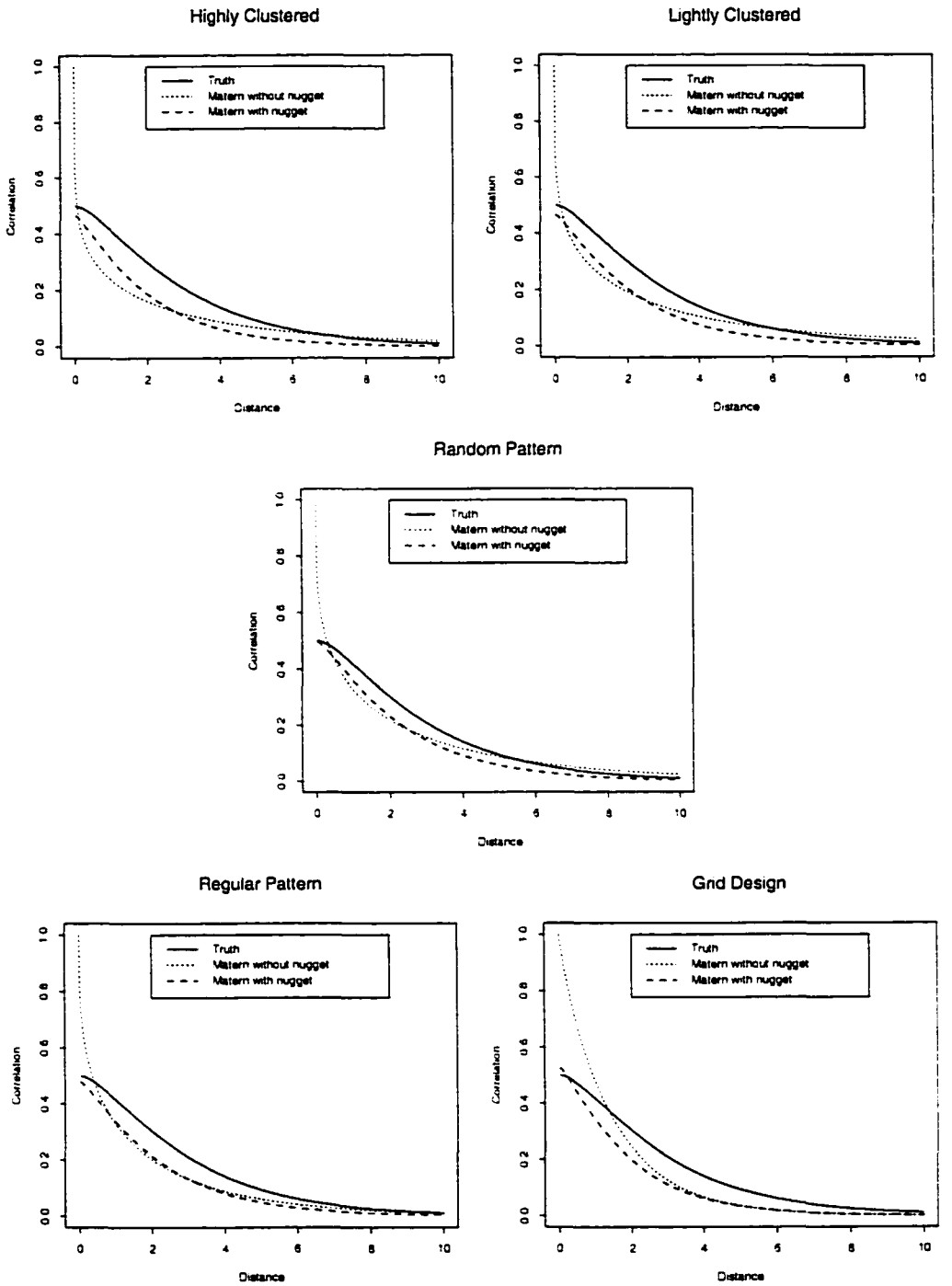

Figure 5.15: Comparison of Median Autocorrelation Functions for the Matern with Nugget and Matern without Nugget when $\theta_3 = 0.5$. The solid line is the true function, the dashed line is Matern with nugget, and the dotted line is Matern without nugget. The median autocorrelation function is based on the median values of θ_1, θ_2, and θ_3.

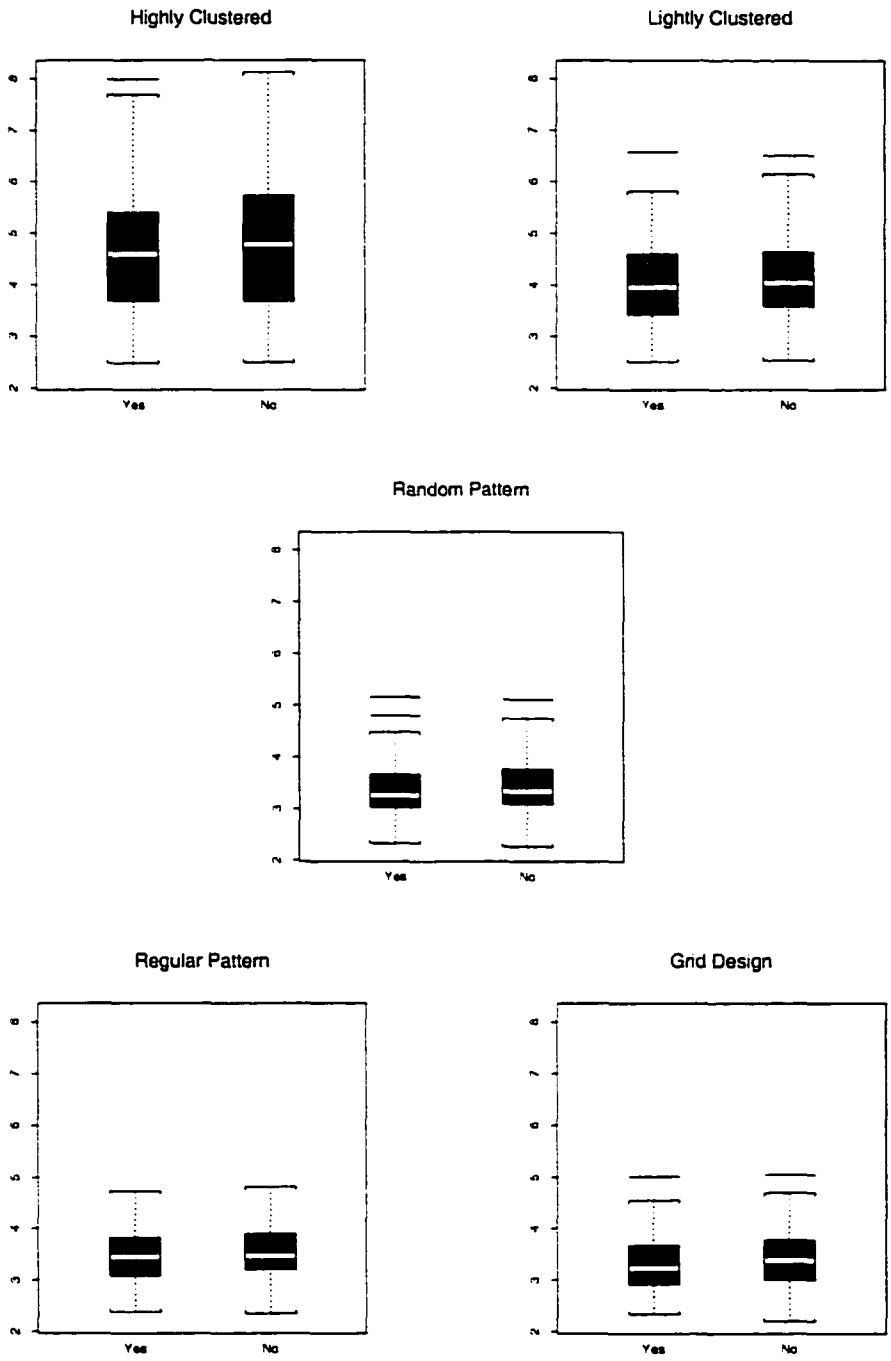

Figure 5.16: Comparison of MSPE for Matern Nugget and Matern no Nugget when the True Nugget is 0.20. Yes refers to $\theta_3 \geq 0$ and no refers to $\theta_3 = 0$.

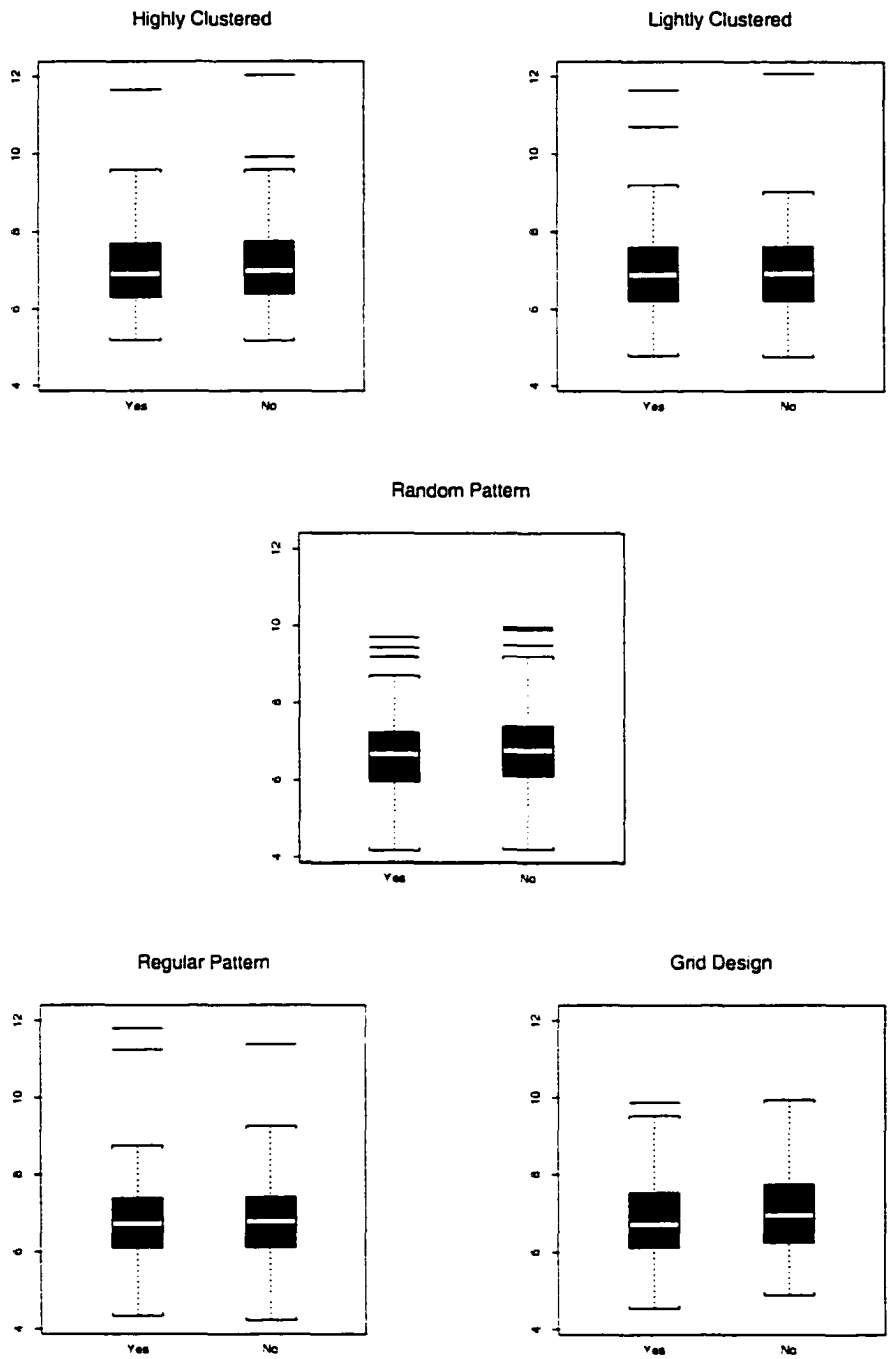

Figure 5.17: Comparison of MSPE for Matern Nugget and Matern no Nugget when the True Nugget is 0.50. Yes refers to $\theta_3 \geq 0$ and no refers to $\theta_3 = 0$.

not indicative of an effect from the pattern. The random pattern has a lower EPC than desired for both with nugget and without, although this again may be due to the simulation. For comparison, when the nugget is 50%, the simulations for the first three spatial patterns are at the nominal 95% level.

The regular and grid patterns display different behavior. Examining the plots of EPC with nugget plotted against EPC without nugget for the grid pattern (Figure 5.20), not including a nugget results in smaller EPC consistently for all replicates. This can also be seen for the regular pattern when the measurement error is 50%. Therefore, when there are few observations at small distances, it is more important to include a nugget term in order to have better predictive coverage.

Conclusion

It is surprising how similar the prediction results of including a nugget term versus not including a nugget term. Although not including a nugget term may result in very different estimates of the spatial parameters, the predictive ability are identical for many of the spatial patterns. It is only when a lack of information about the autocorrelation at small distances such as the regular and grid patterns when the inclusion of nugget is important for predictive coverage.

5.4.3 Sensitivity

I performed a sensitivity analysis to determine the impact of parameter estimation and prediction to changes in hyperparameters. Hoeting (1994, Section 5.1.3) reports on a sensitivity analysis on $\boldsymbol{\beta}$ and σ^2, the same normal-gamma set-up used in this thesis. I will not repeat this sensitivity analysis. I investigate each parameter in $\boldsymbol{\theta}$ individually by simulating ten data sets following the set-up in Section 5.1. When investigating the sensitivity on $\theta_i, i = 1, 2, 3$, the hyperparameters for the other two parameters are held constant.

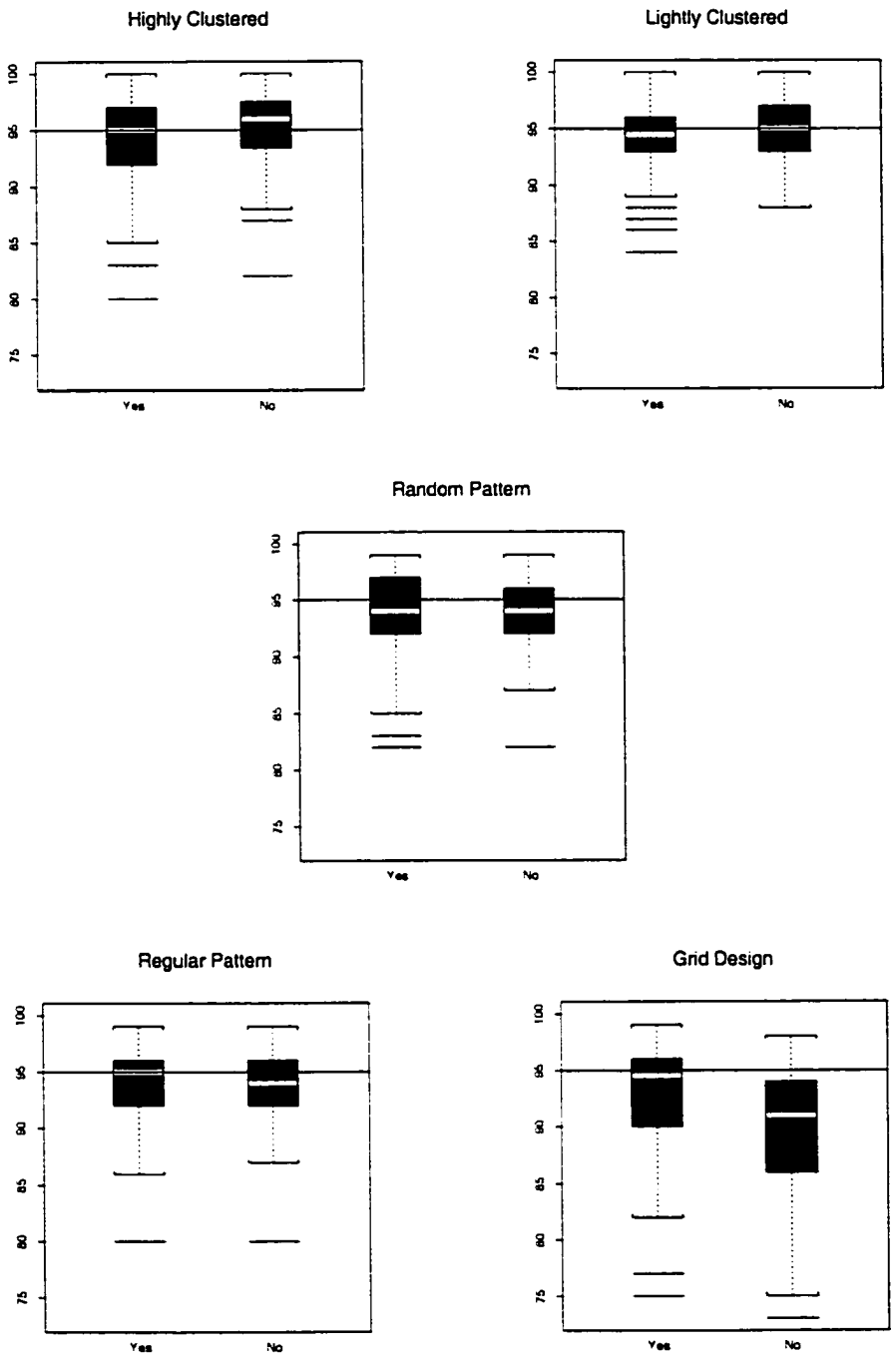

Figure 5.18: Comparison of EPC for Matern Nugget and Matern no Nugget when the True Nugget is 0.20. Yes refers to $\theta_3 \geq 0$ and no refers to $\theta_3 = 0$.

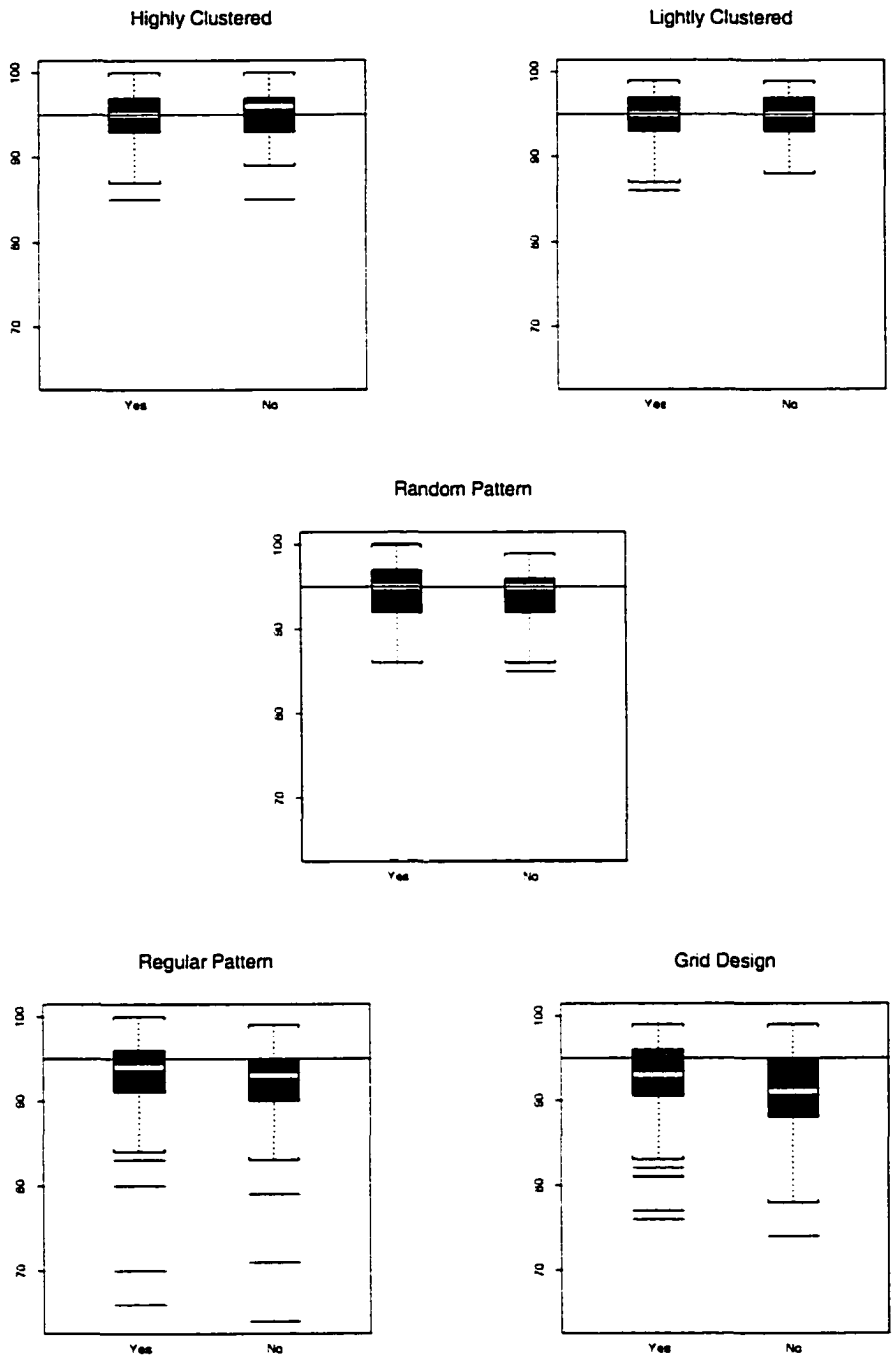

Figure 5.19: Comparison of EPC for Matern Nugget and Matern no Nugget when the True Nugget is 0.50. Yes refers to $\theta_3 \geq 0$ and no refers to $\theta_3 = 0$.

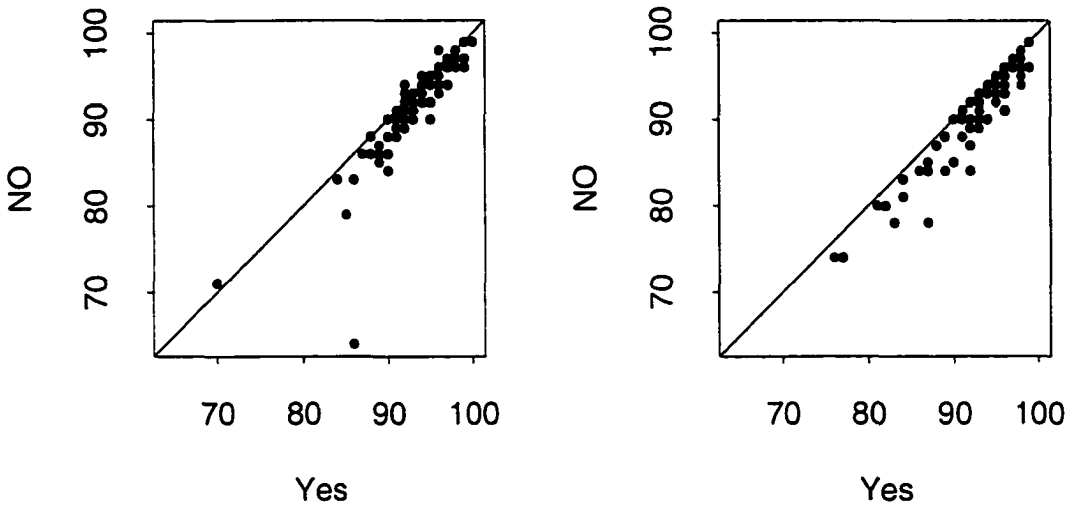

Figure 5.20: Comparison of EPC for Matern Nugget and Matern no Nugget when the True Nugget is 0.50 for the Regular and Grid Patterns. Yes refers to $\theta_3 \geq 0$ and no refers to $\theta_3 = 0$.

In investigating θ_1, I compare parameter estimates of $\boldsymbol{\theta}$ using four different sets of hyperparameters (Figure 5.21). These sets provide a variety of distributions that either place more weight on smaller values of θ_1 (Set 1) or weigh equally the variety of distances (Set 4).

The conclusions from this simulation are: first, different hyperparameter sets have some influence on estimates of θ_1. As the prior mean of θ_1 increases, the posterior mode also increases. Second, there is a slight impact on estimates of both θ_2 and θ_3. Third, there is variability in the PMP, but the impact on prediction is slight. Forth, there is little influence on MSPE and no impact on EPC or PC. In conclusion, the hyperparameters for θ_1 have very little impact, using a uniform distribution on θ_1 does not alter the predictive statistics.

In selecting hyperparameters for θ_2, the prior distribution should downweight large smoothness values. I examine three different hyperparameter sets (Figure 5.22). The results show hyperparameters on θ_2 impact estimation of θ_1 and

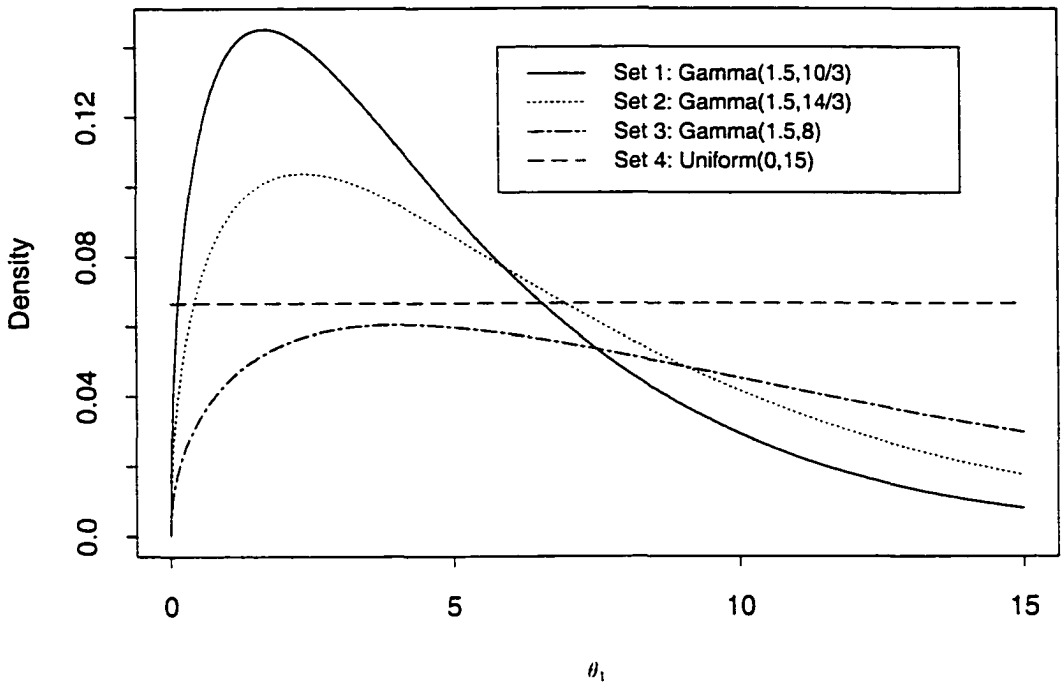

Figure 5.21: Prior Distributions for θ_1.

θ_2. In this simulation, as the prior mean of θ_2 decreases, the posterior mode of θ_2 also decreases but the posterior mode of θ_1 increases. This trade-off between estimates of θ_2 and θ_1 requires further investigation. The posterior mode of θ_3 does not seem to be affected by changes in the hyperparameters for θ_2. There are differences in PMP across the hyperparameter sets for θ_2 are smaller than for θ_1.

In terms of predictive statistics, the impact of changes from different sets of hyperparameters on θ_2 are slight. There is some small differences on the order of 0.1 between the different sets for MPSE. The measures of predictive coverage (EPC, meanPC, maxPC, minPC, and stdPC) are almost identical for the three hyperparameter sets.

The sensitivity analysis on θ_3 investigated how the posterior mode of $\boldsymbol{\theta}$ changes for different sets of hyperparameters (Figure 5.23). In order to determine the impact of hyperparameters on the posterior distribution, the data was simulated without a nugget term. These prior distributions were chosen because they represent a variety

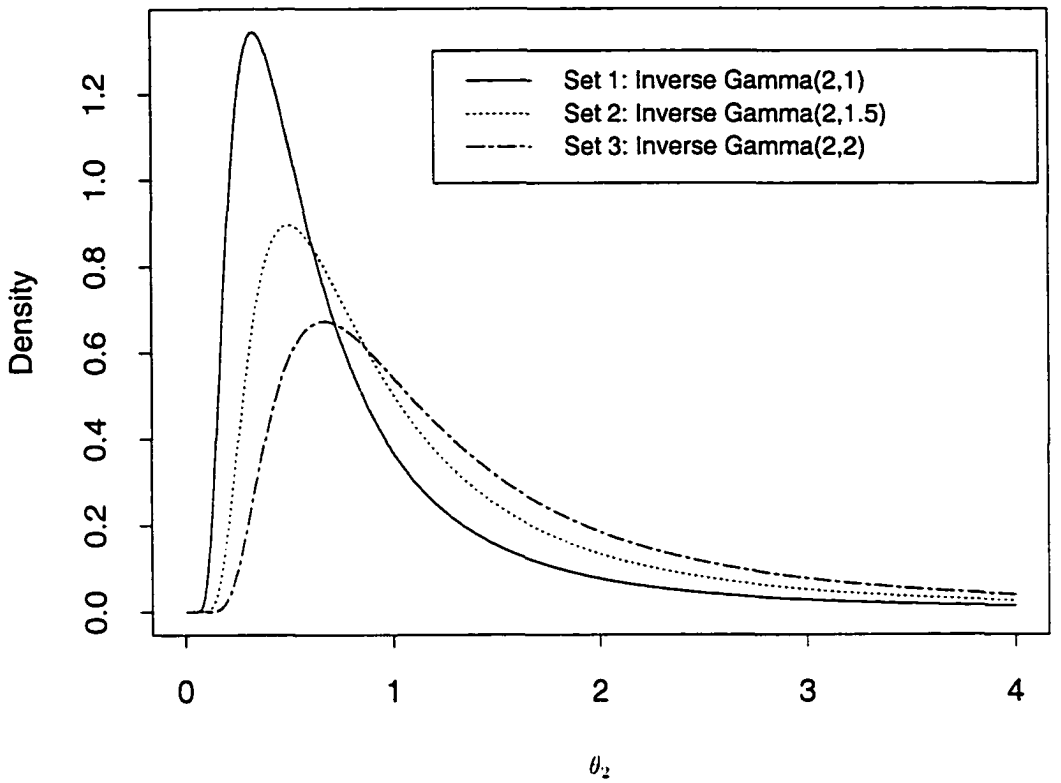

Figure 5.22: Prior Distributions for θ_2.

of beliefs about the lack of measurement error, from no opinion about nugget (Set 1) to a strong belief that $\theta_3 = 0$ (Set 2).

For all ten replications and the four hyperparameter sets, estimates of θ_3 are less than 5% for the model with the correct explanatory variables. There is quite a bit of variability in estimates of θ_1 and θ_2 for different hyperparameters sets for θ_3. For some of the replicates, there are slight differences between hyperparameter sets ($\theta_1 = 2.8$ for Set 2, $\theta_1 = 2.8$ for Set 4), while other replicates differer greatly ($\theta_1 = 3.6$ for Set 2, $\theta_1 = 2.0$ for Set 4).

Overall, parameter estimation for $\boldsymbol{\theta}$ is highly sensitive to selection of hyperparameters, although the impact on prediction statistics is slight. There seems to be significant dependence between the prior distributions on $\boldsymbol{\theta}$. This may indicate that the initial assumption of independence between spatial parameters is not valid, more research on this issue needs to be carried out.

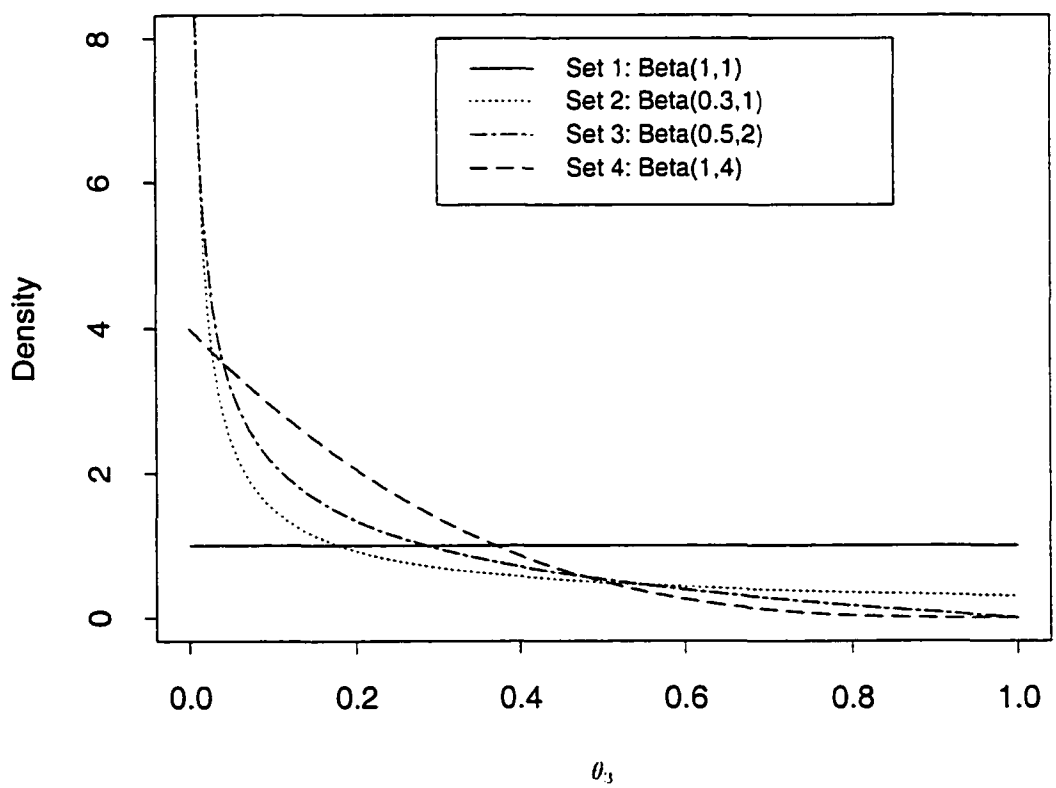

Figure 5.23: Prior Distributions for θ_3.

5.5 REML versus ML Estimation

The purpose of this simulation is to investigate the bias of ML and REML estimation. Zimmerman and Zimmerman (1991) investigated REML, ML and other estimation methods for bias for a model with constant mean, small sample size ($n = 16$ and $n = 36$), and autocorrelation structure either exponential or linear. As discussed in Section 2.3.4, their research has shown that the bias in REML estimation tends to be positive and larger than the bias in ML estimation which tends to be negative.

Comparing the profile ML equation (2.27) to the profile REML equation (2.29),

$$\begin{aligned} \ell_{profile}\left(\theta; \hat{\beta}, \hat{\sigma}^2, \boldsymbol{Z}\right) &= -\frac{n}{2}\log\left(\hat{\sigma}^2\right) - \frac{1}{2}\log|\boldsymbol{\Psi}| + c_1, \\ \ell_{REML}\left(\theta; \tilde{\sigma}^2, \boldsymbol{Z}\right) &= -\frac{n-q}{2}\log(\tilde{\sigma}^2) - \frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\log\left|\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}\right| + c_2, \end{aligned}$$

where c_1 and c_2 are constants, the differences arise in the coefficient of $\log(\tilde{\sigma}^2)$ and $\log\left|\boldsymbol{X}'\boldsymbol{\Psi}^{-1}\boldsymbol{X}\right|$. The differences between the two techniques are greatest when the number of columns of $\boldsymbol{X}$ is large. In Zimmerman and Zimmerman, the use of $\boldsymbol{X} = (1, \ldots, 1)'$ may reduce the differences between the techniques. This simulation study uses a design matrix, $\boldsymbol{X}$, that includes a mean and three explanatory variables (Section 5.1.2). Due to convergence problems discussed below, θ_2 was limited to be between 0 and 10.

5.5.1 Details

For each spatial point pattern, 500 independent replicates of $\boldsymbol{Z}$ were generated from the model $\boldsymbol{Z} = 2 + 0.75\boldsymbol{X}_1 + 0.50\boldsymbol{X}_2 + 0.25\boldsymbol{X}_3 + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ follows the Matern autocorrelation function with $\theta_1 = 4$ and $\theta_2 = 1$. For each replicate, locations were generated following the steps in Section 5.1.1. For the mean function with explanatory variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$, I used REML and ML techniques to estimate θ_1 and θ_2 assuming that $\theta_3 = 0$, the no nugget model. I also used REML and ML techniques to estimate θ_1, θ_2, and θ_3 to the same data.

Table 5.14: Number of Simulations and Large Estimates of θ_2 when θ_3 is Non Zero. n_c is the number of replicates which converged, while n_l is the number of replicates where $\theta_2 > 10$.

Estimation Method	Highly Clustered		Lightly Clustered		Random Pattern		Regular Pattern		Grid Design	
	n_c	n_l	n_c	n_l	n_c	n_l	n_c	n_l	n_c	n_l
ML	390	2	496	1	500	1	500	5	500	70
REML	396	1	498	2	500	3	500	7	500	49

5.5.2 Results

In this section, I summarize the results of the simulation comparing a) REML and ML estimation, b) including a nugget term (θ_3) versus no nugget term ($\theta_3 = 0$), and c) how REML and ML estimation depends on the sampling pattern. First, I discuss how the sampling pattern affects the convergence of estimation. Second, I focus on the average estimates for each sampling pattern, comparing and contrasting the different estimation techniques. Third, I focus on each sampling pattern and comment on the bias and standard error of each parameter.

Convergence Issues

For the REML versus ML comparisons there were some convergence difficulties for the simulations when θ_3 was allowed to be nonzero. In some cases the REML estimation did not converge, while for others ML did not converge. For all non-convergence cases, I did not include the pair (REML,ML) of observations in the results shown in Figure 5.24 and Tables 5.16 through 5.19.

Table 5.14 lists the number of replicates that converged as well as the number of replicates with $\theta_2 \geq 10$, the boundary value for smoothness for these simulations. When θ_2 is allowed to be larger than 10, some estimates θ_3 were large (0.25). The maximum value of 10 was selected because the differences in the autocorrelation curves between $\theta_2 = 10$ and $\theta_2 = 100$ are slight. Replicates with $\theta_2 \geq 10$ were not included in the analysis. The highly clustered spatial pattern has the largest

number of convergence problems in the simulation, with a few failed replicates with the lightly clustered pattern. The highly clustered pattern is unique when compared to the other patterns in that it can have a large number of observations at very small distances. (See Figure 5.3 for one example). The observations within a cluster are all highly correlated and may not include much independent information (Cressie *et al.*, 1990).

Additionally, the grid design has the largest number of simulations where $\theta_2 \geq 10$. The grid design provides no information about the autocorrelation structure at distances less than one unit. Therefore, it makes sense that accurate estimation of θ_2 is difficult for samples with a grid design.

In terms of convergence and large values of θ_2 when $\theta_3 \geq 0$, the results in Table 5.14 indicate similar convergence behavior for REML and ML estimation except for the grid design. REML estimation for the grid design has fewer cases where $\theta_2 \geq 10$. Again, the grid design has no observations at distances less than one unit apart, although I am unable to explain the difference in convergence behavior between REML and ML for the grid design.

Comparison of Average Autocorrelation Curves

Figure 5.24 demonstrates the average autocorrelation curves for all five sampling patterns. The solid line is the true autocorrelation curve, Matern with $\theta_1 = 4$ and $\theta_2 = 1$. The dotted line is the ML estimate with nugget ($\theta_3 \geq 0$), while the short dashed line represents the REML estimate nugget($\theta_3 \geq 0$). The ML estimate without nugget ($\theta_3 = 0$) is the long dashed line, and REML estimate without nugget ($\theta_3 = 0$) is the dash-dot line.

For each plot, there are two areas of concern. First, I focus on the behavior at the origin. The grid design is the only pattern which includes a difference in curves near the origin between with nugget and without nugget. Including a nugget

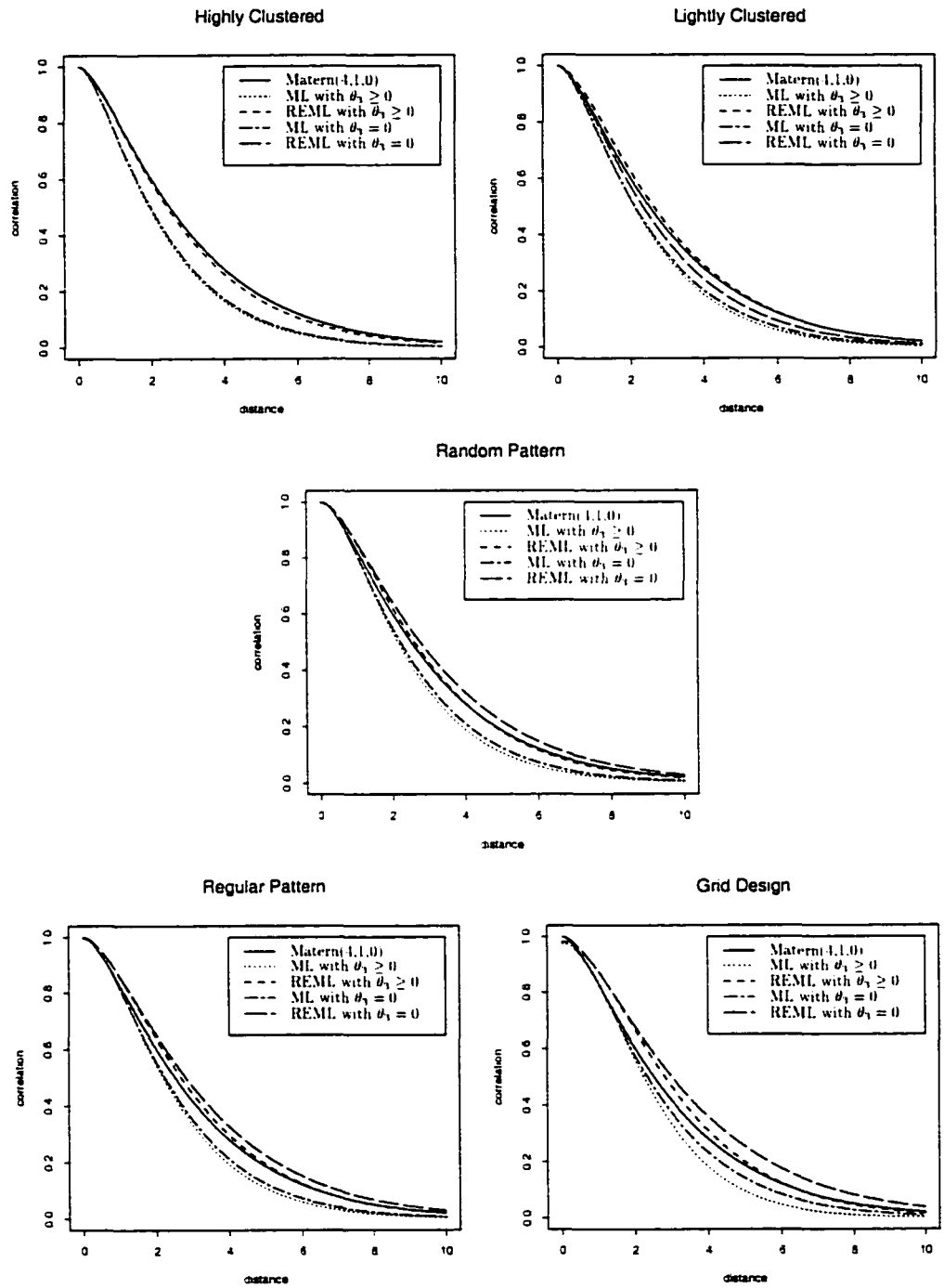

Figure 5.24: Average Autocorrelation Curves for Five Sampling Patterns. The solid line is the true autocorrelation function (Matern(4.1)). The other lines correspond to the other methods of estimation. The parameters can be determined from the bias tables (Tables 5.16, 5.17, 5.15, 5.18, and 5.19).

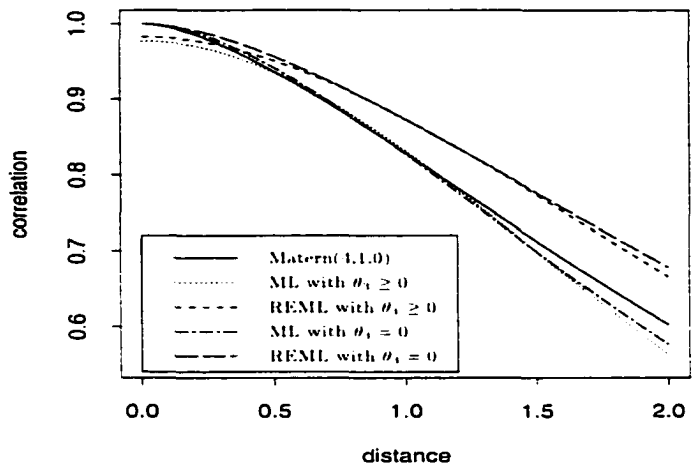

Figure 5.25: Average Autocorrelation Curve for the Grid Pattern at Small Distances.

term results in estimates of θ_3 that alter the curve (Figure 5.25). The grid design does not have any observations closer than one unit apart, although I will want to predict the random process at locations at smaller distances. In Sections 5.3 and 5.4, I investigated how the spatial sampling pattern can effect prediction.

Second, how well does each estimation method model the range of the function? Based on my simulations, in the highly clustered design, it is clear that REML methods do better than ML methods. The lightly clustered, random pattern and regular pattern results are similar. ML methods underestimate the range and therefore underestimate the autocorrelation function. REML methods are closer to the true model, and do not seem to be consistently under or over estimating based on these three sampling patterns. The grid design is different from the first four patterns, REML with $\theta_3 = 0$ overestimates the range (Figure 5.24 grid design long dashed line), while REML with $\theta_3 \geq 0$ estimates overestimate the autocorrelation at distances 0.5 to 5.0. ML estimation underestimates the autocorrelation with ML for $\theta_3 \geq 0$ doing worse than when $\theta_3 = 0$.

Table 5.15: Bias results for Random Pattern. The first two rows allow for nonzero θ_3, while rows three and four require $\theta_3 = 0$.

Estimation	θ_1		θ_2		θ_3		Replicates
Method	Bias	SE	Bias	SE	Bias	SE	Converged
ML	-0.800	0.051	0.350	0.027	0.0017	0.00021	496
REML	-0.058	0.084	0.250	0.025	0.0022	0.00022	496
ML	-0.630	0.056	0.210	0.021	NA	NA	500
REML	0.320	0.095	0.072	0.012	NA	NA	500

In conclusion, REML estimation does better than ML estimation in terms of average autocorrelation curves. Including θ_3 results in better estimates of the autocorrelation curve even when the data were simulated with $\theta_3 = 0$. This is most important for the grid design where there may be no information at small distances, the added flexibility of $\theta_3 \geq 0$ lets the estimate match the truth. The next section focuses on bias of the parameter estimates, followed by the bias in each individual sampling pattern.

Bias results for the Random Pattern

For the random pattern, ML estimation is less biased when $\theta_3 = 0$ compared to $\theta_3 \geq 0$ (Table 5.15). Overall, ML estimates of θ_1 have a negative bias, resulting in underestimation of the autocorrelation curve. Second, ML estimation of θ_1 with $\theta_3 \geq 0$ has a larger absolute bias when compared to ML estimation with $\theta_3 = 0$. Estimates of θ_2 have a positive bias.

For REML estimation there is a trade-off between bias of θ_1 and θ_2. Using $\theta_3 = 0$ results in a larger positive bias on θ_1, it also results in a smaller bias for θ_2. In support of REML with $\theta_3 \geq 0$, the visible difference in autocorrelation curves indicate that the bias in θ_2 is less important than the bias in θ_1. The trade-off between bias of θ_1 and θ_2 needs to be studied further to determine which is more important for estimation and prediction. Overall the REML estimates are less biased than ML estimates for θ_1 and θ_2. The ML estimates are less biased for θ_3. In future studies this should be investigated further for simulations where $\theta_3 \geq 0$.

Table 5.16: Bias results for Highly Clustered Design. The first two rows allow for nonzero θ_3, while rows three and four require $\theta_3 = 0$.

Estimation Method	θ_1		θ_2		θ_3		Replicates Converged
	Bias	SE	Bias	SE	Bias	SE	
ML	-0.970	0.058	0.110	0.0086	0.000091	0.000019	381
REML	-0.180	0.096	0.069	0.0110	0.000200	0.000029	381
ML	-0.900	0.055	0.084	0.0063	NA	NA	500
REML	0.014	0.090	0.014	0.0057	NA	NA	500

Table 5.17: Bias results for Lightly Clustered Design. The first two rows allow for nonzero θ_3, while rows three and four require $\theta_3 = 0$.

Estimation Method	θ_1		θ_2		θ_3		Replicates Converged
	Bias	SE	Bias	SE	Bias	SE	
ML	-0.810	0.056	0.240	0.0230	0.00065	0.000087	494
REML	0.045	0.095	0.160	0.0190	0.00081	0.000089	494
ML	-0.670	0.058	0.120	0.0079	NA	NA	500
REML	0.360	0.098	0.035	0.0068	NA	NA	500

Bias results for the other sampling patterns

The simulation results for the highly clustered design are shown in Table 5.16. The number of replicates that converged differ greatly between including θ_3 and fixing $\theta_3 = 0$. More than any other information, this supports restricting $\theta_3 = 0$ or estimating using both cases for θ_3.

The results from both the lightly clustered pattern (Table 5.17) and the regular pattern (Table 5.18) are practically identical to the random pattern. The grid design differs from the other spatial patterns in the size of the bias for REML estimated

Table 5.18: Bias results for Regular Pattern. The first two rows allow for nonzero θ_3, while rows three and four require $\theta_3 = 0$.

Estimation Method	θ_1		θ_2		θ_3		Replicates Converged
	Bias	SE	Bias	SE	Bias	SE	
ML	-0.770	0.053	0.440	0.028	0.0037	0.00043	492
REML	0.056	0.091	0.280	0.025	0.0043	0.00042	492
ML	-0.620	0.056	0.260	0.017	NA	NA	500
REML	0.380	0.100	0.093	0.013	NA	NA	500

Table 5.19: Bias results for Grid Design. The first two rows allow for nonzero θ_3. while rows three and four require $\theta_3 = 0$.

Estimation	θ_1		θ_2		θ_3		Replicates
Method	Bias	SE	Bias	SE	Bias	SE	Converged
ML	-0.860	0.056	1.200	0.077	0.023	0.0012	430
REML	0.084	0.100	0.710	0.049	0.017	0.0010	430
ML	-0.470	0.064	0.250	0.023	NA	NA	500
REML	0.670	0.110	0.099	0.018	NA	NA	500

θ_1 when $\theta_3 = 0$ is larger than the ML estimated bias when $\theta_3 = 0$. although REML bias is positive versus negative ML bias. The grid design also has a much larger problem with large estimates of θ_2 (Table 5.14). If these simulations where $\theta_2 = 10$. the boundary value. were included in the summary results. the bias for θ_2 would be larger for both REML and ML estimation when $\theta_3 \geq 0$ suggesting that requiring $\theta_3 = 0$ may be very important.

5.5.3 Conclusion

Summarizing the results of this simulation. REML estimation produces estimates with smaller bias as compared to than ML estimation. For these simulations. the differences between convergence of ML versus REML were small. When there is no information about the presence of a nugget in the data. it may be useful to include θ_3 because it allows more flexibility in the autocorrelation curves and generally results in less bias for θ_1. On the other hand. including θ_3 can cause problems in convergence depending on the spatial pattern.

Chapter 6

REAL DATA EXAMPLES

This chapter examines the application of spatial BMA using the Matern autocorrelation function to two data sets. The first is Davis's (1986) topographic data set which has been widely analyzed in many spatial applications. The second example is the Whiptail Lizard data set from Hollander et al. (1994).

6.1 Topographic Data Set

In this section I analyze the Davis's topological elevation data (1986). This data set has been analyzed in a variety of papers (Ripley, 1981; Warnes and Ripley, 1987; Mardia and Watkins, 1989; Handcock and Stein, 1993) as a standard spatial data set. Methods used to analyze this data include universal kriging (Ripley, 1981) and Bayesian kriging (Handcock and Stein, 1993). I analyze the data using BMS with the Matern autocorrelation function to determine which model best describes the data.

This data set consists of 52 locations originally surveyed for a class assignment. Each observation consists of the north-south and east-west coordinates, and the elevation at each location (the response). Handcock and Stein (1993) incorporate an additional explanatory variable, distance to the nearest stream. The locations are plotted in Figure 6.1 and the eight possible models are listed in Table 6.1.

Assuming a stationary isotropic autocorrelation function, I adopt the Matern autocorrelation function described in Section 2.2.2. Furthermore, I assume the data

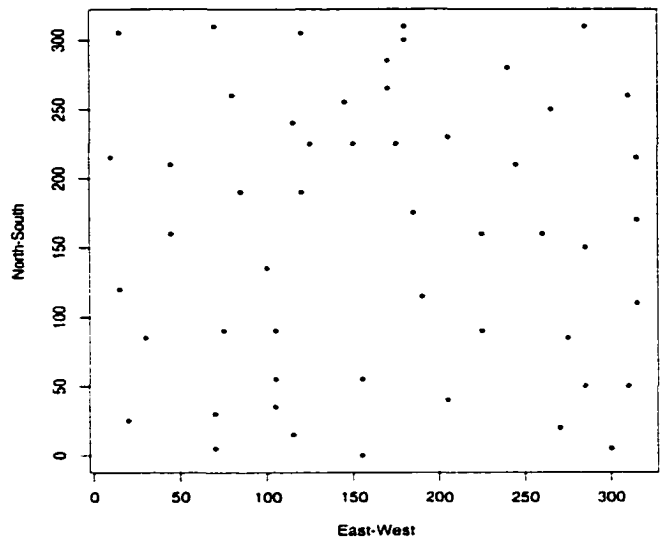

Figure 6.1: Locations for the Topographic Data Set.

Table 6.1: Models for Topographic Data Set. The three explanatory variables are listed in columns with • representing the inclusion of a variable in that model.

Model	East-West	North-South	Stream Distance	PMP (%)
M_1				< 0.1
M_2	•			< 0.1
M_3		•		< 0.1
M_4	•	•		0.2
M_5			•	< 0.1
M_6		•	•	99.7
M_7	•		•	< 0.1
M_8	•	•	•	< 0.1

can be modeled using a Gaussian random field. Handcock (1989) investigated this data set and found these assumptions to be reasonable.

6.1.1 Prior Distributions

The prior distributions on the parameters follow the description in Section 3.1. using proper priors on $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$. In order to simplify hyperparameter specification, the explanatory variables are standardized to have a mean of zero and variance

1. Additionally, the response variable (elevation) is rescaled to have variance equal to 1. $Z_i^* = Z_i/s_Z$, where s_Z is the standard deviation of the elevation.

The prior distribution on β is Gaussian with hyperparameter $\nu = 2.85$. The hyperparameters for σ^2 are $\nu = 2.58$ and $\lambda = 0.28$. The prior distribution on θ_1 is uniform over the interval $(0, d_{max})$ where $d_{max} = 414$. The prior distribution on θ_2 is also uniform over the interval $(0, 10)$. Previous analysis of these data assumes that measurement error of the process is small, therefore no nugget is included here (Handcock, 1989, p.161).

Handcock and Stein (1993) use a different set of prior distributions when modeling the data via Bayesian kriging with the Matern autocorrelation function. They include two explanatory variables, north-south coordinate and stream distance. They use the priors

$$f(\beta, \sigma^2, \theta) \propto \frac{f(\theta)}{\sigma^2}. \tag{6.1}$$

This corresponds to Jeffrey's prior on both σ^2 and β. Additionally, they use uniform priors on both θ_1 and θ_2. Section 3.1.3 addressed some of the difficulties with these priors on σ^2 and β and the resulting improper posterior distributions.

6.1.2 Results

Following the methodology in Chapters 3 and 4, I estimate the spatial parameters and PMP. Because this data set includes only 52 observations, I do not split the data into an estimation set and a prediction set. Instead, I focus on BMS (Section 4.5). As seen in Table 6.1, the model which includes both the north-south coordinate and stream distance accounts for almost all of the PMP (99.7%). Therefore, no significant weight is placed on the other models. In this situation, the PMP values are useful for specifying the lack of model uncertainty based on this set of models while still incorporating the uncertainty in parameter estimates.

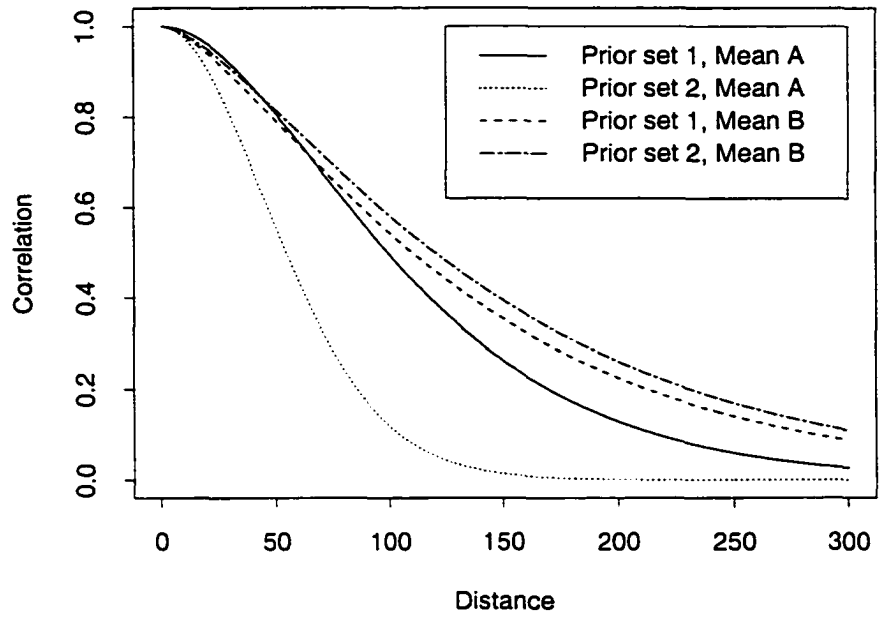

Figure 6.2: Autocorrelation Curves for the Topographic Data Set. Mean A corresponds to including North-South direction and the Stream Distance. Mean B is a constant mean. Prior set 1 consists of the conjugate priors described in Section 6.1.1 while prior set 2 consists of the prior distributions used in Handcock and Stein (1993).

The prior structure of the parameters in Bayesian kriging can have a dramatic effect on parameter estimation. Figure 6.2 plots autocorrelation curves based on fitted models with two mean functions, constant mean and the best model, and two prior distributions. The two prior distributions are discussed in Section 6.1.1, the Jeffrey's priors used by Handcock and Stein (1993) versus the proper conjugate priors on β and σ^2 discussed in this thesis. The priors on θ are identical. Comparisons between prior distributions are useful for investigating sensitivity.

Figure 6.2 demonstrates the effect the prior distributions on β and σ^2 can have on estimates of the spatial parameters. When a constant mean function is used, the difference between the priors is slight, both have similar mean and smoothness values. The differences when using a mean that includes both North-South coordinate and stream distance is more dramatic. Jeffrey's prior results in a much smaller θ_1 and a much larger smoothness value. Since the data are not simulated there is no

true autocorrelation curve to compare with the estimated curves. Instead, other methods such as cross validation can be used to determine which curve is more appropriate.

BMS indicates strong support for the model with explanatory variables North-South coordinate and stream distance (99.7% PMP). In this case BMA would not be useful because there is little model uncertainty. Second, the prior distributions on the parameters can have a large impact on the autocorrelation curves. More importantly, the prior distributions on β and σ^2 impact the estimates of the spatial parameters, θ.

6.2 Lizard Data Set

Previously analyzed by Hollander et al. (1994) and Ver Hoef et al. (1999). I reanalyze abundance data for the orange-throated whiptail lizard in southern California. A total of 256 locations at 21 sites were used for trapping. Each observation consists of the average lizards caught per day at each location. Following the methodology of Ver Hoef et al. (1999), I analyze lizard abundance, ignoring the locations where no lizards were observed. There are a total of 148 observations for the abundance analysis, one additional observation was removed as an outlier following Ver Hoef et al. (1999). Additionally, a log transformation is used on the response, average lizards caught per day, to allow for the use of a Gaussian random field.

The data set includes 37 explanatory variables along with the locations of each observation. Figure 6.3 shows the locations of the 148 observations. The sampling locations are highly clustered. Modeling difficulties associated with highly clustered data are discussed in Chapter 5.

The explanatory variables include information on vegetation layers, vegetation types, topographic position, soil types, and abundance of ants. This thesis focuses

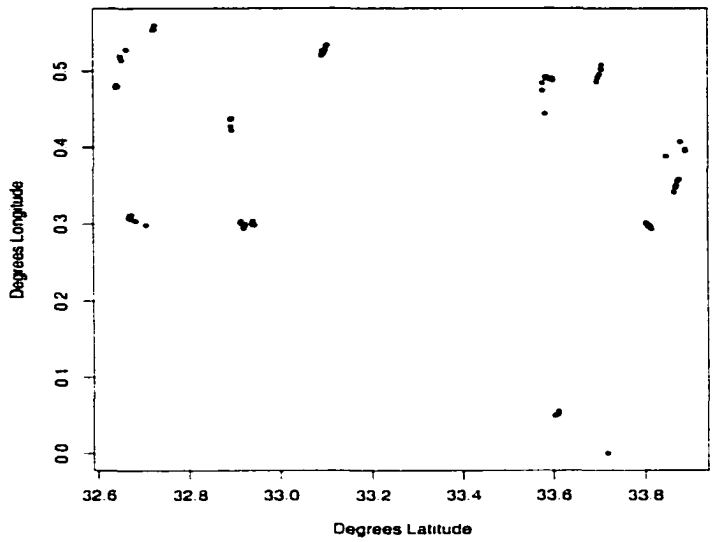

Figure 6.3: Locations for the Lizard Data Set.

Table 6.2: Lizard example: Summary Statistics.

	Explanatory Variable	Mean	S.D.	Min	Max
X_1	Crematogaster ant abundance	1.70	0.81	1.00	3.00
X_2	Log percent sandy soils	1.10	0.28	0.30	1.90
X_3	Elevation	270	200	12	590
X_4	Barerock Indicator	0.39	0.49	0.00	1.00
X_5	% Cover	82	11	18	99
X_6	Log Percent chapparal plants	0.53	0.67	0.00	1.90

on methodology for fitting all possible models for a given set of explanatory variables. For the lizard data, this corresponds to fitting 2^{37} or $1.37 * 10^{11}$ total models. Therefore the number of explanatory variables was reduced. Ignoring the spatial correlation, I used a stepwise procedure to select the best model. This model included seven explanatory variables, six of which were significant at the 0.10 level.

I restrict the set of possible explanatory variables to be the six with significance at 0.10. Following the results in Section 5.2, there are problems with ignoring the spatial autocorrelation when choosing explanatory variables. Incorporation of all of the explanatory variables may lead to more model uncertainty, but this requires research into the methods for exploring the model space for spatial models.

This is discussed further in Chapter 7. These six explanatory variables include the abundance of Crematogaster ants, elevation, log of percent sandy soils, indicator of bare rock, percent cover, and log percent chapparal plants (Table 6.2). Using six explanatory variables, there are a total of 64 models to be considered.

6.2.1 Prior Distributions

In order to simplify prior specification, the six explanatory variables are standardized to have zero mean and variance one. The response, log average counts, is also standardized to have variance one. The hyperparameters for the lizard data are identical to the hyperparameters for the topographic data described in Section 6.1.1.

This data set includes significant measurement error so θ_3 is included in the model. The prior distribution on θ_3 is inverse gamma with hyperparameters $\nu = 2$ and $\lambda = 1.5$. This distribution has a mean of 1.5 and an infinite variance. As discovered in Section 5.3, when a nugget is included in the analysis, θ_2 needs a prior distribution with more weight on smaller values of θ_2 to reduce the probability of θ_2 converging to ∞.

6.2.2 Results

The analysis of the lizard data consists of two parts. First, I analyze the entire data set. I compare BMA with the Matern autocorrelation function to modeling the data using the spherical autocorrelation function with two explanatory variables. Second, I perform a cross validation analysis, removing each observation one at a time and refitting the BMA analysis.

Full Model

The results from parameter and PMP estimation show there is some model uncertainty in the data. The best model accounts for 61.4% of the model probability and includes explanatory variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The next four models ranked by

Table 6.3: Top Five Models for the Lizard Data Set by PMP (%).

X_1	X_2	X_3	X_4	X_5	X_6	PMP
•	•					61.4
	•					10.3
•	•	•				8.6
•	•			•		4.9
•	•				•	3.3

PMP build off of this first model, either not including X_1 or including one additional explanatory variable. This gives a strong indication that including X_2 in the model is the most important, followed by X_1. Over all models, first two explanatory variables account for 83.9% of the PMP. This does not indicate much model uncertainty.

Cross Validation

To do cross validation, I remove one observation at a time and repeat the entire procedure, including parameter and PMP estimation, followed by the prediction of the removed observation. This allows us to calculate prediction diagnostics described in Section 5.1.3. Ver Hoef et al. (1999) also performed a cross validation study of the data using the spherical autocorrelation function for the model with X_1 and X_2.

There are some difficulties with cross validation on spatial data. First, the observations are typically correlated, the removal of one observation does not have the same impact as removal of one independent observation. Second, this data set is very highly clustered (Figure 6.3). Therefore there is a large amount of information about observations at small distances yet there is considerable measurement error. The estimated nugget (θ_3) is at least 36%. In view of these issues, the results in this section should be considered with care.

I compare BMA to the model adopted by Ver Hoef et al. (1999). The model is determined as follows. The residuals from fitting the full model with six explanatory

Table 6.4: Prediction statistics for Cross Validated Lizard Data

Method	Explanatory Variables	EPC	MSPE
Independent	$X_1, X_2, X_3, X_4, X_5, X_6$	97.3	0.942
Spherical	X_1, X_2	93.2	0.649
BMA	(See Table 6.3)	93.9	0.677

variables are used to estimate the spherical autocorrelation function parameters. Based on this fit of the full model, only the first two explanatory variables are significant at the 0.05 level. Using only the first two explanatory variables, the model is refit and the spatial parameters reestimated. For the cross validation, the set of explanatory variables are assumed fixed, with the removal of each observation, the parameters are reestimated and a prediction interval is computed for the removed observation.

The results for BMA, the spherical autocorrelation model, and a model assuming independent errors are summarized in Table 6.4. In terms of MSPE, the spatial models do much better than the independent model, the spherical model has a smaller MSPE than the BMA model. In terms of EPC, the two spatial models are very similar and slightly underestimate the 95% level, while the independence model overestimates the prediction interval level.

The similarity of the spatial models may be due to a variety of factors. First, there are some concerns about cross validation for spatial data in general, and highly clustered data in particular. Second, there is little model uncertainty in the dataset. The PMP values for cross validation show very little variability. Third, there is a large amount of nugget for all of the models in BMA which limits the overall effectiveness of spatial models.

Chapter 7

CONCLUSION AND FURTHER RESEARCH

7.1 Discussion

Although spatial data analysis is an active area of research, little work has been done on methods for model selection. The impact of model selection techniques on spatial models can be costly in terms of the accuracy of prediction. In this thesis, I explore spatial analysis and propose two methods that improve upon available model selection techniques. First, I propose simultaneous model selection which estimates the spatial parameters while selecting the explanatory variables. Second, I propose methodology for Bayesian model averaging with spatial models. Together, these two methods serve as tools that may offer improvements in the predictive performance of spatial methods.

Simulation results presented here show the improvement of simultaneous model selection over the selection of explanatory variables assuming independent observations in both predictive coverage and mean square predictive error. Simulation results also demonstrate the improvement of Bayesian model averaging over other methods in terms of predictive coverage.

7.2 Future Research

The following sections describe extensions of this thesis that I would like to pursue.

BMA over Autocorrelation Functions

In this thesis, I focus on using the Matern autocorrelation function to model the autocorrelation structure of the data. Other autocorrelation functions such as the spherical and exponential have been used to model spatial data with good results. BMA can be expanded to incorporate other autocorrelation functions, resulting in a methodology that does not depend on the spatial functions. Additionally, BMA can average over both isotropic and anisotropic, stationary and potentially non-stationary autocorrelation functions.

BMA over Different Model Families and Transformations

BMA can be used in a wide variety of situations. As seen in the lizard data set, some situations require transformations of the response in order to meet the assumptions of a Gaussian random field. There are two methods for incorporating transformations into BMA. First, transformations can be included into the model space, so that one model may use the square root transformation of a predictor while another model may use the log transformation of that same predictor. Second, using the Box-Cox class of transformations of the response, a transformation parameter can be estimated simultaneously with the other parameters of the models so that each model may use a different transformation. This method was used by Hoeting (1994) for BMA for linear regression models. Even with the use of transformations, some spatial data sets do not meet the Gaussian assumptions. Expanding on the Gaussian linear models, BMA can be used to model average over different model classes such as logistic regression models.

Improvement of Integral Approximations

With the expansion of the model space described in the previous two sections, integral approximations of (3.15) and (4.3), the posterior predictive distribution

and the integrated likelihood are more important. When direct calculations are not possible due to the prior structure as described in this thesis, the MLE, BIC, and Laplace approximations form a starting point (Sections 3.2.3 and 4.4.3). Further work in this area should focus on finding faster and more accurate methods of approximation when necessary. The addition of other model classes as described above may require more approximations in calculation of the posterior predictive distribution and the integrated likelihood.

BMA with Large Model Space

Methods need to be developed for averaging over large model spaces when the data are correlated. Section 4.3.2 discussed some of the techniques used for large model spaces in generalized linear models. The interaction between explanatory variables and autocorrelation structure is an additional challenge that can not be ignored in reducing the sample space. Extensions to the leaps and bounds algorithm (Volinsky, 1997) may offer one research direction.

Prior Distributions

Research is needed into the specification and sensitivity of prior distributions and hyperparameters for Bayesian kriging and spatial BMA. In this thesis I used conjugate prior distributions to simplify the integrals in the analysis and rescaled the explanatory variables to ease hyperparameter selection.

The prior distributions of the parameters are assumed independent in this thesis. Work needs to be done to explore the potential dependence between autocorrelation parameters as well as dependence with the variance of the process and the regression parameters. There is some evidence that the hyperparameters, and therefore the prior distributions of $\boldsymbol{\beta}$ and σ^2 affect the estimation of $\boldsymbol{\theta}$. I would like to explore the interaction in prior distributions in future work.

Computer Software

I would like to develop computer software that would simplify the use of BMA for spatial data. The package would allow for specification of prior distributions and autocorrelation family selection. Most importantly the software would include graphic representations of prediction. as well as methods for assessing how different models fit the data.

Bibliography

Abramowitz, M. and Stegun, I. A., editors (1965). *Handbook of Mathematical Functions*. Dover: New York.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrox, B. and Caski, F., editors, *Second International Symposium on Information Theory*, page 267.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forcasts. *Operational Research Quarterly*, 20:451–468.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag: New York, second edition.

Berger, J. O., Oliveira, V. D., and Sanso, B. (2000). Objective Bayesian analysis of spatially correlated data. Technical Report 00-12, Institute of Statistics and decision Sciences, Duke University.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer-Verlag: New York, second edition.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference A Practical Information Theoretic Approach*. Springer: New York.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, 158:419–466.

Clyde, M. (1999). Comment on 'Bayesian model averaging: A tutorial'. *Statistical Science*, 14:401–404.

Cochran, W. G. (1937). Problems in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society (suppliment)*, 4:102–118.

Cressie, N., Gotway, C. A., and Grondona, M. O. (1990). Spatial prediction from networks. *Chemometrics and Intelligent Laboratory Systems*, 7:251–271.

Cressie, N. and Reed, T. R. C. (1989). Spatial data analysis of regional counts. *Biometrics Journal*, 31:699–719.

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* Wiley: New York.

Cressie, N. A. C. and Grondona, M. O. (1992). A comparison of variogram estimation with covariogram estimation. In Mardia, K. V., editor, *The Art of Statistical Science*, pages 191–208. Wiley: New York.

Davis, J. C. (1986). *Statistics and Data Analysis in Geology.* Wiley: New York. second edition.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, 57:45–97.

Draper, D. (1999). Comment on 'Bayesian model averaging: A tutorial'. *Statistical Science*, 14:405–409.

Ecker, M. D. and Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological, and Environmental Statistics*, 2:347–369.

Ecker, M. D. and Gelfand, A. E. (1999). Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology*, 31:67–83.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, 16:499–511.

George, E. I. and McCulloch, R. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

Griffith, D. A. and Layne, L. J. (1999). *A Casebook for Spatial Statistical Data Analysis.* Oxford University Press: New York.

Haining, R. (1990). *Spatial data analysis in the social and enviromental sciences.* Cambridge University Press: Cambridge.

Handcock, M. S. (1989). *Inference for Spatial Gaussian Random Fields When the Objective is Prediction.* PhD thesis, University of Chicago.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics*, 35:403–410.

Handcock, M. S. and Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meterological fields. *Journal of the American Statistical Association*, 89:368–390.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61:383–385.

Hoeting, J. A. (1994). *Accounting for Model Uncertainty in Linear Regression*. PhD thesis, University of Washington.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial with discussion. *Statistical Science*, 14:382–417.

Hollander, A. D., Davis, F. W., and Stoms, D. M. (1994). Hierarchical representations of species distributions using maps, images and signting data. In Miller, R. I., editor, *Mapping the diversity of Nature*. Chapman and Hall: London.

Huijbregts, C. J. and Matheron, G. (1971). Universal kriging(an optimal method for estimating and contouring in trend surface analysis). In McGerrigle, J. I., editor, *Proceedings of Ninth International Symposium on Techniques for Decision-Making in the Mineral Industry*, pages 159–169.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press: London.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kitanidis, P. K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research*, 22:499–507.

Laplace, P. S. (1818). *Deuxiem Supplement a la Theorie analytique des Probabilities*. Gauthier-Villars: Paris. Reprinted (1847) in Oeuvers Completes de Laplace, Vol 7.

Lauritzen, S. L., Thiesson, B., and Spiegelhalter, D. J. (1994). Diagnostic systems created by model selection methods: a case study. In Cheeseman, P. and Oldford, W., editors, *Uncertainty in Artificial Intelligence*, pages 143–152. Springer: Berlin.

Leamer, E. E. (1978). *Specification Searches.* Wiley: New York.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146.

Mardia, K. V. and Watkins, A. J. (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika*, 76:289–295.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models.* Irwin: Chicago, fourth edition.

Omre, H. and Halvorsen, K. B. (1989). The Bayesian bridge between simple and universal kriging. *Mathematical Geology*, 21:767–786.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, NEED.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83:251–266.

Ripley, B. D. (1981). *Spatial Statistics.* Wiley: New York.

Roberts, H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Association*, 60:50–62.

Schwartz, G. (1978). Estimating the dimension of a model. *Journal of the American Statistical Association*, 6:461–446.

Smith, R. L. (2000). Spatial statistics in environmental science. In *Nonlinear and Nonstationary Signal Processing.* Cambridge University Press.

Spiegelhalter, D. J., Dawid, A., Lauritzen, S., and Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8:219–283.

Stein, M. L. (1999). *Interpolation of Spatial Data.* Springer: New York.

Stigler, S. M. (1973). Studies in the history of probability and statistics. XXXII: Laplace, fisher and the discovery of the concept of sufficiency. *Biometrika*, 60:439–445.

Ver Hoef, J. M., Cressie, N., Fisher, R. N., and Case, T. J. (1999). Uncertainty and spatial linear models for ecological data.

Volinsky, C. T. (1997). *Bayesian Model Averaging for Censored Survival Models.* PhD thesis. University of Washington.

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Journal of the Royal Statistical Society, Series C*, 46:433–448.

Warnes, J. J. and Ripley, B. D. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, 74:640–642.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41:434–449.

Zimmerman, D. L. and Zimmerman, M. B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, 33:77–91.